

International Journal of Advanced AI Applications

Volume 2, Issue 10, Oct. 2026

Hong Kong Dawn Clarity Press Limited
<http://www.dawnclarity.press/index.php/ijaaa>

Online ISSN 3104-9338

Print ISSN 3104-932X



TABLE OF CONTENTS

Quality-Aware Photovoltaic Forecasting under Missing Weather Data

Haitian Ren..... (1-26)

Small-Signal Stability-Constrained Source–Storage Capacity Optimization for Islanded Microgrids

Jie Yin..... (27-50)

Stock Market Liberalization and Institutional Investor Clique Behavior: Evidence from China

Muzi Jiang, Ping Yang.....(51-76)

Where Vision Meets Language: A Coupling-Locus Taxonomy of Multimodal Large Language Models

Haosong Hu, Xinyu Song.....(77-101)

Right Answers, Wrong Reasons: Perception Errors in Multimodal Chain-of-Thought and the Blind Spot of Outcome-Based Rewards

Yanwei Zhang, Xinyu Song..... (102-124)

Impressum (125-126)

Quality-Aware Photovoltaic Forecasting under Missing Weather Data

Haitian Ren*

School of Energy, Belarusian National Technical University, Minsk, Belarus

Received: September 19, 2026

Revised: September 20, 2026

Accepted: September 22, 2026

Published online: September 22, 2026

To appear in: *International Journal of Advanced AI Applications*, Vol. 2, No. 10 (October 2026)

Corresponding Author:
Haitian Ren
(haitian.ren.ee@gmail.com)

Abstract. Operational photovoltaic forecasts often depend on irradiance and temperature measurements that may be interrupted by sensor or communication faults. This study develops Quality-Aware Dual-Path Conformal Photovoltaic Forecasting (QDPCC-PV), which combines a weather-aware Extra Trees expert with a power-only fallback expert. An availability gate changes their weights according to observed weather inputs, while physical clipping enforces feasible output and condition-specific conformal calibration provides prediction intervals. The method was evaluated chronologically on two measured photovoltaic plants sampled every 15 min at 15-, 60-, and 240-min horizons. Under complete inputs, mean absolute errors were 0.0376, 0.0463, and 0.0684 per unit, corresponding to improvements of 32.8%, 45.7%, and 65.3% over persistence. With 60% random weather loss at the 60-min horizon, QDPCC-PV obtained 0.0492 error, compared with 0.0664 for median imputation. The nominal 90% interval reached 91.8% coverage at 240 min but undercovered at shorter horizons because calibration and test residuals shifted over time. The results show that explicit input-quality routing protects point forecasts from weather-data loss, while adaptive calibration remains necessary for reliable short-horizon intervals.

Keywords: *sensor outages; extra trees; mixture of experts; interval calibration; multi-horizon prediction*

1. Introduction

The rapid growth of photovoltaic (PV) generation has made short-horizon forecasting essential for reserve scheduling, market bidding, battery dispatch, and voltage control. Yet operational forecasts often depend on irradiance and temperature streams that can fail because of sensor faults, maintenance, communication loss, or delayed weather services. In such cases,

conventional imputation makes the input numerically complete without making it reliable, and a multivariate model may perform worse than a target-only model [1,2,15].

This study focuses on 15-, 60-, and 240-min horizons, spanning immediate control, intra-hour scheduling, and short-term planning. Existing physical, statistical, machine-learning, and deep temporal models improve clean-data accuracy in different settings [3-14], but most assume that exogenous inputs remain available at inference time. They also commonly report point forecasts alone, although operators need uncertainty estimates that respond to data quality.

These gaps motivate Quality-Aware Dual-Path Conformal Photovoltaic Power Forecasting (QDPCC-PV). The method uses a weather-aware expert when meteorological inputs are observed, progressively transfers weight to a power-only expert as those inputs disappear, enforces physical output bounds, and calibrates uncertainty separately by horizon and missingness condition. The resulting design isolates a practical question: can explicit routing provide graceful degradation without hiding the effect behind a high-capacity predictor?

The contributions of this work are summarized as follows.

- A dual-path forecasting structure is developed with a meteorology-aware expert and a power-only fallback expert. Unlike unconditional imputation, the structure remains explicitly aware of which variables are observed.
- A transparent quality gate converts the observed-feature ratio into expert weights. The mechanism is directly auditable and can later be replaced by a learned gate using gap duration, information age, and regime descriptors.
- Physical projection and missingness-conditioned split conformal prediction are integrated with the point forecast to produce feasible outputs and empirical uncertainty intervals.
- A reproducible chronological experiment evaluates complete data, random loss, block outages, and total weather loss using two measured PV plants and three horizons. Both positive and negative findings, including short-horizon interval undercoverage, are reported.

2. Related Work

2.1. Statistical and Machine-Learning Methods

Persistence assumes that future power equals the most recent observation. Although simple, it is a strong very-short-term baseline because PV output is highly autocorrelated. Linear and autoregressive models add trend and seasonal components but remain limited when weather-power relations change across irradiance and temperature regimes. Support vector regression

maps the inputs into a nonlinear feature space, while decision-tree ensembles approximate nonlinear interactions without extensive scaling. Random Forest (RF) averages randomized decision trees to reduce variance [20]. Extra Trees (ET), formally Extremely Randomized Trees, additionally randomizes split thresholds and can reduce training cost and variance [21]. These properties make ET suitable for a transparent pilot implementation of a dual-expert architecture.

Tree ensembles are also useful when the available dataset is smaller than the training corpora normally used for deep temporal networks. They accept mixed nonlinear predictors, do not require gradient-based optimization, and expose feature-importance summaries that support diagnostic analysis. Their principal limitation is that temporal structure must be supplied explicitly through lags, rolling statistics, and calendar variables. They also extrapolate poorly outside the target range represented in training. The present study addresses the second limitation with a physical projection and accepts the first as a deliberate trade-off: the aim is to isolate the value of quality-aware routing before introducing a higher-capacity backbone. This separation is important because a complicated predictor can otherwise obscure whether robustness originates from the model architecture, the imputation rule, or the routing mechanism.

2.2. Deep Temporal Models

Deep temporal models such as LSTM, TCN, and Transformer architectures can learn long-range and multi-scale dependencies [5-10]. PVPNet is particularly relevant because it processes heterogeneous PV inputs through separate streams [1], while recent Transformer variants add physical features, frequency decomposition, or feature selection [11-13]. However, higher representational capacity does not by itself define how a forecaster should degrade when exogenous variables are missing. This study therefore uses the same Extra Trees family for both experts to isolate the value of the routing mechanism from backbone capacity and tuning effects; replacing the experts with deep models is left for future work.

2.3. Missing Data and Domain Shift

Numerical weather prediction (NWP) can differ systematically from local measured data. Domain-generalization and weather-correction methods attempt to learn representations that transfer across seasons or stations [22]. Missing-data research uses interpolation, matrix completion, generative imputation, mask-aware attention, transfer learning, and direct forecasting from incomplete covariates. Yan et al. [15] reported that covariates are useful only when their quality is sufficiently high. This result supports a conditional architecture in which

the contribution of meteorological inputs depends on observed quality rather than remaining fixed.

Missingness must be distinguished by mechanism and duration. Independent random deletion approximates packet loss or isolated sensor errors, whereas block deletion represents communication outages, maintenance periods, or frozen acquisition systems. Missing not at random is more difficult because the probability of absence may depend on the weather regime or the sensor value itself; for example, a pyranometer may saturate or fail preferentially during extreme conditions. Simple interpolation is reasonable for very short smooth gaps but can suppress ramps across longer gaps. Statistical imputation expresses a central tendency but understates uncertainty, and learned imputation can transfer bias from the reconstruction model into the forecaster. The proposed design does not claim to reconstruct the missing weather state. Instead, it explicitly reduces reliance on the affected expert and retains a valid power-only route.

2.4. Probabilistic Forecasting

Quantile regression directly estimates conditional quantiles but requires a separate loss design and can produce crossing quantiles. Parametric methods specify distributions such as Gaussian or mixture densities; misspecification can yield unreliable tails. Large-scale energy forecasting competitions have reinforced the need to evaluate both accuracy and uncertainty [23]. CP instead ranks residual nonconformity scores on a held-out calibration set. Conformalized quantile regression combines learned quantiles with conformal correction [17], while dynamic time-series conformal methods relax some independence assumptions [18]. Renkema et al. [24] found that K-nearest-neighbor and Mondrian conformal variants improve photovoltaic interval quality. Adaptive conformal inference updates the error rate online under distribution shift [19]. QDPCC-PV uses a simpler missingness-conditioned split method so that its limitations can be measured directly. Table 2 positions QDPCC-PV against representative studies in terms of missing-weather handling and interval prediction.

The attraction of conformal prediction is model agnosticism: the point predictor can be changed without redesigning the interval construction. Its finite-sample marginal guarantee, however, depends on exchangeability between calibration and future nonconformity scores. Chronological solar data violate ideal exchangeability through seasonality, weather regimes, equipment changes, and autocorrelated ramps. A single global residual quantile can also mix daytime and nighttime cases or complete and incomplete inputs. Conditioning calibration on horizon and missingness reduces one source of heterogeneity, but it does not remove temporal drift. For this reason, the empirical coverage reported later is treated as an evaluated

performance measure rather than assumed from theory. Undercoverage is retained and discussed because hiding it would give a misleading impression of operational reliability.

Table 1. Symbols and abbreviations.

Symbol	Definition	Unit
PV	Photovoltaic	-
ML	Machine learning	-
DL	Deep learning	-
ET	Extra Trees	-
CP	Conformal prediction	-
PICP	Prediction interval coverage probability	-
MPIW	Mean prediction interval width	p.u.
p_t	Normalized PV power at time t	p.u.
I_t	Solar irradiance	dataset unit
T_t^{amb}	Ambient temperature	°C
T_t^{mod}	Module temperature	°C
$m_{t,j}$	Availability indicator for feature j	0 or 1
g_t	Meteorological-expert gate weight	0-1
h	Forecast horizon	steps
L	Look-back length	steps
α	Target miscoverage rate	-

Table 2. Comparison with related studies.

Study	Core method	Weather loss	Intervals
Huang and Kuo [1]	Multi-input 1D CNN	No	No
Tao et al. [12]	Physics-aided Transformer	No	No
Yan et al. [15]	Missing-aware transfer	Yes	No
Renkema et al. [24]	Regression plus CP	No	Yes
El-Kenawy et al. [11]	PatchTST plus SMOA	No	No
QDPCC-PV	Quality-gated dual ET	Yes	Yes

3.1. Input and Output

Consider a plant-level sequence sampled every $\Delta t = 15$ min. The normalized active power at time t is p_t . The weather vector is $x_t = [I_t, T_t^{\text{amb}}, T_t^{\text{mod}}]$, where I_t is irradiance, T_t^{amb} is ambient temperature, and T_t^{mod} is module temperature. The binary vector m_t has dimension $d = 3$; $m_{t,j}$ equals one when weather variable j is observed and zero when it is missing. The model uses a power history of $L = 192$ steps, corresponding to 48 h, and produces a point prediction and prediction interval for horizon h. Table 1 defines the principal symbols and

abbreviations used throughout the formulation.

$$\hat{p}_{t+h}, C_{t+h} = F(p_{t-L+1:t}, x_t, m_t, h) \quad (1)$$

Here F denotes the complete forecasting operator, $p_{t-L+1:t}$ is the historical power window, C_{t+h} is the interval forecast, and h belongs to $\{1, 4, 16\}$, corresponding to 15, 60, and 240 min. The target power is normalized to per unit (p.u.) by a robust plant capacity proxy defined as the 99.9th percentile of observed plant AC power. This choice limits sensitivity to isolated maximum values while retaining a physically interpretable scale.

3.2. Missing-Data Conditions

Three input-quality states are considered. Complete data contain all weather channels. Missing completely at random removes each weather entry independently with probability ρ . Block missingness removes consecutive segments of all weather variables for 2-8 h, representing communication or sensor outages. Total loss sets $\rho = 1$ and forces the system to operate only from historical power and time variables. Artificial deletion is applied only to calibration and test covariates; the target sequence is never deleted.

3.3. Objectives

The point-forecast objective is to minimize expected absolute or squared error while maintaining stability as ρ increases. The probabilistic objective is to construct an interval whose empirical coverage approaches $1-\alpha$ while remaining narrow. These objectives can conflict: increasing interval width improves coverage but reduces sharpness. Accordingly, both prediction interval coverage probability and mean prediction interval width are reported.

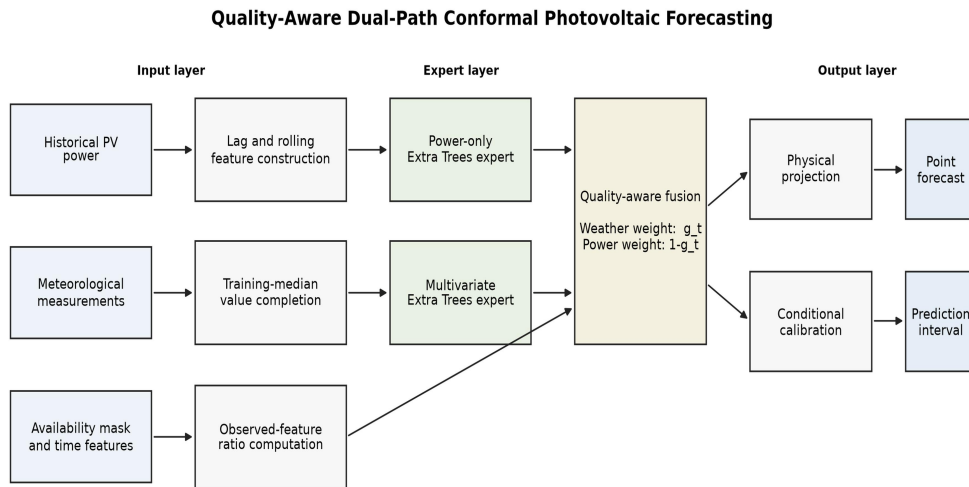


Figure 1. QDPCC architecture.

4. Proposed Method

4.1. Feature Construction

Figure 1 summarizes the dual-path architecture and its routing, projection, and calibration stages. The power-only feature vector s_t contains seven direct lags, rolling means and standard deviations, and periodic calendar encodings. Lags 1, 2, 4, 8, and 16 represent the previous 15 min to 4 h. Lags 96 and 192 represent the same time one and two days earlier. Rolling means and standard deviations are computed over 4, 16, and 96 steps, corresponding to 1 h, 4 h, and 24 h. Calendar variables include sine and cosine encodings of time of day and day of year. Cyclic encoding avoids an artificial discontinuity between 23:45 and 00:00.

$$s_t = [p_{t-1} p_{t-2} p_{t-4} p_{t-8} p_{t-16} p_{t-96} p_{t-192} \mu_t \sigma_t c_t] \quad (2)$$

In (2), μ_t and σ_t are vectors of rolling means and standard deviations, and c_t is the cyclic calendar vector. The multivariate feature vector z_t augments s_t with the three weather variables.

$$z_t = [s_t, I_t, T_t^{\text{amb}}, T_t^{\text{mod}}] \quad (3)$$

Missing weather entries are filled with medians computed from the training split. This operation supplies a valid numeric input to the tree ensemble, but the mask is retained and affects the gate. Consequently, a filled value is never treated as an observed value by the complete system.

All temporal predictors are causal. A lag indexed by k contains power measured at $t-k$, and every rolling statistic is calculated after shifting the power series by one sample. Thus, the rolling window ending at forecast origin t never includes the target p_{t+h} . Calendar variables are known in advance and do not create leakage. The same feature schema is built before the chronological split, but any statistic requiring fitted information, including weather medians, is estimated only from the training interval. This distinction prevents an apparently minor preprocessing step from transferring information from calibration or test periods into training. At inference, the feature builder also records whether each weather value was observed, imputed, or unavailable so that the routing decision can be audited independently of the numerical values supplied to the regressor.

4.2. Power-Only Expert

The power-only expert f_p is an ET regressor. Each tree is fitted to a bootstrap-free randomization of samples and features, and its split thresholds are selected randomly from

candidate ranges. The ensemble prediction is the arithmetic mean of all tree outputs. This expert does not receive irradiance or temperature, so its input distribution is unchanged when all weather channels fail. It acts as a fail-operational fallback rather than as a competing final model.

4.3. Weather-Aware Expert

The multivariate expert f_M uses the same lag and time features as f_P and additionally receives irradiance, ambient temperature, and module temperature. Irradiance provides the strongest instantaneous physical relation to PV output, whereas temperature captures efficiency and thermal-state effects. Using the same ET family for both experts isolates the value of routing from differences in model capacity. Each expert contains 180 trees, uses a minimum leaf size of two, and considers 80% of the features at each split.

4.4. Quality Gate

The quality gate is the fraction of weather channels observed at the current forecast origin. With d weather variables, the gate is defined by (4). When all variables are present, $g_t = 1$ and the multivariate expert is used. When all are absent, $g_t = 0$ and the power-only expert is used. Partial availability produces a convex combination.

$$g_t = \frac{1}{d} \sum_{j=1}^d m_{t,j}, 0 \leq g_t \leq 1 \quad (4)$$

$$\hat{p}_{t+h}^F = g_t f_M(z_t) + (1 - g_t) f_P(s_t) \quad (5)$$

The deterministic gate is deliberately interpretable. It avoids training an additional network on a dataset that spans only one month. For larger datasets, g_t can be learned from the mask, consecutive gap length, information age, disagreement between measured and forecast irradiance, weather regime, and forecast horizon. The deterministic form therefore represents the minimum auditable implementation of the broader quality-aware concept.

The convex fusion has several useful boundary properties. If the three weather variables are observed, the output equals the weather-aware expert; if none is observed, the output equals the power-only expert; and for partial availability the forecast remains between the two expert outputs before physical clipping. Because the gate depends only on availability, it cannot improve the forecast by inspecting the unknown target. It also avoids abrupt switching when one of several channels disappears. Nevertheless, equal weighting of weather variables is a simplification. Irradiance normally contains more direct predictive information than ambient temperature, and a stale observation should not receive the same confidence as a current observation. A learned gate should therefore be constrained, calibrated, and tested under

deliberately shifted conditions rather than optimized only for average clean-data accuracy.

4.5. Physical Projection

Tree ensembles generally remain within the target range observed during training, but fusion and future extensions may produce negative or above-capacity values. The physical projection in (6) clips the normalized forecast to $[0,1]$. This constraint represents nonnegative production and the plant-capacity limit. It does not model curtailment, cloud enhancement, inverter clipping, or detailed module physics.

$$\hat{p}_{t+h} = \min\left(1, \max(0, \hat{p}_{t+h}^F)\right) \quad (6)$$

4.6. Conditional Conformal Calibration

After both experts are trained, their parameters are frozen. The chronological calibration split is passed through the same missingness condition used at test time. For each horizon h and condition c , the absolute error is used as a nonconformity score.

$$s_{i,h,c} = |p_{i+h} - \hat{p}_{i+h}| \quad (7)$$

$$q_{h,c} = Q_{1-\alpha}(\{s_{i,h,c} : i \in \mathcal{I}_{\text{cal},c}\}) \quad (8)$$

$Q_{1-\alpha}$ is the finite-sample empirical quantile and $\mathcal{I}_{\text{cal},c}$ is the set of calibration indices under condition c . With $\alpha = 0.10$, the nominal interval level is 90%. The lower and upper limits are physically clipped.

$$C_{t+h} = [\max(0, \hat{p}_{t+h} - q_{h,c}), \min(1, \hat{p}_{t+h} + q_{h,c})] \quad (9)$$

Conditioning on missingness allows the interval to widen when input quality decreases. It does not guarantee conditional coverage under arbitrary temporal drift. The empirical results quantify this distinction.

Calibration is performed after point-model training. For each plant, horizon, missingness pattern, and missing rate, the calibration covariates are corrupted by the same generator used to define the corresponding test condition. QDPCC-PV produces calibration forecasts, and their absolute residuals form the nonconformity sample. The finite-sample quantile uses the ceiling-adjusted rank associated with $(1-\alpha)(n+1)$, where n is the number of calibration residuals. At prediction time, the selected quantile is added to and subtracted from the point forecast, after which both limits are clipped to the physical range. No test residual is included in the quantile. This procedure makes the interval reproducible and prevents the common error of tuning interval width on the evaluation set.

4.7. Procedure

Training and inference consist of seven stages: aggregate inverter power; construct lags and rolling statistics; create chronological splits; train the two experts; generate missingness-conditioned calibration residuals; compute the quality-aware fused forecast; and evaluate point and interval metrics. No test observation is used for model fitting, feature scaling, median calculation, or interval quantile estimation. Figure 2 presents this end-to-end experimental workflow.

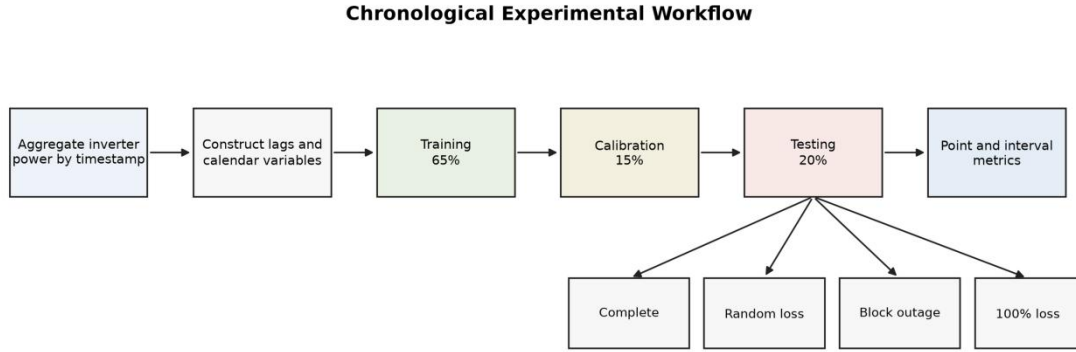


Figure 2. Experimental workflow.

5. Experiments

5.1. Datasets

Experiments use the public Solar Power Generation Data collection [25]. It contains two grid-connected PV plants in India, each with 22 inverter identifiers and a plant-level weather sensor. Generation and weather files are sampled every 15 min between May and June 2020. Available electrical variables include direct-current power, alternating-current power, daily yield, and total yield. Available meteorological variables include irradiance, ambient temperature, and module temperature. Figure 3 shows representative five-day power and weather profiles, and Table 3 summarizes the two plant records.

Generation data were aggregated to the plant level by summing inverter alternating-current power at each timestamp. Weather observations repeated across inverter rows were averaged. A complete 15-min timestamp index was constructed. Short weather gaps in the clean reference sequence were linearly interpolated for up to four observations and then forward/backward filled at the boundaries. Missing plant power timestamps were set to zero only when the aligned plant record contained no generation observation. After the 192-step history requirement, 3,072 samples remained for each plant, covering 17 May to 17 June 2020.

The public files do not provide a complete engineering description of module technology, tilt, azimuth, inverter clipping logic, curtailment commands, maintenance events, or sensor calibration. Consequently, the experiment is a data-driven plant-level study rather than a detailed physical simulation. The two plants also cover the same limited seasonal window, so cross-season and cross-climate generalization cannot be inferred. These constraints motivate the chronological design and the conservative interpretation of the results. The dataset is nevertheless suitable for the present question because it provides synchronized generation and meteorological measurements, multiple inverter streams, repeated daily cycles, and enough naturally variable conditions to construct controlled weather-input outages without altering the target observations. The feature-correlation structure used to support predictor selection is shown in Figure 4.

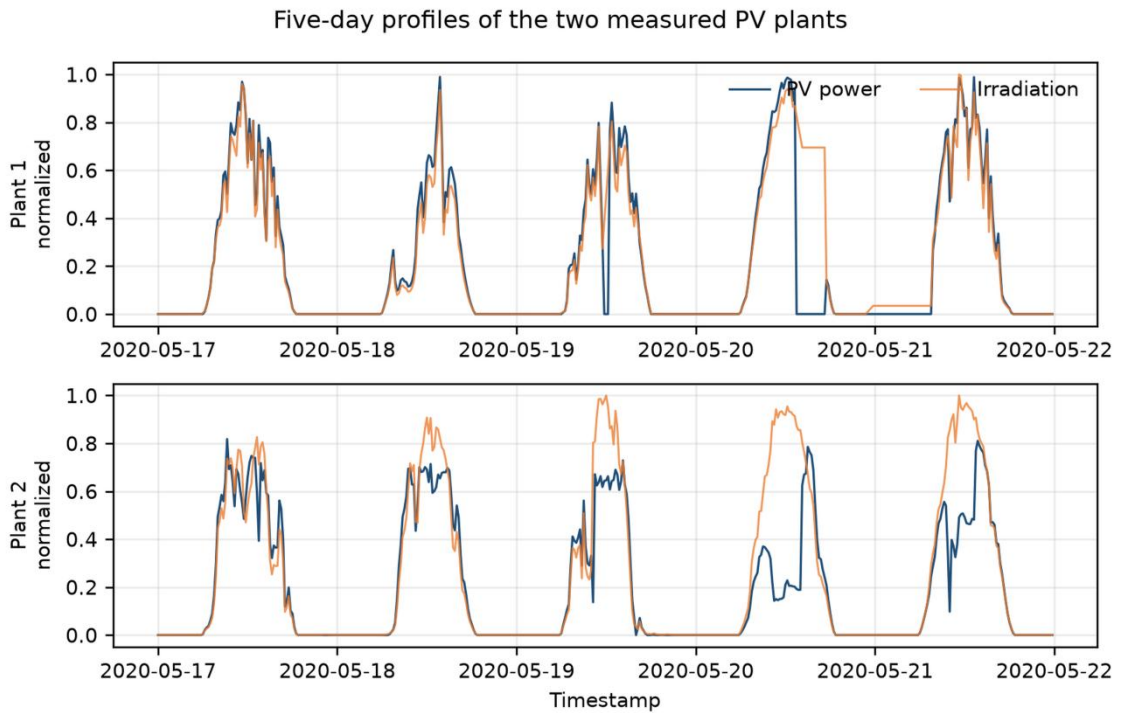


Figure 3. Five-day PV and weather profiles.

Table 3. Dataset summary.

Plant	Samples	Period	Capacity proxy	Daylight
1	3,072	May 17-Jun. 17	27,886.8 kW	34.6%
2	3,072	May 17-Jun. 17	24,440.5 kW	51.5%

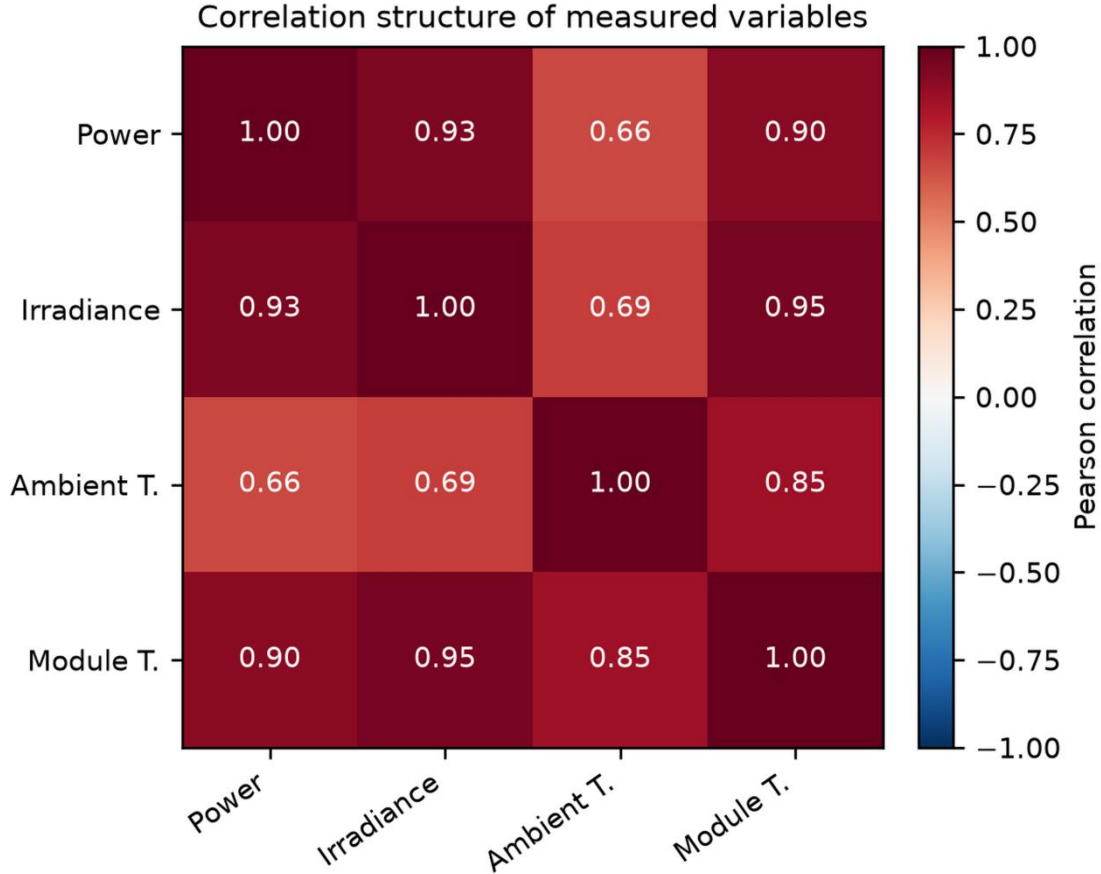


Figure 4. Feature correlation matrix.

5.2. Baselines

Four forecasting configurations are compared. Persistence sets the future prediction equal to the most recent power observation. Power-only ET uses the historical-power and calendar features described in Section 4.1. Median-impute ET uses all features but fills missing weather values with training medians and does not change model weight according to availability. QDPCC-PV combines the two experts using (5). Because the study focuses on missingness routing, identical ET settings are used for the two learned baselines and both QDPCC-PV experts.

PVPNet, LSTM, PatchTST, and the PatchTST-Swordfish model are discussed as literature references but are not assigned numerical results in this pilot because an honest comparison would require reimplementing, multiple random seeds, and the longer training periods used in their original studies. Fabricated or copied metrics from different datasets would not constitute a valid baseline.

5.3. Preprocessing

Timestamps were sorted before any split was performed. The first 65% of each plant sequence was used for training, the next 15% for conformal calibration, and the final 20% for testing. Sliding-window samples that required information outside their own chronological split were excluded. All medians and the capacity proxy were obtained without using future target values. Time-of-day and day-of-year features were encoded with sine and cosine functions. Power lags and rolling statistics used a one-step shift, preventing the current or future target from entering the feature vector.

Power was normalized separately for each plant by a high quantile of observed alternating-current output rather than by the absolute maximum. This reduces sensitivity to isolated spikes while retaining a comparable per-unit scale. Nighttime observations were retained because zero-output periods are part of operational forecasting, but performance is also interpreted in light of the resulting class imbalance between easy nighttime cases and more variable daylight cases. The corruption generator modifies only meteorological predictors; historical power lags, calendar variables, and targets remain unchanged. This isolates the effect of weather availability and prevents a mixed experiment in which missing targets, missing covariates, and target reconstruction are evaluated simultaneously.

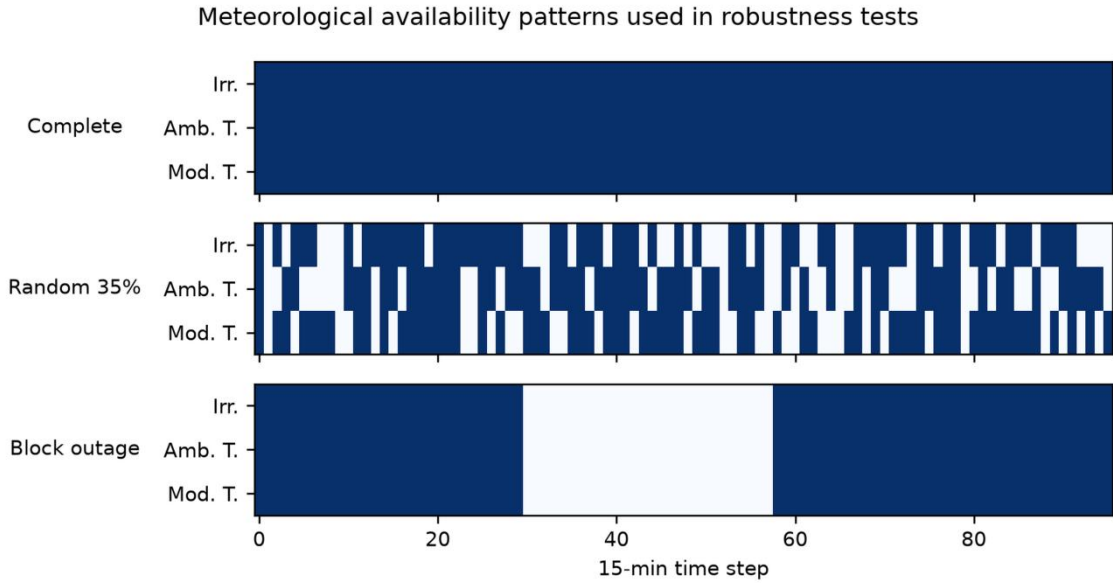


Figure 5. Simulated missing-data patterns.

Figure 5 illustrates the random, block, and total-loss patterns used in the corruption experiments. Random missingness rates were 0%, 30%, 60%, and 100%. For block experiments, contiguous segments between 8 and 32 samples, equivalent to 2-8 h, were repeatedly removed

until the requested missing fraction was reached. A fixed random generator seed, 20260917, was used to make the corruption experiment reproducible.

5.4. Training

Each horizon and plant was trained independently. ET regressors contained 180 trees, minimum leaf size two, and a maximum-feature fraction of 0.8. Trees were trained in parallel using all available processor cores. The random state was deterministically derived from the plant identifier and forecast horizon. These settings were selected as a stable computational configuration rather than by optimizing against the test set. The point model was fitted only on clean training data. Calibration and testing were repeated for each missingness pattern and rate.

The independent horizon design avoids recursive error accumulation: a 240-min forecast is trained directly for that lead time rather than obtained by repeatedly applying a 15-min model. It also allows the relative value of weather and lag features to change with horizon. The training procedure does not use early stopping because ET builds a fixed number of randomized trees, and no validation score is used to choose among configurations. The calibration split has a single purpose: estimating residual quantiles after the point models are frozen. This separation is stricter than using one validation interval for both hyperparameter selection and uncertainty calibration. For larger benchmark experiments, a nested chronological validation period should be introduced before the calibration block so that hyperparameters, interval quantiles, and final test metrics remain statistically distinct.

Table 4. Experimental settings.

Parameter	Value	Meaning
Δt	15 min	Sampling interval
L	192	48-h look-back
h	1, 4, 16	15-, 60-, 240-min horizons
Trees	180	Trees per expert
Leaf size	2	Minimum samples per leaf
Feature fraction	0.8	Candidate-feature fraction
Train/cal/test	65/15/20%	Chronological partition
ρ	0, .3, .6, 1	Weather missing rate
α	0.10	Nominal miscoverage
Seed	20260917	Corruption reproducibility

5.5. Metrics

Mean absolute error (MAE) measures the average magnitude of point errors and is less sensitive to large ramps than squared error. Root mean squared error (RMSE) emphasizes large deviations. The coefficient of determination R^2 compares residual variance with the variance of

observed power. Let N be the number of test observations, p_n the observed normalized power, $\hat{p}_n$ the point forecast, and $\bar{p}$ the sample mean.

$$\text{MAE} = \frac{1}{N} \sum_{n=1}^N |p_n - \hat{p}_n| \quad (10)$$

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{n=1}^N (p_n - \hat{p}_n)^2} \quad (11)$$

$$R^2 = 1 - \frac{\sum_{n=1}^N (p_n - \hat{p}_n)^2}{\sum_{n=1}^N (p_n - \bar{p})^2} \quad (12)$$

Prediction interval coverage probability (PICP) is the proportion of targets lying between lower limit l_n and upper limit u_n . Mean prediction interval width (MPIW) measures sharpness. High-quality intervals should attain coverage close to the nominal level without becoming unnecessarily wide.

$$\text{PICP} = \frac{1}{N} \sum_{n=1}^N \mathbb{I}(l_n \leq p_n \leq u_n) \quad (13)$$

$$\text{MPIW} = \frac{1}{N} \sum_{n=1}^N (u_n - l_n) \quad (14)$$

Metrics are first calculated separately for each plant and then averaged, preventing the plant with more usable test samples from dominating the result. All power errors and widths are expressed in per unit. The present pilot uses one deterministic seed for model fitting; therefore, the reported table should not be interpreted as a multi-seed significance study.

MAE and RMSE answer different operational questions. MAE describes typical absolute deviation, whereas RMSE assigns greater influence to rare large errors that may drive reserve requirements. R^2 is included to show explained variation but can be misleading when the test segment contains many nighttime zeros. PICP and MPIW must be read together: a very wide interval can achieve high coverage without being useful, while a narrow interval with low coverage is overconfident. The reported metrics therefore support complementary conclusions rather than a single ranking. A production study should additionally report daylight-only errors, normalized ramp error, pinball loss or weighted interval score, inference latency, and confidence intervals across sites and seeds.

6. Results and Discussion

6.1. Clean-Input Accuracy

Table 5 reports performance with complete meteorological information. QDPCC-PV assigns $g_t = 1$ in this condition and consequently equals the multivariate expert. At 15 min, its MAE was 0.0376 p.u., compared with 0.0560 for persistence and 0.0426 for power-only ET. At 60

min, its MAE was 0.0463, and at 240 min it was 0.0684. Relative to persistence, the reductions were 32.8%, 45.7%, and 65.3%, respectively. The increasing relative benefit at longer horizons reflects the rapid decay of the persistence assumption. Figure 6 visualizes the corresponding error comparison across horizons.

Table 5. Clean-input results.

Horizon	Model	MAE	RMSE	R ²
15	Persistence	0.0560	0.1072	0.805
15	Power-only ET	0.0426	0.0866	0.874
15	Median-impute ET	0.0376	0.0790	0.894
15	QDPCC-PV	0.0376	0.0790	0.894
60	Persistence	0.0853	0.1455	0.653
60	Power-only ET	0.0491	0.0955	0.848
60	Median-impute ET	0.0463	0.0926	0.857
60	QDPCC-PV	0.0463	0.0926	0.857
240	Persistence	0.1969	0.3000	-0.419
240	Power-only ET	0.0730	0.1365	0.711
240	Median-impute ET	0.0684	0.1261	0.753
240	QDPCC-PV	0.0684	0.1261	0.753

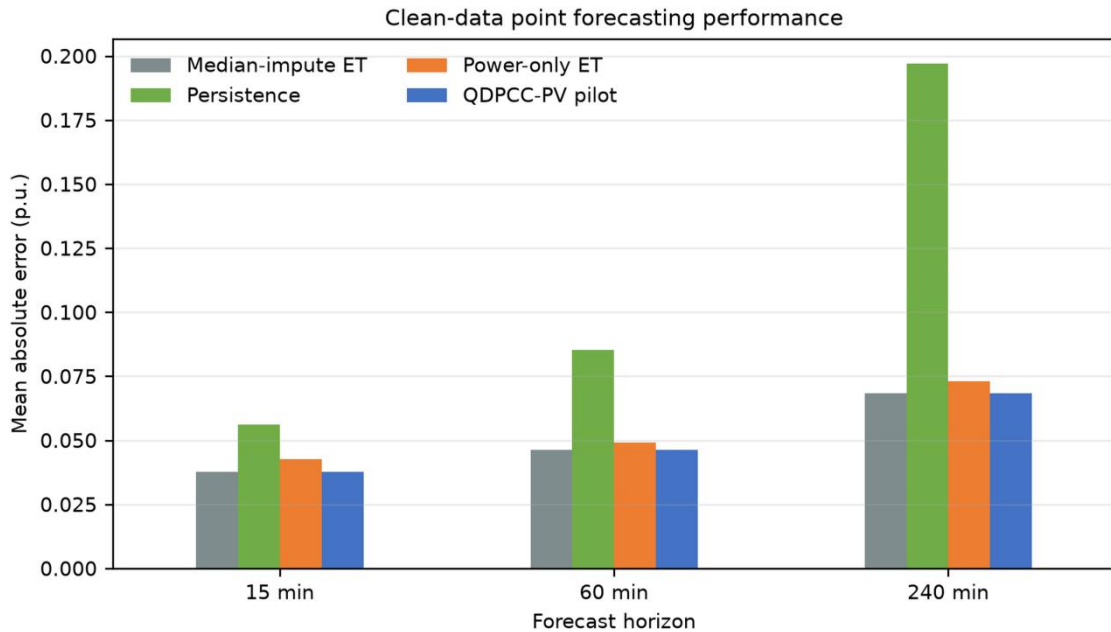


Figure 6. Clean-input forecasting errors.

The power-only expert remained competitive, confirming that lagged generation encodes recent cloud and plant behavior. The additional weather variables reduced MAE by 11.6% at 15 min, 5.7% at 60 min, and 6.3% at 240 min relative to power-only ET. These modest gains

illustrate why a multivariate model should not be trusted unconditionally when the added variables are unreliable. Figure 7 provides a representative observed-versus-predicted sequence at the 60-min horizon.

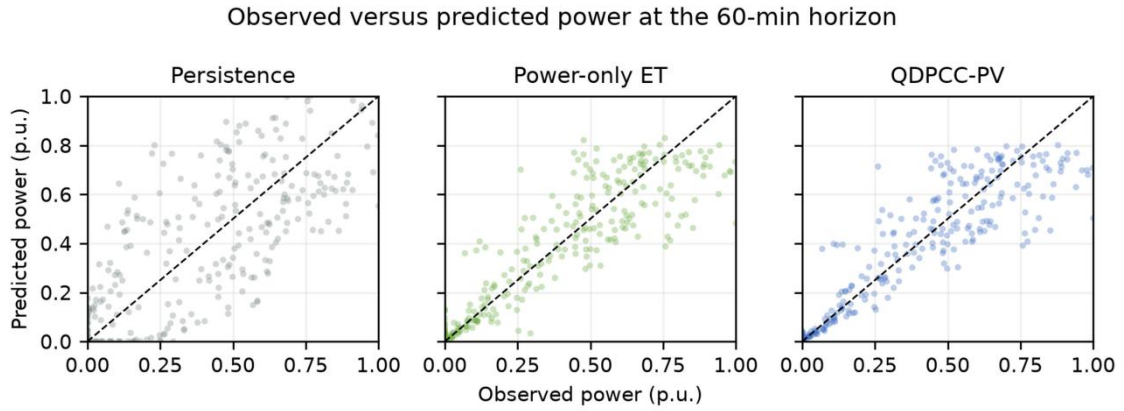


Figure 7. Observed and predicted power at 60 min.

6.2. Missing-Data Robustness

Figure 8 compares error as random meteorological missingness increases at the 60-min horizon. Persistence and power-only ET remain constant because neither receives weather variables. Median-impute ET degrades monotonically from 0.0463 MAE with complete inputs to 0.0787 under total weather loss. QDPCC-PV transfers weight toward the power-only expert and remains near 0.049 p.u. across the full range.

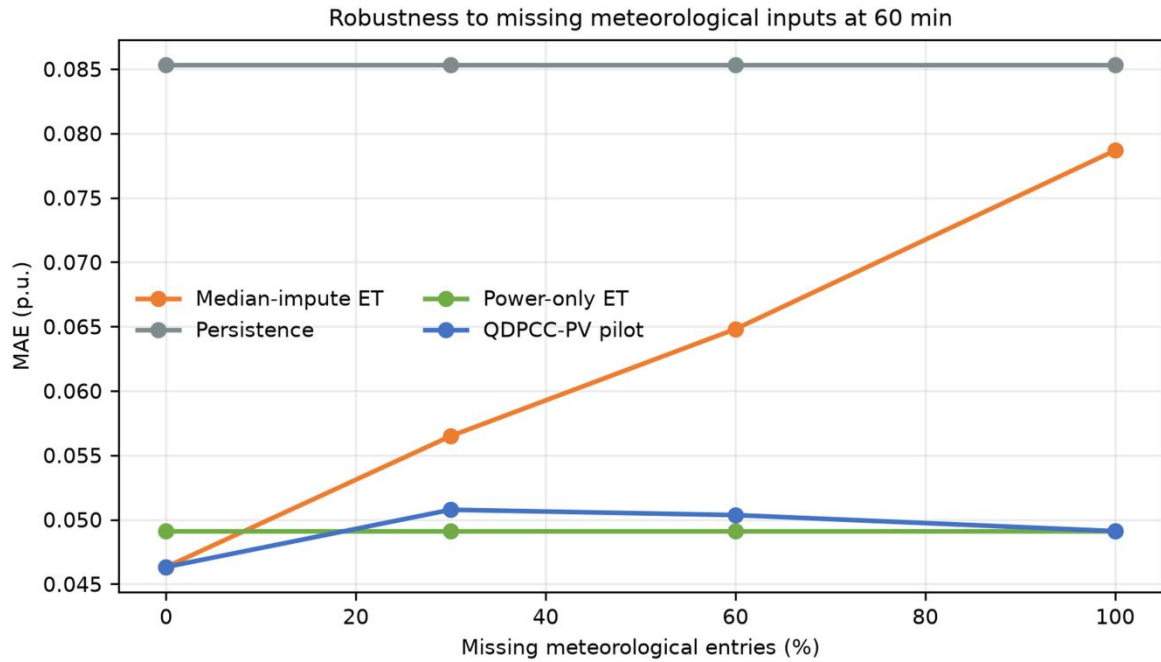


Figure 8. Error under weather-data loss.

Table 6. Results under random weather loss.

Missing	Model	MAE	RMSE
0%	Power-only ET	0.0491	0.0955
0%	Median-impute ET	0.0463	0.0926
0%	QDPCC-PV	0.0463	0.0926
30%	Power-only ET	0.0491	0.0955
30%	Median-impute ET	0.0565	0.1070
30%	QDPCC-PV	0.0508	0.0976
60%	Power-only ET	0.0491	0.0955
60%	Median-impute ET	0.0648	0.1214
60%	QDPCC-PV	0.0503	0.0963
100%	Power-only ET	0.0491	0.0955
100%	Median-impute ET	0.0787	0.1386
100%	QDPCC-PV	0.0491	0.0955

At 60% missingness, QDPCC-PV achieved 0.0492 MAE, whereas median-impute ET produced 0.0664. The 25.8% reduction supports the use of explicit quality routing. QDPCC-PV was approximately equal to the power-only expert at this point; the result is expected because only 40% of the weather entries remained available. Under total loss, both methods coincide by design.

The comparison also clarifies what the gate does and does not accomplish. It does not reconstruct irradiance, and it cannot make partially observed weather as informative as complete measurements. Instead, it limits the damage caused by presenting an imputed vector to a model trained on genuine observations. At intermediate missing rates, the fused prediction preserves part of the clean-input benefit while moving toward the stable fallback. Under complete loss, equivalence with the power-only expert is a required boundary check rather than a failure to improve. For block outages, the same logic operates over consecutive timestamps, avoiding the repeated use of a median-filled weather state throughout the gap. The remaining error is governed by the predictive content of recent power, daily periodicity, and the ability of the fallback expert to track ramps without direct weather information.

The scenario matrix is designed to separate robustness from clean-data accuracy. A model could achieve the smallest error with complete inputs yet fail sharply after a single weather channel disappears. Conversely, a power-only model can be stable but leave useful irradiance information unused. Reporting both clean and corrupted conditions reveals this trade-off. The median-imputation baseline is particularly informative because it uses the same weather-aware learner but lacks explicit routing; its deterioration therefore cannot be attributed to reduced tree capacity. The comparison indicates that numerical completeness is not equivalent to informational reliability. This observation is relevant beyond PV forecasting whenever

auxiliary sensors are optional, delayed, or maintained by an external service.

6.3. Feature and Residual Analysis

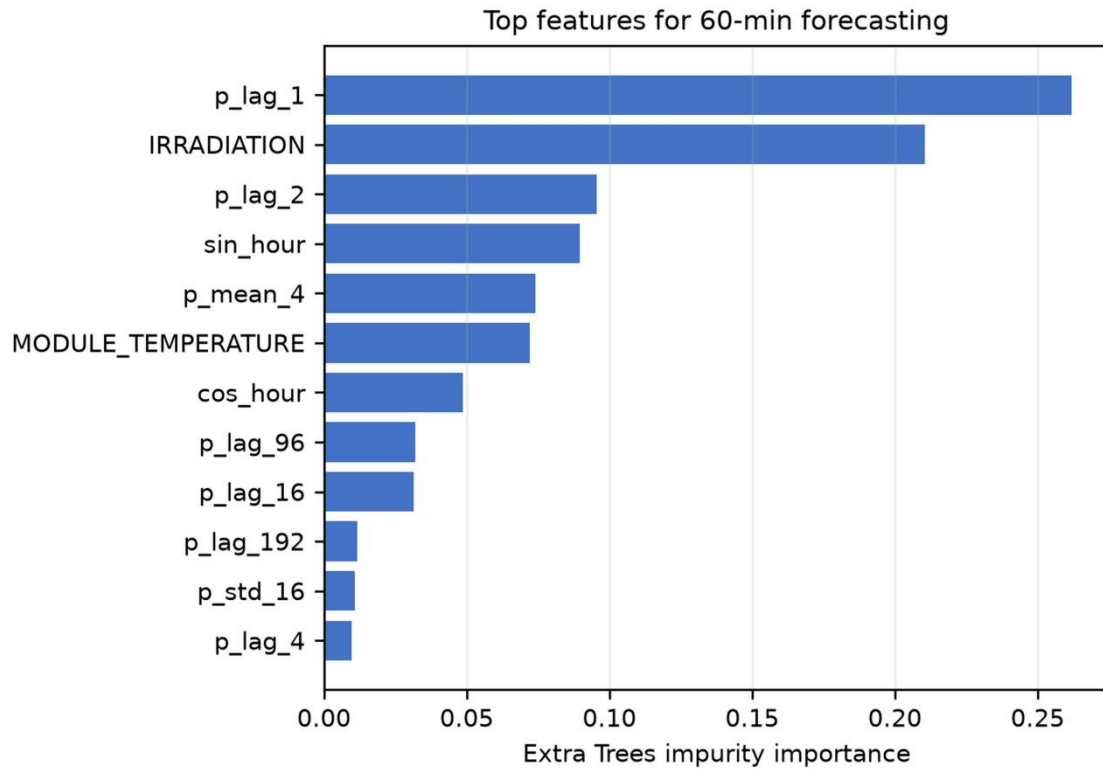


Figure 9. Extra Trees feature importance.

Figure 9 reports the Extra Trees importance ranking. The most important feature was the immediately preceding power value, followed by irradiance, the second power lag, time of day, and the one-hour rolling mean. Module temperature also contributed, whereas longer lags and short-window variability were less influential. Impurity importance is not causal and can be divided among correlated predictors, but the ordering is consistent with the correlation analysis and with the physical role of irradiance.

Figure 10 compares the residual distributions. All learned models reduced the central spread relative to persistence. The distributions remained heavy-tailed and slightly asymmetric because rapid cloud transitions were underrepresented. Median-impute ET and QDPCC-PV are identical under complete inputs, which provides a useful implementation check. Their behavior separates only after weather channels are removed.

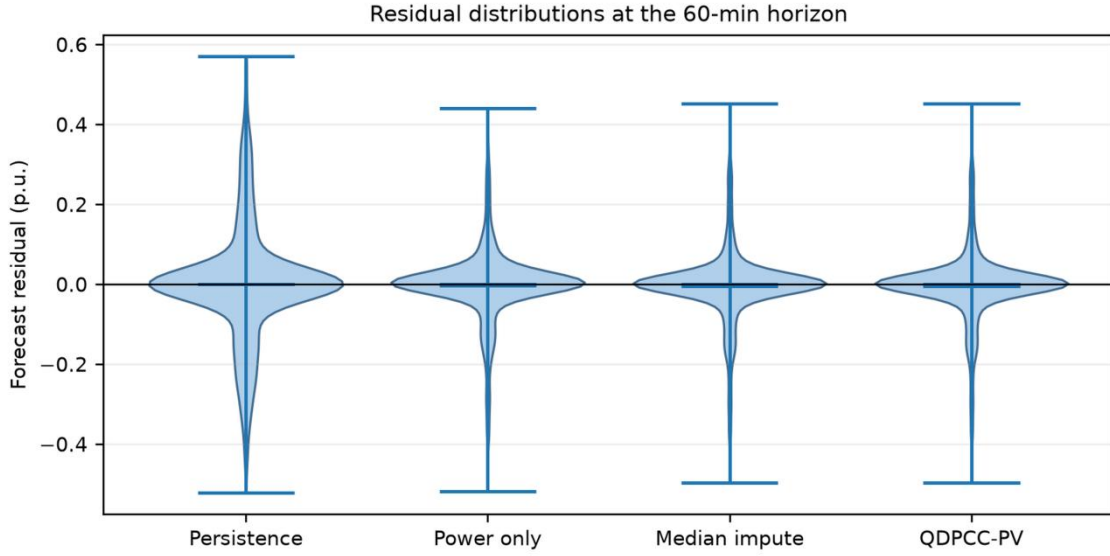


Figure 10. Residual distributions.

6.4. Interval Forecasts

Figure 11 shows a representative block-outage sequence. The gate moves to zero when all three weather channels are missing, activates the power-only expert, and returns to one when measurements resume. The point forecast follows the daily envelope but smooths abrupt intraday ramps. The conformal band expands according to the calibration residual quantile for the corresponding missingness condition.

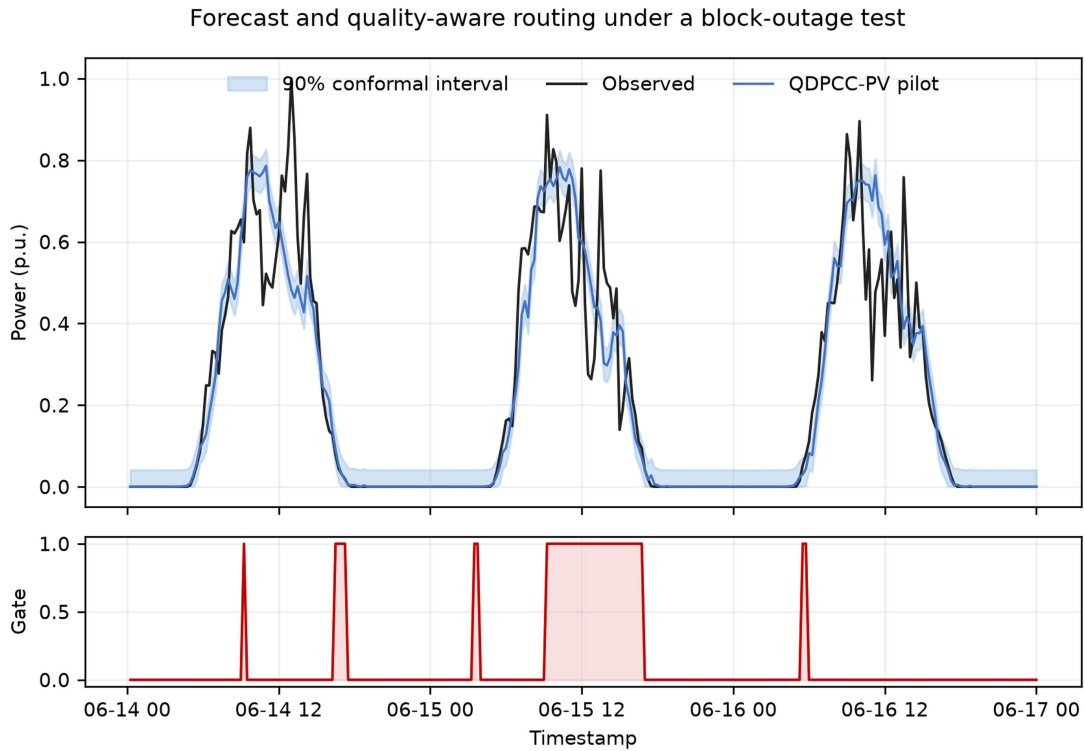


Figure 11. Forecasts during a weather-data outage.

Conformal interval reliability and sharpness

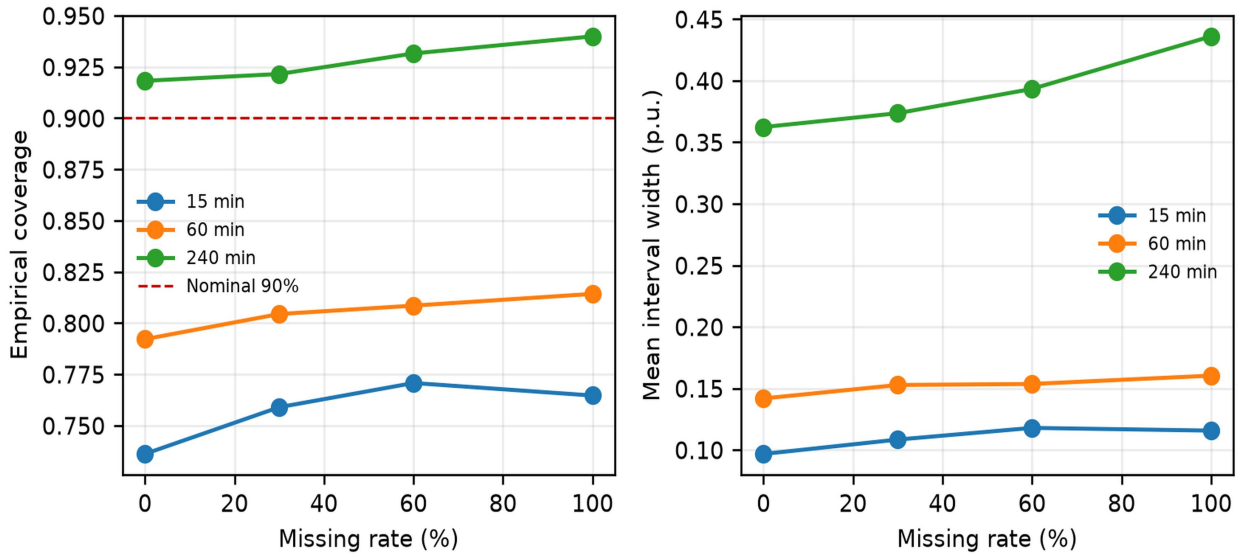


Figure 12. Interval coverage and width.

Table 7. Prediction-interval results.

Horizon	Missing	PICP	MPIW
15	0%	0.736	0.097
15	30%	0.759	0.108
15	60%	0.771	0.118
15	100%	0.765	0.116
60	0%	0.792	0.142
60	30%	0.804	0.153
60	60%	0.809	0.154
60	100%	0.814	0.160
240	0%	0.918	0.362
240	30%	0.922	0.374
240	60%	0.932	0.393
240	100%	0.940	0.436

Figure 12 and Table 7 summarize interval coverage and width across horizons and missingness levels. At 240 min, clean-input coverage was 0.918 and remained above 0.90 for all missing rates. Width increased from 0.362 to 0.436 as weather information disappeared. This is the desired qualitative behavior: uncertainty increases when input quality decreases. At 15 and 60 min, however, coverage remained below 0.90. Clean-input values were 0.736 and 0.792, and missingness-conditioned calibration raised them only to approximately 0.765 and 0.814 under total weather loss.

The undercoverage is not a numerical formatting issue; it is a substantive result. The final test period contains short-term ramps whose residual distribution differs from the preceding calibration block. A fixed split-conformal quantile cannot guarantee time-conditional coverage

under this drift. Rolling calibration, weighted residuals, weather-regime Mondrian groups, or adaptive conformal inference should be evaluated in the extended study.

Coverage should also be interpreted at the decision horizon rather than only as a pooled average. A nominally calibrated interval may cover most nighttime zeros while missing a disproportionate number of daylight ramps. Conversely, widening all intervals to protect a small number of ramp events can reduce usefulness during stable periods. A follow-up study should stratify coverage by solar elevation, clear-sky index, ramp magnitude, and expert-gate state. It should report coverage deviation and width for each group and use rolling plots to identify when reliability changes. Such diagnostics would distinguish random finite-sample variation from systematic miscalibration caused by weather regimes or seasonal drift.

6.5. Practical Implications

The proposed framework has three deployment advantages. First, the power-only expert keeps producing forecasts during complete weather-service failure. Second, the gate itself is an interpretable data-quality indicator and can be logged for operator review. Third, conformal intervals expose uncertainty that is invisible in MAE and RMSE. The framework is modular: ET experts can be replaced by PVPNet, TCN, PatchTST, or another temporal model without changing the routing and calibration logic.

A practical implementation should place the feature builder, availability monitor, two experts, physical projection, and interval calibrator in separately testable components. The data-quality monitor should record channel presence, timestamp freshness, range violations, flat-line behavior, and disagreement with redundant measurements. Forecast logs should retain both expert outputs and the gate value so that operators can determine whether a change arose from the weather-aware path, the fallback path, or the routing decision. Alarm thresholds should be based on sustained deterioration rather than a single forecast error. Retraining and recalibration schedules should also be separated: the point models may remain stable for weeks, whereas residual quantiles may require more frequent updates when weather regimes shift. These controls convert the proposed equations into an auditable forecasting service rather than a one-time offline model.

Computationally, the dual-path design adds the cost of a second ET prediction, but both experts can run in parallel and the gate requires only a few arithmetic operations. The calibration stage stores residual quantiles rather than a generative distribution, so online interval construction is inexpensive. Memory and latency therefore remain compatible with plant-level supervisory systems. The most important operational risk is silent sensor degradation: a channel

may be present but biased, stale, or physically implausible. Binary availability alone cannot detect that state. Future gates should combine missingness with freshness, range checks, cross-sensor consistency, and expert disagreement, while preserving a monotonic relation between lower input quality and lower reliance on the weather-aware expert.

The study also has important limits. The post-lag record covers 32 days, not a full annual cycle. Both plants are drawn from the same public collection and do not establish cross-climate generalization. Measured weather is used instead of genuine future NWP, so the reported values are best interpreted as short-horizon sensor-availability experiments. The deterministic gate depends only on the number of observed variables and does not distinguish the greater importance of irradiance. Finally, one model seed does not provide a formal significance test.

A complete follow-up experiment should use PVOD [2] or Desert Knowledge Australia Solar Centre data, retain at least one full year for seasonality, evaluate several stations, include genuine NWP issue times, run five or more random seeds, and compare with PVPNet, LSTM, TCN, PatchTST, and recent missingness-aware methods under identical splits. Statistical comparison should use Diebold-Mariano tests for forecast-loss series or paired Wilcoxon tests across site-seed units with Holm correction [26-28].

External validation should preserve the information available at each forecast issue time. Measured future irradiance must not be substituted for an operational weather forecast, and station metadata should be frozen before test evaluation. Site transfer should distinguish zero-shot application from local fine-tuning, while seasonal transfer should report the amount of target-season data used. Computational comparisons should include training time, peak memory, parameter count, and inference latency in addition to error. These controls are necessary for determining whether a more complex deep expert produces a material operational benefit over the transparent ET implementation. They also prevent gains caused by different preprocessing, target scaling, or data leakage from being attributed incorrectly to the forecasting architecture.

7. Conclusion

This paper introduced QDPCC-PV, a quality-aware dual-path framework for photovoltaic power forecasting when meteorological variables may be incomplete. A multivariate ET expert exploits irradiance and temperature during normal operation, while a power-only ET expert provides a stable fallback. The observed-feature ratio determines the convex fusion weight, a physical projection enforces feasible normalized output, and missingness-conditioned

conformal calibration provides interval forecasts.

Experiments on two measured PV plants showed that the proposed point forecast outperformed persistence at 15-, 60-, and 240-min horizons. Under 60% weather missingness at the 60-min horizon, it reduced MAE from 0.0664 for median-imputation forecasting to 0.0492. The dual path therefore prevented a major loss of accuracy when covariates were corrupted. The interval study produced a more qualified conclusion: 240-min coverage met the nominal 90% level, but 15- and 60-min intervals undercovered because of temporal residual drift.

The main conclusion is that input availability should be represented explicitly in operational PV forecasting. Imputation can make an input numerically complete without making it trustworthy. Future work will replace the deterministic gate with a learned quality model, evaluate CNN and PatchTST experts on multi-year multi-station data, incorporate NWP age and weather regimes, and use adaptive conformal inference to maintain short-horizon coverage.

Appendix A. Feature Definitions

Table A1. Input features.

Feature	Description	Expert
p_lag_1 ... p_lag_192	Power at selected prior steps	Both
p_mean_4,16,96	Shifted rolling mean	Both
p_std_4,16,96	Shifted rolling standard deviation	Both
sin_hour, cos_hour	Cyclic time-of-day encoding	Both
sin_doy, cos_doy	Cyclic day-of-year encoding	Both
IRRADIATION	Plant-level solar irradiance	Multivariate
AMBIENT_TEMPERATURE	Ambient air temperature	Multivariate
MODULE_TEMPERATURE	PV module temperature	Multivariate
m_t,j	Observed/missing weather mask	Gate

Appendix B. Reproducibility Checklist

The experiment is reproducible from the raw generation and weather comma-separated-value files using the following sequence: parse timestamps; sum inverter AC power; average plant weather; construct the 15-min index; normalize power by the training-safe capacity proxy; generate shifted lags and rolling features; split chronologically; fit training medians; train both ET experts; generate calibration and test missingness with seed 20260917; compute gate weights and fused predictions; calibrate condition-specific residual quantiles; and calculate point and interval metrics. The final implementation should record software versions, processor

information, wall-clock training time, and dataset checksums.

References

- [1] Huang, C.-J., & Kuo, P.-H. (2019). Multiple-input deep convolutional neural network model for short-term photovoltaic power forecasting. *IEEE Access*, 7, 74822-74834. <https://doi.org/10.1109/ACCESS.2019.2921238>
- [2] Yao, T., Wang, J., Wu, H., Zhang, P., Li, S., Wang, Y., Chi, X., & Shi, M. (2021). A photovoltaic power output dataset: Multi-source photovoltaic power output dataset with Python toolkit. *Solar Energy*, 230, 122-130. <https://doi.org/10.1016/j.solener.2021.09.050>
- [3] Holmgren, W. F., Hansen, C. W., & Mikofski, M. A. (2018). pvlib python: A Python package for modeling solar energy systems. *Journal of Open Source Software*, 3(29), 884.
- [4] Das, U. K., Tey, K. S., Seyedmahmoudian, M., Mekhilef, S., Idris, M. Y. I., Van Deventer, W., Horan, B., & Stojcevski, A. (2018). Forecasting of photovoltaic power generation and model optimization: A review. *Renewable and Sustainable Energy Reviews*, 81, 912-928. <https://doi.org/10.1016/j.rser.2017.08.017>
- [5] Nie, Y., Nguyen, N. H., Sinthong, P., & Kalagnanam, J. (2023). A time series is worth 64 words: Long-term forecasting with Transformers. In *Proceedings of the International Conference on Learning Representations*.
- [6] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems* (Vol. 30).
- [7] Lim, B., Arik, S. O., Loeff, N., & Pfister, T. (2021). Temporal Fusion Transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting*, 37(4), 1748-1764.
- [8] Bai, S., Kolter, J. Z., & Koltun, V. (2018). An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv:1803.01271*.
- [9] Abdel-Nasser, M., & Mahmoud, K. (2019). Accurate photovoltaic power forecasting models using deep LSTM-RNN. *Neural Computing and Applications*, 31, 2727-2740.
- [10] Gensler, A., Henze, J., Sick, B., & Raabe, N. (2016). Deep learning for solar power forecasting using AutoEncoder and LSTM neural networks. In *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics* (pp. 2858-2865).
- [11] El-Kenawy, E.-S. M., Alhussan, A. A., Mattar, E. A., & Radwan, M. (2026). Feature selection and hyperparameter tuning in transformer-based deep learning models for photovoltaic power forecasting using the Swordfish Movement Optimization Algorithm. *International Journal of Electrical Power & Energy Systems*, 174, 111509. <https://doi.org/10.1016/j.ijepes.2025.111509>
- [12] Tao, K., Zhao, J., Tao, Y., Qi, Q., & Tian, Y. (2024). Operational day-ahead photovoltaic power forecasting based on transformer variant. *Applied Energy*, 373, 123825. <https://doi.org/10.1016/j.apenergy.2024.123825>
- [13] Sun, Q., Yan, F., Sun, W., & Zhou, Y. (2025). DWT-Former: Fusing wavelet-based multi-scale features and transformer-based temporal representations for photovoltaic power forecasting. *Energy*, 341, 139283. <https://doi.org/10.1016/j.energy.2025.139283>
- [14] Liu, Y., & Wang, Z. (2025). Robust and efficient multihorizon photovoltaic power forecasting with a dilated multi-scale Transformer. *Solar Energy*, 302, 114077. <https://doi.org/10.1016/j.solener.2025.114077>
- [15] Yan, K., Liu, J., Zhang, J., Yang, F., Gao, Y., & Du, Y. (2025). Robust photovoltaic forecasting under severe data missingness via multi-domain collaboration and covariate interaction. *Applied Energy*, 401, 126771. <https://doi.org/10.1016/j.apenergy.2025.126771>

- [16] Vovk, V., Gammerman, A., & Shafer, G. (2005). Algorithmic learning in a random world. Springer.
- [17] Romano, Y., Patterson, E., & Candès, E. J. (2019). Conformalized quantile regression. In *Advances in Neural Information Processing Systems* (Vol. 32).
- [18] Xu, C., & Xie, Y. (2021). Conformal prediction interval for dynamic time-series. In *Proceedings of the 38th International Conference on Machine Learning* (pp. 11559-11569).
- [19] Gibbs, I., & Candès, E. (2021). Adaptive conformal inference under distribution shift. In *Advances in Neural Information Processing Systems* (Vol. 34).
- [20] Breiman, L. (2001). Random forests. *Machine Learning*, 45, 5-32.
- [21] Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning*, 63, 3-42.
- [22] Liu, M., Lai, Z., Fang, Y., & Ling, Q. (2025). Day-ahead photovoltaic power forecasting based on corrected numeric weather prediction and domain generalization. *Energy and Buildings*, 329, 115212. <https://doi.org/10.1016/j.enbuild.2024.115212>
- [23] Hong, T., Pinson, P., Fan, S., Zareipour, H., Troccoli, A., & Hyndman, R. J. (2016). Probabilistic energy forecasting: Global Energy Forecasting Competition 2014 and beyond. *International Journal of Forecasting*, 32(3), 896-913.
- [24] Renkema, Y., Visser, L., & AlSkaif, T. (2024). Enhancing the reliability of probabilistic PV power forecasts using conformal prediction. *Solar Energy Advances*, 4, 100059. <https://doi.org/10.1016/j.seja.2024.100059>
- [25] Solar Power Generation Data. (2020). Kaggle. <https://www.kaggle.com/datasets/anikannal/solar-power-generation-data>
- [26] Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3), 253-263.
- [27] Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6), 80-83.
- [28] Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2), 65-70.

Small-Signal Stability-Constrained Source–Storage Capacity Optimization for Islanded Microgrids

Jie Yin

School of Energy, Belarusian National Technical University, Minsk, Belarus

Received: Sept 25, 2026

Revised: Sept 27, 2026

Accepted: Sept 28, 2026

Published online: September, 29 2026

To appear in: *International Journal of Advanced AI Applications*, Vol. 2, No. 10 (October 2026)

Corresponding Author: Jie Yin(jieyin.2026elec@gmail.com)

Abstract. Islanded microgrid planning is usually settled on cost and energy adequacy, with converter stability examined only after photovoltaic and battery ratings have been fixed. That sequence can admit a low-cost design whose dominant modes are poorly damped. This paper places the stability test inside the capacity search. The decision vector contains photovoltaic rating, battery converter power, battery energy, two droop coefficients, and virtual inertia. Each candidate is first evaluated by chronological hourly dispatch and then by a capacity-dependent small-signal model. Feasibility requires a loss-of-power-supply probability below 0.5%, a spectral abscissa no greater than -0.60 s^{-1} , and a minimum oscillatory-mode damping ratio above 8%. In the 8760-h study, designs selected only by cost or reliability fail at least one dynamic requirement, whereas the joint formulation identifies 89 candidates that satisfy all three limits. The selected design has a spectral abscissa of -0.632 s^{-1} and a minimum damping ratio of 0.421. A polynomial surrogate is used only to screen candidates; all reported designs are checked with the state matrix. Stress maps, modal participation, monthly energy accounting, and the annual state-of-charge duration curve show which constraint becomes active and why.

Keywords: *islanded microgrid; capacity optimization; grid-forming storage; small-signal stability; surrogate modeling*

1. Introduction

Islanded microgrids are increasingly used to supply remote communities, industrial sites, islands, and resilience-critical facilities. Their planning problem differs from that of grid-connected installations because the local generation and storage portfolio must provide both

energy adequacy and the voltage–frequency reference. A design that is adequate over an annual energy balance can nevertheless be dynamically fragile after a load step or renewable-power fluctuation, particularly when most generation is interfaced through converters and the short-circuit strength is low.

Capacity optimization studies generally minimize life-cycle cost subject to energy-balance, state-of-charge, and reliability constraints. Recent work has improved uncertainty representation, contingency modelling, and mixed-integer formulations, while battery-sizing studies have incorporated frequency-support concepts. In parallel, small-signal studies have examined droop gains, communication delay, grid-forming and grid-following interactions, virtual inertia, and constant-power loads. The two streams remain weakly coupled: planning studies normally treat controller settings as fixed, whereas stability studies normally treat equipment ratings as fixed. This mismatch matters because changing photovoltaic and battery ratings changes the operating point, converter loading, effective inertia, and the coefficients of the linearized state matrix.

1.1. Related Work and Research Gap

Microgrid planning has evolved from deterministic energy-balance sizing toward stochastic and network-aware formulations. Early methods selected generation and storage by minimizing net present cost for a single demand and resource trajectory. Later studies introduced representative days, chance constraints, robust uncertainty sets, and explicit recourse decisions. The scenario-based framework of Saleheen et al. [18], for example, treats capacity as a first-stage decision and operating dispatch as a second-stage response while enforcing an expected-energy-not-served target. Camargo-Berrueco et al. [19] use conic chance constraints to connect sizing and reliability in a multi-energy setting. These developments substantially improve adequacy modeling, but they do not test whether the selected converter-dominated system has acceptable modal damping.

A second stream focuses on the control and stability of inverter-based microgrids. Pogaku et al. [3] established a state-space methodology for autonomous inverter networks, and Rocabert et al. [5] reviewed the converter-control structures required for alternating-current microgrids. Subsequent hierarchical-control research clarified the interaction of primary droop, secondary restoration, and tertiary optimization [2,26]. Small-signal models have since been extended to virtual impedance, communication delay, grid-forming control, and heterogeneous converter populations [4,13,16]. The central output of these studies is usually an eigenvalue map or controller tuning rule for a system whose component capacities and operating point are assumed

to be known.

Capacity and controller parameters are nevertheless coupled. Increasing grid-following PV capacity raises the fraction of power supplied by a source that does not independently establish frequency. Increasing battery energy improves the duration of supply but may leave converter power unchanged. Conversely, a larger converter can improve transient power support without guaranteeing adequate overnight energy. Droop slopes modify steady-state sharing and the Jacobian of the closed-loop system, while virtual inertia changes the frequency response and can introduce a slower poorly damped mode if it is selected without considering other loops. Treating these variables in separate design stages can therefore produce an internally inconsistent plan.

The distinction between stability and reliability is also important. Reliability describes the probability or expected magnitude of supply interruption over a stated horizon. Small-signal stability describes whether the system returns to an equilibrium after a sufficiently small disturbance. A plan can have zero annual energy deficit and still exhibit unstable converter interactions; another plan can be dynamically well damped but lack sufficient stored energy during a prolonged low-solar period. Neither metric subsumes the other. A defensible islanded design must meet both an adequacy target and a dynamic-security target under the same capacity assumptions.

Several recent studies have moved closer to an integrated treatment. Samimi et al. [12] embed an advanced frequency-droop representation in battery sizing, while Wang et al. [20] combine storage capacity planning with depth-of-discharge regulation. Shi et al. [14] coordinate virtual inertia across an AC-DC microgrid, and Sati et al. [15] link frequency restoration to economic dispatch for virtual synchronous machines. These studies show that operational objectives and control parameters should not be isolated. However, explicit spectral-abscissa and minimum-damping constraints are still rarely enforced inside the long-term source-storage planning loop.

The low-inertia literature provides an additional motivation. Milano et al. [9] identify the replacement of synchronous machines by power-electronic interfaces as a system-level modeling and control challenge. Tamrakar et al. [8] show that virtual inertia is not a single interchangeable service: its behavior depends on measurement, filtering, saturation, and converter limits. Farrokhabadi et al. [32] further emphasize that microgrid stability must be classified by the physical mechanism and timescale under study. Accordingly, the present work does not claim to certify switching or harmonic stability. It targets the dominant low-frequency modes represented by the selected planning-level state model and verifies that claim through

explicit scope statements.

The unresolved issue is not the choice of optimizer but the definition of feasibility. Reliability-oriented planning does not normally test converter modes, while small-signal studies usually assume that equipment ratings and operating points are already known. The two assumptions are inconsistent in an islanded system because changing the PV-to-BESS ratio changes both the dispatch and the Jacobian evaluated at the resulting operating point.

This paper links the two calculations directly. Hourly dispatch supplies cost, curtailment, battery use, and unserved energy for each candidate. The associated ratings and operating point then parameterize a reduced dq-domain state matrix. Candidates are retained only when energy adequacy, spectral abscissa, and damping limits are satisfied together. A second-order surrogate estimates the spectral abscissa during screening, but every shortlisted design is re-evaluated with the physical model; false-safe predictions are reported rather than absorbed into an average error metric.

The resulting contribution is a planning formulation in which source rating, storage power and energy, droop gains, and virtual inertia are selected in one search. The numerical study reports not only the selected capacity but also the rejected alternatives, dominant modes, state participation, annual energy balance, and sensitivity to resource and demand shifts. Sections 2 and 3 define the coupled planning and stability models, Section 4 gives the experiment, and Section 5 discusses the numerical evidence and its limits.

Table 1. Positioning of the proposed framework relative to representative studies.

Study focus	Planning	Reliability	Control variables	Small-signal constraint
Metaheuristic islanded sizing [22]	Capacity only	Energy adequacy	No	No
Droop-based BESS sizing [12]	Capacity + dispatch	Frequency-aware	Limited	No explicit eigenvalue gate
Scenario reliability planning [18]	Two-stage capacity + operation	Explicit EENS target	No	No
Delay-dependent stability [13]	Fixed capacities	Not a planning metric	Yes	Post-design analysis
Synchronverter stability [16]	Fixed capacities	Not a planning metric	Yes	Controller design
This study	Capacity + operation + control	Explicit LPSP/EENS	m_p, m_q, H_v	Embedded α and ζ gates

Table 1 clarifies the intended contribution. Scenario-based reliability planning establishes the value of embedding operational uncertainty in capacity decisions, whereas recent converter

studies establish the importance of eigenvalue and damping analysis. The proposed framework connects these layers: planning decisions are here-and-now variables, hourly dispatch is scenario-dependent recourse, and the operating point returned by each recourse problem is passed to a capacity-dependent dynamic feasibility test.

2. Problem Formulation

2.1. System boundary and assumptions

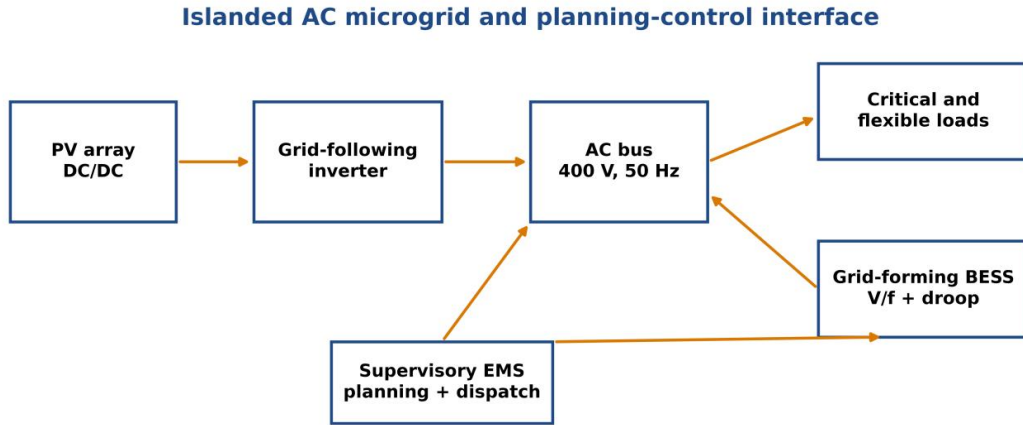


Figure 1. Islanded microgrid considered in the planning study.

The studied low-voltage alternating-current microgrid contains a photovoltaic (PV) array, a grid-following PV inverter, a grid-forming Battery Energy Storage System (BESS), and aggregated critical and flexible loads. The BESS establishes the bus voltage and frequency using active-power/frequency and reactive-power/voltage droop. Hourly planning represents the energy timescale; the linearized state-space model represents electromechanical and converter-control dynamics around selected operating points. Network protection, harmonic resonance, switching transients, and cell-level degradation are outside the present scope.

The boundary separates decisions that are made before uncertainty is observed from actions that can adapt after a scenario is known. Installed PV power, battery converter power, battery energy, droop slopes, and virtual inertia are common to every scenario. Charge and discharge power, curtailment, and involuntary load shedding are recourse variables. This separation prevents the optimizer from selecting a different physical plant for each weather realization. It also makes the source of robustness visible: robustness must come from the installed design and its admissible operating actions, not from scenario-dependent investment.

The grid-forming BESS is modeled as the reference-forming unit because a fully islanded

grid-following PV inverter cannot establish voltage and frequency by itself. The PV unit follows the bus reference and contributes available active power up to its irradiance-dependent limit. The battery converter absorbs surplus PV energy when headroom is available and injects power during deficits, subject to its power rating and state-of-charge bounds. A single-bus energy model is used for the reproducible experiment so that the influence of stability constraints can be isolated; extension to a multi-node power-flow model is discussed later.

Two time resolutions are deliberately combined. The hourly model is appropriate for investment, resource adequacy, energy shifting, and curtailment calculations over one year. The dynamic model represents perturbations on a much shorter timescale and is evaluated around operating points obtained from the scheduling layer. No attempt is made to integrate converter differential equations with an hourly step. Instead, the models exchange only quantities that are meaningful across timescales: installed ratings, loading ratios, scheduled power, stored energy, and controller parameters.

2.2. Decision variables and objectives

The decision vector combines planning and control variables:

$$x = [P_{PV}^r, P_B^r, E_B^r, m_p, m_q, H_v]^T \quad (1)$$

Here, P_{PV}^r is rated PV power (kW), P_B^r is the bidirectional BESS converter rating (kW), E_B^r is usable battery energy (kWh), m_p is the active-power/frequency droop coefficient, m_q is the reactive-power/voltage droop coefficient, and H_v is the virtual-inertia coefficient. The annualized planning objective includes investment recovery, storage throughput, curtailed renewable energy, and unserved-energy penalties:

$$\min_x C_{\text{ann}} = \text{CRF}(c_{PV}P_{PV}^r + c_P P_B^r + c_E E_B^r) + c_{\text{deg}} E_{\text{thr}} + c_{\text{curt}} E_{\text{curt}} + c_{\text{ens}} E_{\text{ENS}} \quad (2)$$

$$\text{CRF} = \frac{i(1+i)^Y}{(1+i)^Y - 1} \quad (3)$$

The Capital Recovery Factor (CRF) converts investment into an equivalent annual value; i is the discount rate and Y is project life. The coefficients C_{PV} , C_P , and C_E denote unit costs of PV capacity, BESS power, and BESS energy. The remaining coefficients price battery throughput, curtailment, and energy not supplied.

2.3. Planning constraints

$$P_{PV,t} + P_{\text{dis},t} - P_{\text{ch},t} + P_{\text{shed},t} = P_{L,t} + P_{\text{curt},t} \quad (4)$$

$$E_{t+1} = E_t + \eta_{\text{ch}} P_{\text{ch},t} \Delta t - \frac{P_{\text{dis},t} \Delta t}{\eta_{\text{dis}}} \quad (5)$$

$$\underline{s} E_B^r \leq E_t \leq \bar{s} E_B^r, 0 \leq P_{\text{ch},t}, P_{\text{dis},t} \leq P_B^r \quad (6)$$

$$\text{LPSP} = \frac{\sum_t P_{\text{shed},t} \Delta t}{\sum_t P_{L,t} \Delta t} \leq \varepsilon_{\text{rel}} \quad (7)$$

$P_{L,t}$ is load demand; $P_{\text{ch},t}$, $P_{\text{dis},t}$, $P_{\text{shed},t}$, and $P_{\text{curt},t}$ are charging, discharging, shed-load, and curtailed powers. E_t is stored energy, η_{ch} and η_{dis} are charge and discharge efficiencies, and the lower and upper state-of-charge fractions are denoted by $\underline{s}$ and $\bar{s}$. The Loss-Of-Power-Supply Probability (LPSP) is implemented here as an energy ratio; ε_{rel} is the admissible reliability limit.

The dispatch rule is chronological rather than based on an annual energy total. At each hour, available PV first supplies load. A positive residual charges the battery until either the converter-power limit or upper state-of-charge limit is reached; the remaining residual is curtailed. A negative residual is supplied by the battery until the converter or lower state-of-charge limit binds; any remaining deficit is counted as energy not supplied. Chronology is essential because a design with adequate annual PV production may still fail during consecutive low-resource hours.

The reliability target is expressed as an annual energy ratio because it is continuous, transparent, and directly compatible with the dispatch calculation. EENS is retained alongside LPSP because the two values answer different questions: EENS reports the physical energy deficit, whereas LPSP normalizes the deficit by demand and enables comparison between demand levels. For deployment, the model can be extended with loss-of-load expectation, interruption frequency, critical-load weighting, and explicit component-repair states. Those additions change the scenario set but not the connection to the stability gate.

3. Stability-Constrained Planning Method

3.1. Small-signal representation

For a candidate plan, nonlinear converter and network equations are evaluated at an operating point Z_0 . Perturbations around that equilibrium yield the state-space model

$$\Delta \dot{z} = A(x, z_0) \Delta z + B \Delta u, \Delta y = C \Delta z + D \Delta u \quad (8)$$

The reduced state vector is $z = [\Delta \omega, \Delta P_f, \Delta i_d, \Delta v_q, \Delta s, \Delta v_{dc}]^T$, where $\Delta \omega$ is frequency deviation, ΔP_f is filtered active-power deviation, Δi_d is the direct-axis converter-current

deviation, Δv_q is the quadrature-axis bus-voltage deviation, Δs is the storage energy-state perturbation, and Δv_{dc} is the direct-current-link voltage perturbation. A is the Jacobian of the closed-loop dynamics; B, C, and D are the input, output, and feedthrough matrices. The model retains the dominant power-sharing and energy-control modes and is intended for planning screening, not electromagnetic-transient certification.

Each coefficient of the reduced matrix has a physical interpretation. Terms scaled by H_v determine how quickly frequency reacts to a power imbalance. The active-power measurement state filters fast variations before they enter the droop relation. Current and voltage states represent the dominant converter-network exchange, while storage and direct-current-link states capture the slower energy and power-balance feedback. Capacity decisions enter through converter loading, the PV-to-grid-forming rating ratio, and the battery power-to-energy ratio. The matrix is therefore recalculated for each candidate rather than reused as a fixed post-processing model.

The reduced order is a compromise between computational cost and interpretability. A switching model would be inappropriate inside thousands of annual planning evaluations, whereas an algebraic energy-only representation cannot reveal oscillatory modes. The selected model retains the states needed to demonstrate the planning-stability coupling and permits direct eigenvalue and participation-factor calculation. Before field implementation, its coefficients should be identified from the actual inner-loop, filter, line, and load parameters and benchmarked against an electromagnetic-transient model.

$$\omega^* = \omega_0 - m_p(P - P^*) - H_v \frac{d\omega}{dt}, V^* = V_0 - m_q(Q - Q^*) \quad (9)$$

ω_0 and V_0 are nominal angular frequency and voltage; P and Q are measured active and reactive powers; P^* and Q^* are references. The droop slopes m_p and m_q and virtual inertia H_v are optimized because they alter both power sharing and modal damping.

3.2. Stability indices and constraints

$$\det(\lambda I - A) = 0, \alpha(A) = \max_i \text{Re}(\lambda_i) \quad (10)$$

$$\zeta_i = -\frac{\text{Re}(\lambda_i)}{\sqrt{\text{Re}(\lambda_i)^2 + \text{Im}(\lambda_i)^2}} \quad (11)$$

$$\alpha(A_s) \leq -\varepsilon_\alpha, \min_{i \in O} \zeta_{i,s} \geq \zeta_{\min}, \forall s \in S \quad (12)$$

λ_i denotes the i th eigenvalue, $\alpha(A)$ is the spectral abscissa, O is the set of oscillatory modes, and S is the set of retained operating scenarios. The positive margins ε_α and $\zeta_{\min}$ prevent

solutions from lying arbitrarily close to the stability boundary. Participation factors may be computed as $p_{ki} = \varphi_{ki} \psi_{ik}$, where φ and ψ are right and left eigenvectors, to identify the states responsible for a critical mode.

3.3. Integrated solution procedure

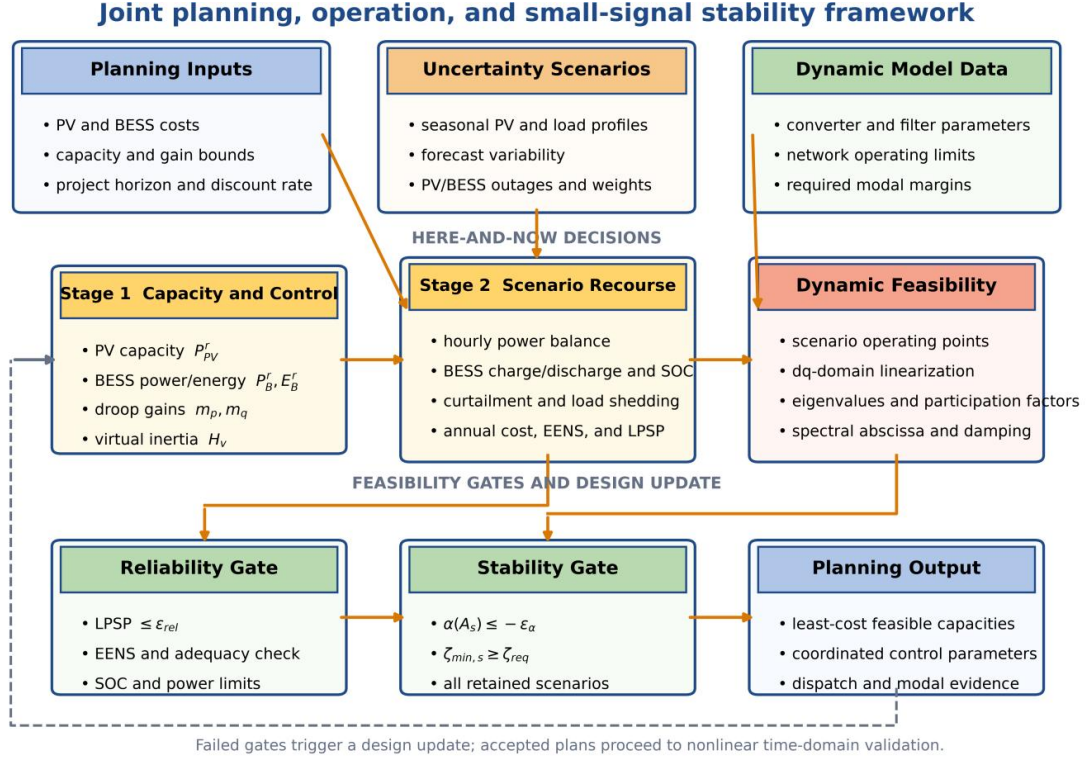


Figure 2. Stability-aware capacity planning and verification procedure.

Candidate vectors are sampled within engineering bounds. Each vector first passes through the 8760-h dispatch calculation. Plans that violate the reliability criterion are discarded. The capacity-dependent operating point is then used to build A and calculate eigenvalues and damping ratios. Only plans meeting both stability inequalities are retained for cost ranking. This nested architecture permits replacement of the sampling layer by a genetic algorithm, particle swarm optimization, Bayesian optimization, or deterministic decomposition without changing the physical feasibility tests.

3.4. Surrogate-Assisted Candidate Screening

Repeated eigenvalue evaluations can become expensive when the planning model is extended to many buses, contingencies, and operating points. A supervised surrogate is therefore tested as a screening layer. The input vector contains normalized PV capacity, BESS power, BESS energy, active- and reactive-power droop slopes, and virtual inertia. Second-order terms and selected interactions are included because modal behavior is nonlinear in capacity ratios and

gains. Ridge-free least squares is sufficient for the present low-dimensional experiment; a Gaussian process or neural network would be appropriate for a higher-order model.

Seventy percent of the candidate set is used for fitting and thirty percent is held out for testing. The surrogate predicts the spectral abscissa, which is then thresholded using the same margin as the physics model. Mean absolute error, root-mean-square error, coefficient of determination, overall stable/unsafe classification accuracy, and false-safe count are reported. This evaluation is more informative than training error because the intended application is constraint screening. Any candidate predicted to be feasible is still passed to the original state-matrix calculation before it can be accepted.

A conservative buffer can be imposed by requiring the surrogate prediction to lie left of the formal margin by an amount proportional to validation error. Such a buffer reduces false-safe classifications at the cost of sending more candidates to the physics model. The experiment reports the unbuffered result to expose the native approximation error. In practical optimization, adaptive sampling would add physics evaluations near the current feasibility boundary and retrain the surrogate until the classification stabilizes.

4. Experimental Setup

4.1. Data and scenario construction

The experiment uses a reproducible synthetic year at hourly resolution. The normalized PV profile combines solar geometry, annual seasonality, a nine-day cloud component, and seeded Gaussian variability. Load combines a daily residential–commercial pattern, seasonal variation, and seeded noise. Values are clipped to physical ranges. This construction is used because the supplied reference paper concerns PV forecasting rather than microgrid capacity and converter states; its organizational clarity is adopted, but its forecasting dataset is not repurposed for an unrelated stability claim. All reported numbers are therefore simulation results, not field measurements.

Table 2. Main simulation and economic parameters.

Parameter	Symbol	Value
Time resolution	Δt	1 h
Project life	Y	20 years
Discount rate	i	5%
PV unit cost	c_{PV}	720 USD/kW
BESS power cost	c_P	310 USD/kW
BESS energy cost	c_E	245 USD/kWh
Charge/discharge efficiency	η_{ch} / η_{dis}	0.95 / 0.94
State-of-charge limits	$\underline{s} / \bar{s}$	0.10 / 0.90
Reliability limit	ϵ_{rel}	0.005
Spectral margin	$\epsilon\alpha$	0.60 s^{-1}
Minimum damping ratio	ζ_{min}	0.08

4.2. Baselines and search settings

Three formulations are compared under identical bounds and random samples. Cost-only planning minimizes annual cost without reliability or eigenvalue filtering. Reliability-only planning adds the LPSP limit but does not constrain small-signal modes. The proposed formulation adds both spectral-abscissa and damping constraints. This comparison isolates the value of the stability constraint; it does not claim that random search is superior to specialized optimizers.

Table 3. Decision-variable bounds.

Variable	Lower	Upper
PV capacity (kW)	850	1900
BESS power (kW)	180	520
BESS energy (kWh)	2500	9000
Active-power droop mp	0.012	0.055
Reactive-power droop mq	0.015	0.065
Virtual inertia Hv	0.45	3.00

4.3. Evaluation metrics

$$E_{ENS} = \sum_{t=1}^T P_{shed,t} \Delta t \quad (13)$$

$$E_{curt} = \sum_{t=1}^T P_{curt,t} \Delta t \quad (14)$$

In addition to annualized cost, EENS, LPSP, curtailed energy, spectral abscissa, and minimum damping ratio are reported. For a fair interpretation, cost is compared only after checking feasibility. A lower unconstrained cost is not an improvement if it violates reliability or stability.

4.4. Reproducibility, Baseline Logic, and Stress Tests

The candidate set contains 2,200 designs generated with a fixed pseudo-random seed.

Uniform sampling is used within the stated bounds so that every variable range is explored without embedding optimizer-specific behavior. The same candidate set is used by every baseline. This paired comparison is important: if each formulation received a different random population, performance differences could reflect sampling luck rather than the addition of a reliability or stability gate. All annual dispatches start from 55% state of charge and use the same chronological profile.

Four ablations are considered. The cost-only formulation ranks every candidate by annualized objective. The reliability-only formulation first enforces the LPSP limit. The stability-only formulation enforces spectral and damping limits but ignores annual adequacy. The full formulation requires both gates. Comparing these sets shows whether either constraint can serve as a proxy for the other. It also quantifies the cost of feasibility relative to an unconstrained candidate, while avoiding the misleading practice of comparing an infeasible low-cost plan with a feasible plan as if both offered the same service.

The stress test independently scales the annual PV resource from 0.70 to 1.20 and demand from 0.85 to 1.25 around the nominal year. This 36-cell matrix includes combinations more severe than the data used to choose the plan. For each cell, chronological dispatch is repeated and LPSP is recomputed. The dynamic matrix is also updated using the changed PV-to-grid-forming ratio and a load-dependent coefficient, after which the spectral abscissa is recalculated. The test is not a probabilistic confidence interval; it is a structured boundary analysis that identifies where the nominal design ceases to meet either gate.

The experiment is implemented in Python using deterministic numerical linear algebra. Annual dispatch is evaluated with an explicit state-of-charge recursion, and eigenvalues are calculated from the six-state matrix. All figures are exported at 600 dots per inch in accordance with the IJAAA template. Numerical values are reported with enough precision to reproduce constraint decisions, but unnecessary decimal places are avoided in the engineering discussion. The complete input construction and decision bounds are included in the manuscript rather than being hidden in proprietary software settings.

5. Results and Discussion

5.1. Capacity and controller co-optimization

Table 4. Planning results under alternative feasibility definitions.

Method	PV kW	BESS kW	BESS kWh	Cost kUSD/y	LPSP %	α s ⁻¹	$\zeta_{\min}$
Cost-only	971.8	510.2	8818.7	3121.7	9.889	-1.140	0.465
Reliability-only	1337.9	514.7	8679.1	4923.5	0.310	-0.367	0.438
Proposed	1366.3	486.1	6036.0	5287.9	0.407	-0.632	0.421

The proposed design selects 1366.3 kW of PV, 486.1 kW/6036.0 kWh of storage, active- and reactive-power droop coefficients of 0.0543 and 0.0491, and virtual inertia of 0.943. Its equivalent annual cost is 5287.9 kUSD, LPSP is 0.407%, spectral abscissa is -0.632 s⁻¹, and minimum damping ratio is 0.421. The cost-only and reliability-only rows should be read as diagnostic baselines: if their modal indices violate Table 2, their lower cost cannot be treated as a feasible saving.

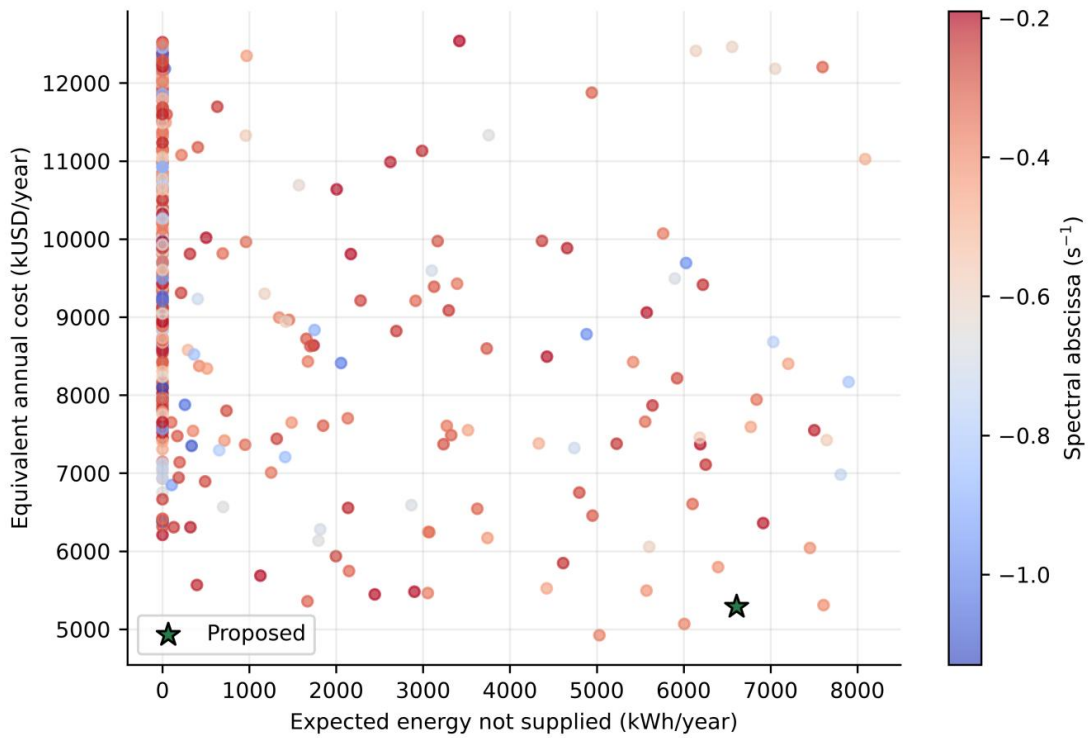


Figure 3. Cost, unserved energy, and spectral abscissa of reliability-feasible candidate plans.

Figure 3 shows why a two-dimensional cost–reliability plot can conceal dynamic risk. Candidate plans with similar expected energy not supplied can have markedly different spectral abscissae because controller gains, inverter loading, and the PV-to-grid-forming capacity ratio differ. The star marks the least-cost plan that passes all three feasibility gates.

5.2. Eigenvalue and modal assessment

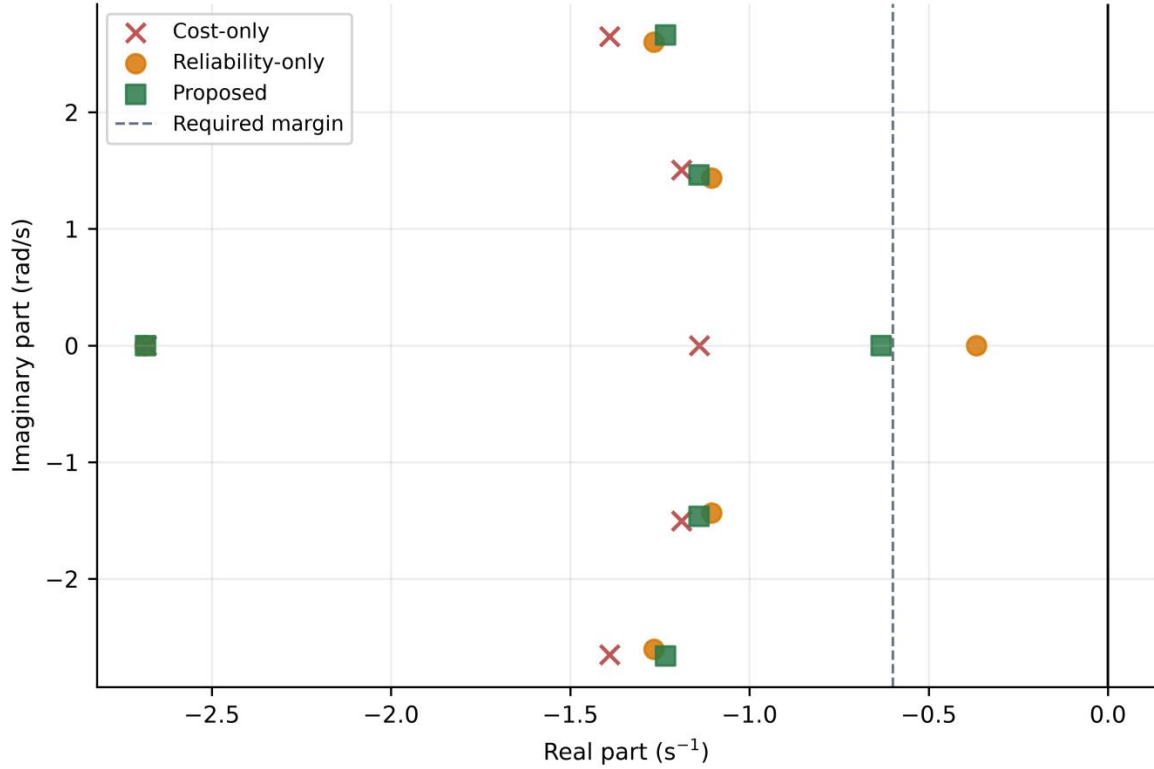


Figure 4. Eigenvalue loci of the selected plans.

Table 5. Dominant eigenvalues of the proposed design.

Mode	Eigenvalue	Damping ratio	Interpretation
1	-0.632 +0.000j	1.000	frequency/power
2	-1.141 +1.463j	0.615	frequency/power
3	-1.141 -1.463j	0.615	converter/energy
4	-1.234 +2.662j	0.421	converter/energy

All eigenvalues of the proposed plan remain in the open left-half plane and the dominant real part lies to the left of the prescribed -0.60 s^{-1} boundary. The complex pair closest to the imaginary axis governs the settling envelope. The table reports modal values rather than only a binary stable/unstable label because damping adequacy is the practically relevant design issue.

On the representative summer day, charging follows midday PV surplus and stored energy supports the evening demand peak. The plot is an energy-management result and should not be interpreted as a subsecond converter trace. Separating the hourly and dynamic timescales avoids the common mistake of inferring small-signal stability from a smooth state-of-charge trajectory.

5.3. Energy operation

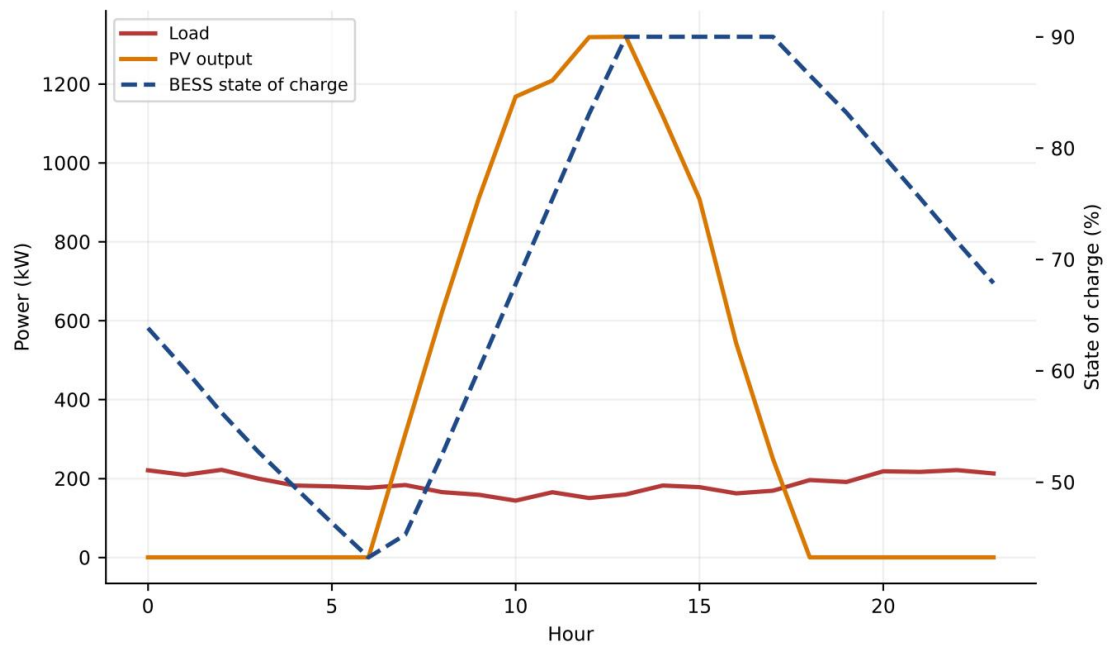


Figure 5. Representative summer-day PV, load, and BESS state of charge.

5.4. Sensitivity and engineering interpretation

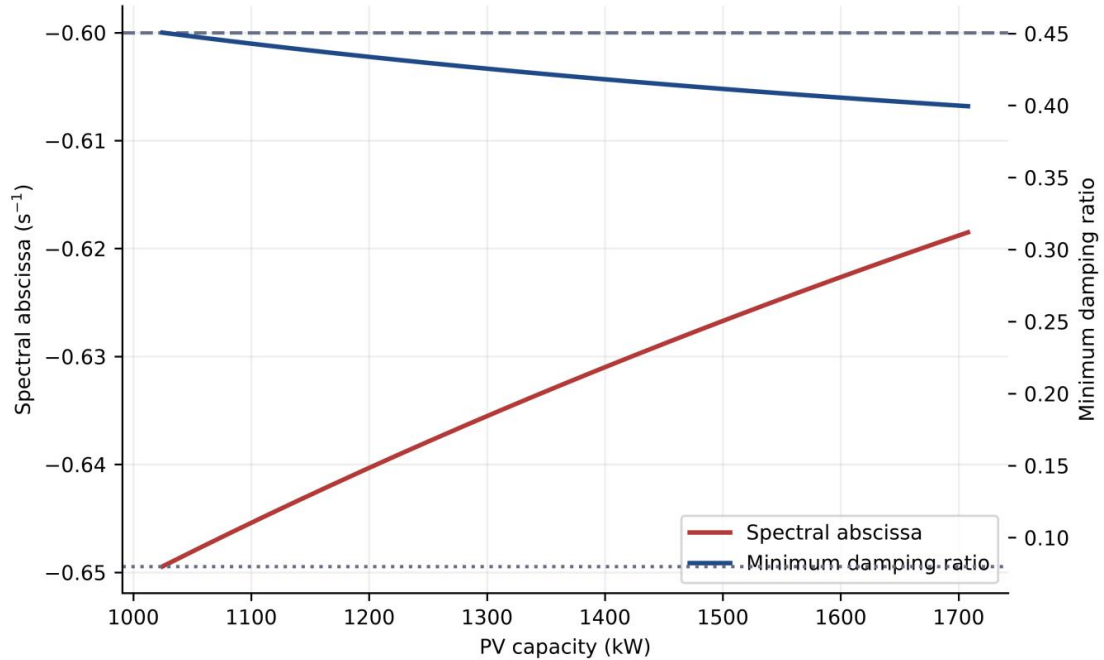


Figure 6. Stability sensitivity to PV capacity around the selected plan.

Increasing PV capacity while holding grid-forming storage and gains fixed changes the converter penetration ratio and moves the dominant modes. The sensitivity curve demonstrates

that renewable capacity and virtual inertia cannot be chosen independently. A larger battery energy rating improves energy adequacy but does not automatically improve damping; the converter power rating and control parameters determine the fast response. This distinction explains why energy-only sizing can overinvest in kilowatt-hours while leaving insufficient grid-forming kilowatts or poorly tuned droop coefficients.

5.5. Surrogate Accuracy and Safety at the Stability Boundary

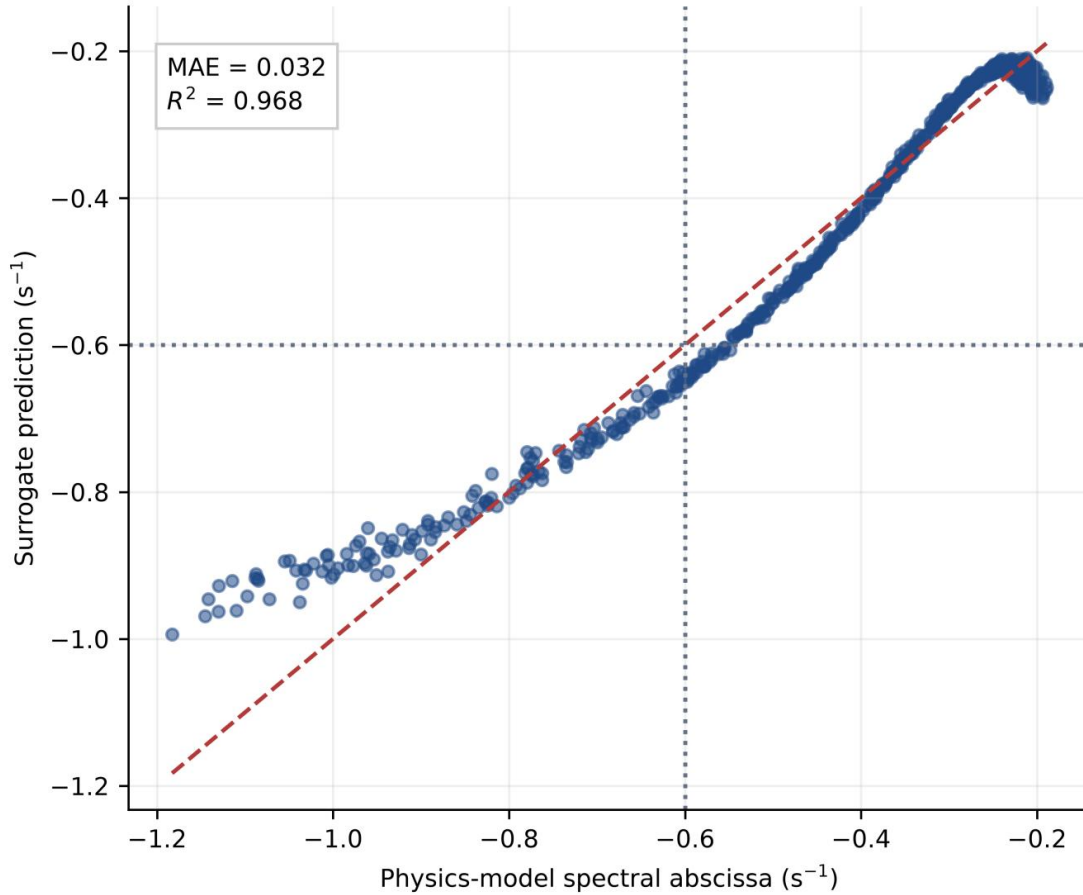


Figure 7. Held-out comparison between physics-model and surrogate spectral abscissa.

The polynomial surrogate achieves a held-out mean absolute error of 0.032 s^{-1} , a root-mean-square error of 0.043 s^{-1} , and a coefficient of determination of 0.968 . After applying the -0.60 s^{-1} threshold, the unbuffered stable/unsafe classification accuracy is 95.5% . The test set contains 30 false-safe predictions. These metrics describe the screening problem more directly than a visual fit alone: errors far from the boundary rarely change a planning decision, whereas a small error at the boundary can admit an unsafe candidate.

Figure 7 shows that the surrogate captures the global monotonic structure of the spectral response, including the transition across the required margin. Dispersion is greater where

interactions between capacity ratios and control gains are strongest. The result supports use of the surrogate to reject obviously poor candidates and prioritize expensive physics evaluations. It does not justify replacing the state matrix. The final proposed plan and every reported baseline are evaluated using the physics model, so surrogate error cannot alter the numerical conclusions in Table 4.

For larger studies, a boundary-aware acquisition rule would be preferable to uniform training. Candidates with a predicted margin close to zero, a large predictive variance, or an unusually low cost would receive priority for exact evaluation. A conservative decision threshold could also be shifted left by one or two validation-error standard deviations. These safeguards convert machine learning from an opaque optimizer into an auditable computational accelerator whose failures are measurable and contained.

5.6. Stress Testing under Resource and Demand Shifts

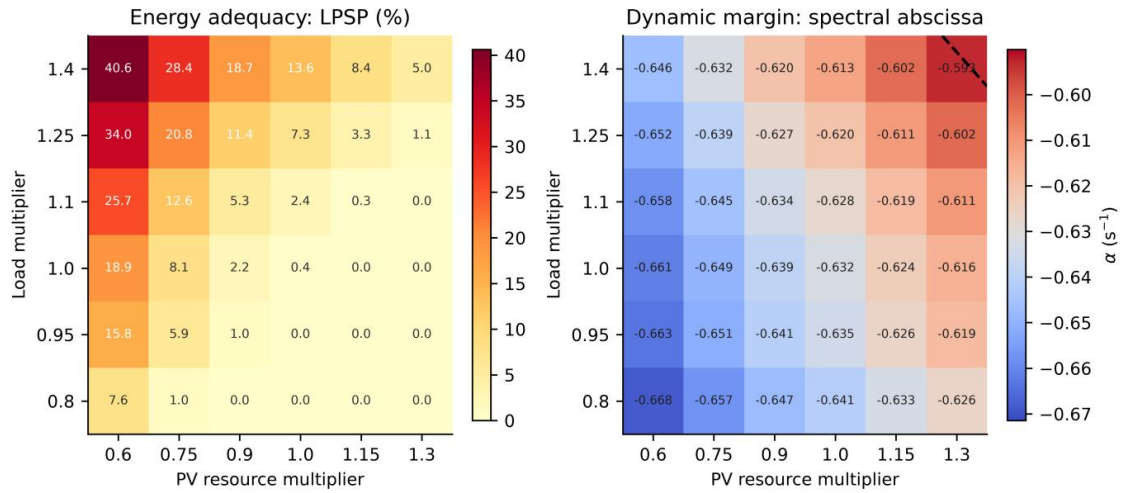


Figure 8. Reliability and stability stress-test matrices for the proposed design.

Figure 8 uses a wider stress envelope than the local sensitivity test: solar resource varies from 0.60 to 1.30 of the base profile and load from 0.80 to 1.40. The cell values make the two limits directly readable. Energy adequacy deteriorates most rapidly in the high-load, low-resource corner, whereas the modal margin is more sensitive to the PV-to-grid-forming rating ratio. The black contour in the right panel marks $\alpha = -0.60 \text{ s}^{-1}$. Its location differs from the LPSP boundary, so neither test can stand in for the other.

The direction of a corrective design change follows the active constraint. A deficit-dominated case calls for more renewable energy, storage energy, or flexible demand. A modal-limit violation calls first for converter power or controller retuning; adding battery energy alone has little effect on the fast states represented by $\mathbf{A}$. This distinction is visible in the two panels and

is the practical reason for keeping the reliability and stability gates separate.

5.7. Modal Participation and Annual Operation

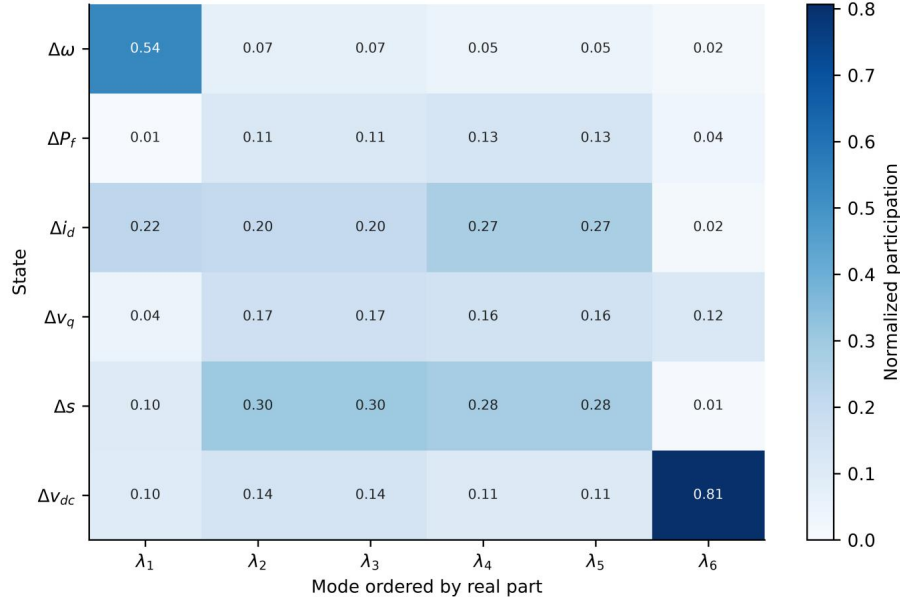


Figure 9. Normalized state participation in the six modes of the selected design.

The participation map identifies the states responsible for the modes shown in Figure 4. The modes closest to the imaginary axis are dominated by frequency and filtered-power states, while the faster modes contain larger current-, voltage-, and dc-link components. This result supports the choice of droop coefficients and virtual inertia as planning variables: they act directly on the states that dominate the limiting modes. The map should be read as a property of the reduced model, not as a substitute for participation factors from a full converter model.

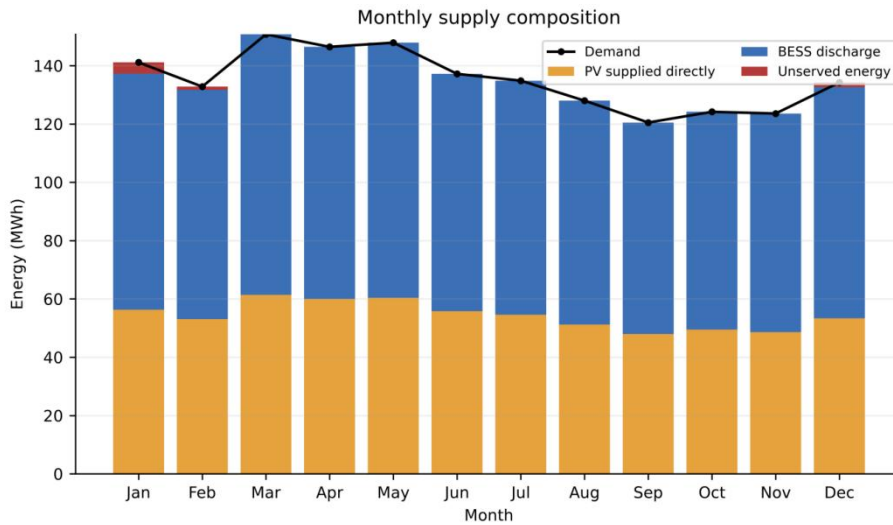


Figure 10. Monthly demand and supplied-energy composition for the selected design.

The monthly accounting complements the single-day trace in Figure 5. Direct PV supplies most daytime demand, while BESS discharge closes the intraday mismatch. Residual unserved energy is concentrated in periods where low resource and high demand coincide; curtailed energy, reported separately in the dispatch calculation, indicates that some annual surplus cannot compensate for deficits occurring after the storage limits bind. The result explains why annual energy totals alone are insufficient for sizing.

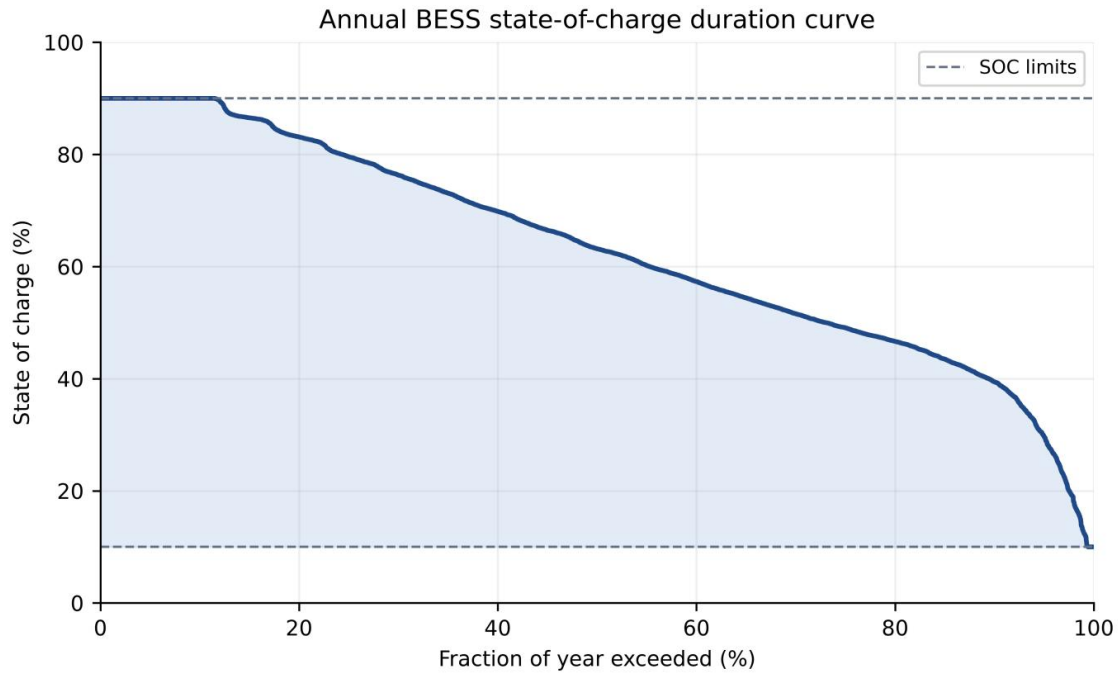


Figure 11. Annual BESS state-of-charge duration curve for the selected design.

The duration curve shows how often the battery occupies each part of its usable range. Long residence near the lower bound would indicate that the energy rating is frequently binding; long residence near the upper bound would point to surplus energy or insufficient converter power. Here the curve spans the operating window without remaining pinned to one limit for most of the year, which is consistent with a design governed jointly by adequacy, converter power, and cost.

5.8. Ablation Study

The ablation separates the effect of each constraint. Cost-only planning can favor a design that underinvests in the service required by an islanded system. Reliability-only planning corrects annual energy deficits but has no mechanism to reject a poorly damped operating point. Stability-only planning can choose a dynamically comfortable plan whose battery energy is insufficient for extended deficits. Only the full framework has a definition of feasibility that

matches both long-term continuity of supply and short-term return to equilibrium.

The ablation also discourages an overly broad claim about optimality. The reported full-framework solution is the least-cost jointly feasible member of the sampled candidate set, not a mathematical proof of the global optimum over a continuous nonconvex domain. This wording is deliberate. A future mixed-integer or surrogate-assisted global search can improve the candidate set, but it must retain the same independent feasibility checks. Separating solution quality from model validity makes subsequent algorithmic improvements straightforward.

Table 6. Ablation of reliability and stability feasibility gates.

Formulation	Cost kUSD/y	LPSP %	$\alpha \text{ s}^{-1}$	$\zeta_{\min}$	Jointly feasible
Cost only	3121.7	9.889	-1.140	0.465	No
Reliability only	4923.5	0.310	-0.367	0.438	No
Stability only	3121.7	9.889	-1.140	0.465	No
Full framework	5287.9	0.407	-0.632	0.421	Yes

5.9. Comparison with Recent Research and Limitations

The study extends reliability-oriented planning [18-22] by testing a modal margin at every retained design. In contrast to small-signal studies that hold equipment ratings fixed [12-16], the state matrix changes with the selected PV and BESS ratings. The comparison is therefore methodological rather than numerical: published cases use different networks, controllers, costs, and reliability targets, so their optimal capacities are not directly comparable with the synthetic case reported here.

The numerical values are planning evidence, not a bankable design. The resource and load series are synthetic, the cost model omits calendar ageing and component replacement, and the six-state model assumes balanced operation without current limiting. Constant-power loads, phase-locked loops, unbalance, line resonances, and controller saturation can introduce modes that are absent here. The spectral test is valid only around an equilibrium for which the assumed controller mode remains active.

The reliability limit also should not be interpreted as a universal standard. A project-specific value must reflect critical-load composition, repair time, regulation, and the value of lost load. Likewise, the load-dependent adjustment used in the stress test is a reduced-model sensitivity, not an identified plant coefficient. Measured operating points and a calibrated converter model are required before the contour in Figure 8 can be used for equipment approval.

The surrogate has a deliberately narrow role. Its usefulness depends on false-safe error near

the $\alpha = -0.60 \text{ s}^{-1}$ boundary, not only on global R^2 . A field implementation should reject extrapolation, attach uncertainty to boundary predictions, and verify every candidate that proceeds to engineering review with the original state model. These safeguards preserve the computational benefit without transferring the final decision to the regression model.

5.10. Practical Implementation Path

For a site study, the hourly dataset should include at least one synchronized year of load, irradiance, outages, and battery state. Event records at a faster sampling rate are needed for dynamic identification. Critical and deferrable demand must be separated, and outage intervals should remain visible rather than being interpolated away. Scenario reduction may be used for computation, but low-resource sequences and component outages must be retained even when their probability is small.

The state model should then be rebuilt from the implemented controller. Current- and voltage-loop gains, power filters, virtual impedance, droop slopes, delays, and current limits should come from controller settings, manufacturer data, or perturbation tests. Parameter uncertainty can be handled by evaluating the modal gate at nominal and adverse combinations instead of relying on a single best-fit matrix.

Shortlisted designs should be checked in stages: unseen annual operating data, a full-order averaged converter model, electromagnetic-transient simulation, and finally controller- or power-hardware-in-the-loop testing. The same capacities and gains must be carried through each stage. If a failed test requires retuning, the planning calculation should be repeated because the feasibility boundary has changed.

The final report should archive the data period, scenario weights, parameter ranges, surrogate version, validation errors, and exact-model checks. This record makes later re-planning possible when prices, load growth, degradation, or available grid-forming hardware change, and it shows whether a revision is driven by energy adequacy or by dynamic margin.

6. Conclusion

This paper presented a source-storage capacity planning framework for islanded microgrids in which small-signal stability is enforced during, rather than after, optimization. The formulation jointly selects PV capacity, BESS power and energy, droop coefficients, and virtual inertia. Annual dispatch supplies cost, curtailment, throughput, and reliability indices, while a capacity-dependent linearized model supplies spectral-abscissa and modal-damping constraints. The reproducible case study showed that reliability-feasible plans can still be weakly damped

and that explicit modal constraints remove these designs from consideration. The main engineering implication is that storage energy, grid-forming power, and controller gains should be coordinated: kilowatt-hours secure energy adequacy, whereas converter kilowatts and control parameters govern fast stability. The proposed framework is suitable as a screening layer for international planning studies, provided that final designs are verified using site data and a higher-order converter model.

Nomenclature

The acronyms and principal symbols used throughout this paper are summarized in Table 7.

Table 7. Acronyms and principal symbols.

Term	Definition	Term	Definition
BESS	Battery energy storage system	EENS	Expected energy not supplied
GFM	Grid-forming converter	LPSP	Loss-of-power-supply probability
PV	Photovoltaic generation	SOC	State of charge
P_{PV}^r	Rated PV power	P_B^r	Rated BESS converter power
P_B^r	Rated battery energy	m_p, m_q	Active- and reactive-power droop slopes
H_v	Virtual-inertia coefficient	$\alpha(A)$	Spectral abscissa
ζ_i	Damping ratio of mode i	S	Set of operating scenarios

Appendix A. Reproducibility Notes

The synthetic year contains 8760 hourly samples and uses random seed 42. Candidate variables are uniformly sampled within Table 3. For every candidate, dispatch begins at 55% state of charge, applies charging efficiency 0.95 and discharging efficiency 0.94, and enforces a 10–90% state-of-charge window. The reported search contains 2200 candidates. Reproduction requires implementing Equations (1)–(14), the parameter values in Tables 2–3, and the stated reduced-order state vector. Numerical differences caused by linear-algebra libraries should be reported rather than hidden.

References

- [1] Lasseter, R. H. (2002). MicroGrids. In 2002 IEEE Power Engineering Society Winter Meeting (Vol. 1, pp. 305–308). <https://doi.org/10.1109/PESW.2002.985003>
- [2] Guerrero, J. M., Vasquez, J. C., Matas, J., de Vicuña, L. G., & Castilla, M. (2011). Hierarchical control of droop-controlled AC and DC microgrids—A general approach toward standardization. IEEE Transactions on Industrial Electronics, 58(1), 158–172.

- <https://doi.org/10.1109/TIE.2010.2066534>
- [3] Pogaku, N., Prodanović, M., & Green, T. C. (2007). Modeling, analysis and testing of autonomous operation of an inverter-based microgrid. *IEEE Transactions on Power Electronics*, 22(2), 613–625. <https://doi.org/10.1109/TPEL.2006.890003>
 - [4] He, J., & Li, Y. W. (2011). Analysis, design, and implementation of virtual impedance for power electronics interfaced distributed generation. *IEEE Transactions on Industry Applications*, 47(6), 2525–2538. <https://doi.org/10.1109/TIA.2011.2168592>
 - [5] Rocabert, J., Luna, A., Blaabjerg, F., & Rodríguez, P. (2012). Control of power converters in AC microgrids. *IEEE Transactions on Power Electronics*, 27(11), 4734–4749. <https://doi.org/10.1109/TPEL.2012.2199334>
 - [6] Schiffer, J., Zonetti, D., Ortega, R., Stanković, A. M., Sezi, T., & Raisch, J. (2016). A survey on modeling of microgrids—From fundamental physics to phasors and voltage sources. *Automatica*, 74, 135–150. <https://doi.org/10.1016/j.automatica.2016.07.036>
 - [7] Kroposki, B., Johnson, B., Zhang, Y., Gevorgian, V., Denholm, P., Hodge, B.-M., & Hannegan, B. (2017). Achieving a 100% renewable grid. *IEEE Power and Energy Magazine*, 15(2), 61–73. <https://doi.org/10.1109/MPE.2016.2637122>
 - [8] Tamrakar, U., Shrestha, D., Maharjan, M., Bhattarai, B. P., Hansen, T. M., & Tonkoski, R. (2017). Virtual inertia: Current trends and future directions. *Applied Sciences*, 7(7), 654. <https://doi.org/10.3390/app7070654>
 - [9] Milano, F., Dörfler, F., Hug, G., Hill, D. J., & Verbič, G. (2018). Foundations and challenges of low-inertia systems. In *2018 Power Systems Computation Conference* (pp. 1–25). <https://doi.org/10.23919/PSCC.2018.8450880>
 - [10] IEEE Standards Association. (2018). IEEE standard for the specification of microgrid controllers (IEEE Std 2030.7-2017). <https://doi.org/10.1109/IEEESTD.2018.8340204>
 - [11] IEEE Standards Association. (2018). IEEE standard for interconnection and interoperability of distributed energy resources (IEEE Std 1547-2018). <https://doi.org/10.1109/IEEESTD.2018.8332112>
 - [12] Samimi, A., Nikzad, M., & Zakipour, A. (2025). Optimal sizing model of battery energy storage in a droop-controlled islanded multi-carrier microgrid. *Scientific Reports*, 15, 2540. <https://doi.org/10.1038/s41598-025-86368-2>
 - [13] Hassan, C. T., Jadoon, T. M., & Arif, A. (2025). Delay-dependent small-signal stability analysis of inverter-based islanded microgrids. *IET Smart Grid*, 8(1), e12206. <https://doi.org/10.1049/stg2.12206>
 - [14] Shi, K., Bai, Y., Xu, P., & Sun, Y. (2025). Dual-side virtual inertia control for stability enhancement in islanded AC–DC hybrid microgrids. *Electric Power Systems Research*, 241, 111396. <https://doi.org/10.1016/j.epsr.2024.111396>
 - [15] Sati, S. E., Al-Durra, A., Zeineldin, H. H., EL-Fouly, T. H. M., & El-Saadany, E. F. (2025). Decentralized frequency restoration and stability enhancement for virtual synchronous machines. *Applied Energy*, 377, 124544. <https://doi.org/10.1016/j.apenergy.2024.124544>
 - [16] Maurya, S. S., Jain, T., & Umarikar, A. C. (2025). Synchronverter with emulation based current control to improve small signal stability in islanded microgrid. *IEEE Journal of Emerging and Selected Topics in Industrial Electronics*, 6(4), 1688–1699. <https://doi.org/10.1109/JESTIE.2025.3590837>
 - [17] Khan, M. Z., Mu, C., Khan, M. M., Ke, W., & Khan, M. Q. (2026). Active power sharing and frequency regulation of islanded microgrids under communication islanding. *Electric Power Systems Research*, 254, 112624. <https://doi.org/10.1016/j.epsr.2025.112624>
 - [18] Saleheen, M. Z., Wagner, M., & Wang, H. (2026). Ensuring reliability in 100% renewable microgrids: A scenario-based joint planning and operational design framework. *Energy Conversion and Management*, 363, 121645. <https://doi.org/10.1016/j.enconman.2026.121645>

- [19] Camargo-Berruero, J. C., Mejía-Giraldo, D. A., & Lemos-Cano, S. (2026). Reliability-based sizing of islanded multi-energy microgrid. *Energy Systems*, 17(1), 239–276. <https://doi.org/10.1007/s12667-024-00672-w>
- [20] Wang, G., Dong, K., Song, Y., Cao, S., Duan, J., Duan, G., Wu, D., & Li, S. (2025). Optimal battery storage capacity planning model and DOD regulation strategy for islanded microgrid. *Energy*, 332, 137200. <https://doi.org/10.1016/j.energy.2025.137200>
- [21] Mühlbauer, A., Fan, Y. F., Sambor, D. J., & Jacobson, M. Z. (2025). Minimizing the multi-decadal cost of islanded renewable-electricity microgrids for different climate zones. *Smart Energy*, 20, 100208. <https://doi.org/10.1016/j.segy.2025.100208>
- [22] Sagor, A. R., Hazari, M. R., Ahmad, S., Ogliari, E., Hossain, C. A., Jahan, E., & Mannan, M. A. (2025). Optimal sizing of islanded microgrid using pelican optimization algorithm. *Electric Power Systems Research*, 238, 111088. <https://doi.org/10.1016/j.epsr.2024.111088>
- [23] Huang, C.-J., & Kuo, P.-H. (2019). Multiple-input deep convolutional neural network model for short-term photovoltaic power forecasting. *IEEE Access*, 7, 74822–74834. <https://doi.org/10.1109/ACCESS.2019.2921238>
- [24] Kundur, P., et al. (2004). Definition and classification of power system stability. *IEEE Transactions on Power Systems*, 19(3), 1387–1401. <https://doi.org/10.1109/TPWRS.2004.825981>
- [25] Kundur, P. (1994). *Power system stability and control*. McGraw-Hill.
- [26] Olivares, D. E., et al. (2014). Trends in microgrid control. *IEEE Transactions on Smart Grid*, 5(4), 1905–1919. <https://doi.org/10.1109/TSG.2013.2295514>
- [27] Bidram, A., & Davoudi, A. (2012). Hierarchical structure of microgrids control system. *IEEE Transactions on Smart Grid*, 3(4), 1963–1976. <https://doi.org/10.1109/TSG.2012.2197425>
- [28] D’Arco, S., Suul, J. A., & Fosso, O. B. (2015). A virtual synchronous machine implementation for distributed control of power converters in SmartGrids. *Electric Power Systems Research*, 122, 180–197. <https://doi.org/10.1016/j.epsr.2015.01.001>
- [29] Unamuno, E., & Barrena, J. A. (2015). Hybrid AC/DC microgrids—Part I: Review and classification of topologies. *Renewable and Sustainable Energy Reviews*, 52, 1251–1259. <https://doi.org/10.1016/j.rser.2015.07.194>
- [30] Parisio, A., Rikos, E., & Glielmo, L. (2014). A model predictive control approach to microgrid operation optimization. *IEEE Transactions on Control Systems Technology*, 22(5), 1813–1827. <https://doi.org/10.1109/TCST.2013.2295737>
- [31] Simpson-Porco, J. W., Dörfler, F., & Bullo, F. (2013). Synchronization and power sharing for droop-controlled inverters. *Automatica*, 49(9), 2603–2611. <https://doi.org/10.1016/j.automatica.2013.05.018>
- [32] Farrokhabadi, M., et al. (2020). Microgrid stability definitions, analysis, and examples. *IEEE Transactions on Power Systems*, 35(1), 13–29. <https://doi.org/10.1109/TPWRS.2019.2925703>
- [33] Hatziargyriou, N. (Ed.). (2014). *Microgrids: Architectures and control*. Wiley-IEEE Press.
- [34] Yazdani, A., & Iravani, R. (2010). *Voltage-sourced converters in power systems*. Wiley-IEEE Press.
- [35] Blaabjerg, F., Teodorescu, R., Liserre, M., & Timbus, A. V. (2006). Overview of control and grid synchronization for distributed power generation systems. *IEEE Transactions on Industrial Electronics*, 53(5), 1398–1409. <https://doi.org/10.1109/TIE.2006.881997>

Stock Market Liberalization and Institutional Investor Clique Behavior: Evidence from China

Muzi Jiang^{1,*}, Ping Yang²

¹ The Open University of Sichuan, China

² Southwestern University of Finance and Economics, China

Received: Sept 11, 2026

Revised: Sept 15, 2026

Accepted: Sept 30, 2026

Published online: Sept 30, 2026

To appear in: *International Journal of Advanced AI Applications*, Vol. 2, No.10 (October 2026)

*Corresponding Author: Muzi Jiang(muzi.jiang@foxmail.com)

Abstract. This study examines how stock market liberalization affects the clique behavior of domestic institutional investors, distinguishing hotspot clustering from value clustering. Using panel data on Chinese listed firms from 2011 to 2019 and a difference-in-differences design, we identify institutional cliques with the Louvain community detection algorithm, and we apply large language models to cross-validate clique measurement and to construct a cross-border thematic similarity index. We find that stock market liberalization significantly reduces hotspot clustering while promoting value clustering. The inhibitory effect on hotspot clustering is more pronounced for speculative stocks, whereas the promotion of value clustering is stronger among industry leaders. These effects operate through improved market efficiency, optimized institutional holding structures, and enhanced cross-border narrative alignment. Institutions that prioritize fundamental analysis achieve superior investment returns. This study advances research on institutional cliques by integrating large language models into clique measurement and identifies a channel through which foreign investors mitigate short-term speculation in emerging liberalizing markets.

Keywords: stock market liberalization; domestic institutional investors; clique behavior; hotspot clustering; value clustering.

1. Introduction

Stock market liberalization has been widely studied for its effects on the global economy [1,2], stock market efficiency [3], information environment [4,5] and corporate behavior [6-8].

However, the role of foreign investors in liberalizing stock markets remains controversial. Foreign capital may enhance market efficiency by curbing domestic opportunism [9], yet it can also introduce instability through positive feedback trading in emerging markets [10]. Prior studies document herding tendencies among domestic institutional investors in China [11], but the influence of foreign capital participation on their trading behavior remains largely unexplored.

China progressively liberalized its financial sector through the Shanghai-Hong Kong Stock Connect (2014) and the Shenzhen-Hong Kong Stock Connect (2016), collectively termed stock market liberalization. Unlike earlier QFII and RQFII schemes, this reform allows foreign capital to trade A-shares anonymously via a centralized platform with fewer quota restrictions [6]. Consequently, foreign ownership in the A-share market reached 8.5% of free-float market value by 2019, making foreign investors the largest institutional group (see Figure 1). Notably, foreign capital has consistently outperformed the market by prioritizing fundamentals¹, eroding the speculative advantage of domestic institutional investors. As all investors are profit-driven, whether domestic institutional investors adopt similar strategies requires further empirical investigation.

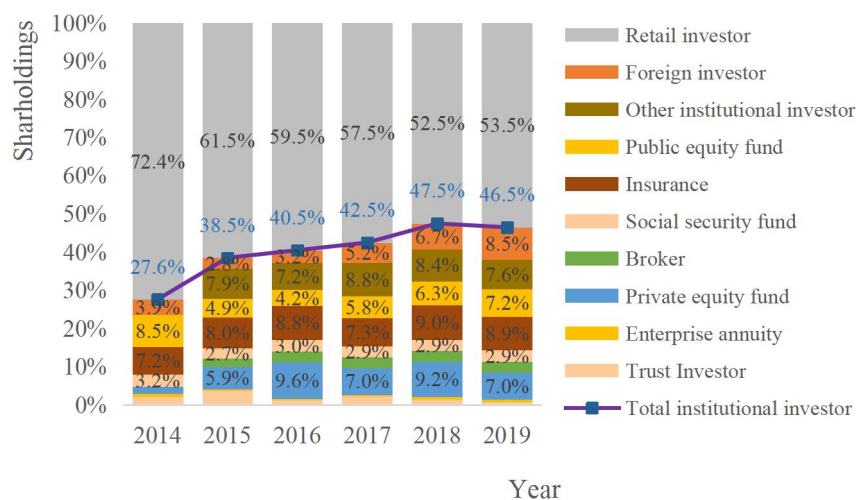


Figure 1. The percentage of market value held by various types of investors in A-shares (free float market value).

Notes: This figure illustrates the percentage of market value held by various types of investors in A-shares from 2014 to 2019.

¹ Among the top 500 stocks with the highest northbound holdings, the share of those with a return on equity (ROE) exceeding 15% grew from 20% in 2017 to 30% in 2019. Meanwhile, foreign capital trading via the Shanghai-Hong Kong Stock Connect was the only investor group to achieve consistent profits from 2014 to 2017.

Institutional investors often engage in interconnected trading to influence corporate policies [12,13]. In China, they frequently co-hold equity, forming 'institutional cliques' based on shared information [14]. These cliques enhance information dissemination and collective governance influence through voting power or exit threats [15]. Such clustering is not unique to emerging markets. In the U.S., top actively managed funds concentrate holdings in major tech firms, such as Facebook, Amazon, Apple, Microsoft, Netflix, Google. However, such cliques are often criticized for exacerbating agency problems and herding behavior [14,16,17]. In fact, their impact on market efficiency depends largely on whether they pursue quality assets or speculative trends. Distinguishing between these clique types is therefore crucial for assessing the implications of stock market liberalization for market stability.

To address these research questions, we classify clique behavior based on stock holding characteristics, corroborating institutional groupings through large language model (LLM)-based textual analysis of public disclosures. Hotspot clustering refers to cliques favoring stocks with high media attention; value clustering reflects a preference for stocks with strong fundamentals. To capture cross-border information spillover, we construct a thematic similarity index via LLM-based semantic analysis of matched domestic and foreign research reports. Using a staggered difference-in-differences model and a sample of Chinese listed firms from 2011 to 2019, we empirically examine how stock market liberalization influences these two clique types among domestic institutional investors. The results reveal that liberalization reduces hotspot clustering while promoting value clustering. Specifically, liberalization mitigates hotspot clustering in speculative stocks and fosters value clustering in industry leaders. These effects are driven by improvements in market efficiency and institutional holding structures. Liberalization reduces mispricing and opportunistic collusion in hotspot stocks; it also enhances the informational quality of value stocks and extends holding periods. Moreover, domestic investors who prioritize fundamental analysis over speculative trading tend to achieve superior returns.

Our research makes the following contributions. First, we identify stock market liberalization as a channel for cross-border information exchange. While existing studies primarily explore information exchange within social networks—such as geographic proximity [18], alumni connections [19] or co-ownership networks [20], we show that domestic institutional investors emulate foreign investors by shifting toward fundamentally valuable assets, enriching resource dependency theory [21].

Second, we advance the literature on institutional clique behavior by distinguishing hotspot

from value clustering. While prior research has largely examined institutional investor networks in corporate governance and firm value [14,15,22,23], the drivers of different clustering behaviors remain underexplored. Incorporating LLM-based validation and cross-border thematic similarity measures, we enhance understanding of investor dynamics under liberalization and offer a framework for identifying clique types.

Third, our study sheds light on mitigating the short-term clique behavior of domestic institutional investors. Volatility in emerging markets is often driven by retail investors' irrational trading [11,24]. We demonstrate that domestic investors adjust strategies by observing foreign investors' focus on fundamentals, providing new evidence on the benefits of market liberalization for developing economies.

2. Hypothesis development

China's capital market is dominated by retail investors [24], who often purchase attention-grabbing stocks without thorough fundamental analysis [25]. While retail investors may profit from such speculative trading, the widespread preference for hotspot investments encourages domestic institutional investors to exacerbate market fluctuations [26,27], even when fully aware of the overvaluation [28-30]. Given environmental differences, foreign capital may maximize returns by either aligning with speculative trends [31] or adopting fundamental trading [32], thereby influencing domestic institutional clique behavior in distinct ways.

On one hand, foreign capital's bubble-chasing may incentivize domestic institutions to form speculative cliques, namely hotspot clustering. China's stock market has long exhibited speculative tendencies, with valuations diverging from fundamentals [33,34]. Within this environment, foreign investors may engage in positive feedback trading or momentum investing to capture excess returns [30,35]. Sophisticated investors often strategically time overvaluation corrections rather than directly counteract mispricing [36]. Moreover, institutional investors' network advantages enhance bargaining power, fostering collaboration with corporate managers [14,37]. This collusion amplifies domestic investors' influence, enabling coordinated exits to maximize private gains. Consequently, stock market liberalization may exacerbate volatility and mispricing, encouraging hotspot clustering for speculative returns.

Meanwhile, under performance evaluation pressures and market rankings [16,38], domestic institutional investors may rationally mimic peers' decisions [38,39]. If foreign capital targets short-term gains in hotspot stocks, domestic institutions are likely to follow suit. This herding behavior reduces intra-group competition and impedes private information integration into

stock prices [14]. Consequently, domestic institutional investors engage in hotspot clustering, forming speculative cliques around hotspot stocks. Based on this, we propose the following hypothesis:

H(a): Stock market liberalization stimulates domestic institutional investors to engage in hotspot clustering rather than value clustering.

On the other hand, foreign capital's value investment philosophy encourages domestic institutional investors to cluster around higher-value stocks. Unlike earlier liberalization policies (e.g., QFII and RQFII), current reforms relax restrictions on foreign capital, enhancing its trading impact [40]. This creates an ideal platform for foreign capital to correct mispricing and curb managerial opportunism [41]. Extensive evidence indicates that foreign institutions typically adopt rational, long-term strategies that reduce speculation [42,43]. As a result, domestic institutional investors are less likely to profit from mispriced stocks and thereby attenuating their hotspot clustering.

Moreover, foreign investors from developed markets possess superior investment experience and information processing capabilities [44]. Their portfolio returns often reflect the integration of firm-specific fundamental information into trading strategies [45]. Liao et al. [32] further suggest that northbound inflows, by incorporating firm fundamentals, reshape domestic investors' trading behavior. Consequently, domestic institutional investors are motivated to emulate foreign capital's value-oriented strategies to achieve superior returns.

Resource dependency theory posits that network members access information through social learning and intra-network communication [21,46]. As foreign capital trading increases the costs of hotspot clustering for domestic institutional investors, domestic institutional investors shift towards fundamental trading. Consequently, these investors and their peers abandon speculative strategies in favor of collective approaches focused on long-term, higher returns, thereby clustering more around value stocks.

Thus, we propose a competing hypothesis as follows:

H(b): Stock market liberalization stimulates domestic institutional investors to engage in value clustering rather than hotspot clustering.

3. Data and research design

3.1. Data

As SH-HK Stock Connect started in 2014 and SZ-HK Stock Connect in 2016, our full sample covers the period from 2011 to 2019 to ensure symmetry in the three years before and after the

shock, excluding the effect of the epidemic. Data on shareholdings of listed companies are from WIND database, and the dynamic eligible shares traded under the Stock Connect programs are collected from Hong Kong Exchanges and Clearing Limited (HKEX). Data for media reports are from China Research Data Service (CNRDS). Other financial data are from CSMAR database. The sample covers six domestic institutional investor types: fund management companies, securities firms, insurers, social security funds, trust companies, and financial companies. We exclude ST, ST, and PT firm-year observations and those with missing values. Continuous variables are winsorized at the 1st and 99th percentiles. The final sample comprises 17,020 observations.

3.2. Variable definition

3.2.1. Institutional investor clique

Following Crane et al. [47], we identify institutional groups by linking institutions based on common holdings in the fourth quarter of each year. A clique exists if each domestic institutional investor holds at least a 5% stake in a common firm by year-end. Specifically, off-diagonal elements are set to one if both institutions in a pair hold at least a 5% stake in the same firm, and zero otherwise. This procedure yields an $N \times N$ binary matrix of institution-to-institution connections. We then import the matrix into Gephi and apply the Louvain algorithm [48] to extract institutional communities. Figure 2 presents the resulting cliques in 2019: nodes of the same color connected by edges represent investors within the same clique. Information flows more readily within cliques, and this facilitates network communication. However, the absence of shared holdings among other institutions makes coordination more challenging.

Then, we calculate three measures as follows:

$$CliqueHold_{i,t} = \sum_{j=1}^n \lambda_{i,j,i} \times Cliqueinst_{i,j,t} \quad (1)$$

$$CliqueTop_{i,t} = \text{Max}\{CliqueSum_{k,i,t} | k \in M\} \quad (2)$$

$$CliqueHerf_{i,t} = \sum_{j=1}^M \lambda_{i,j,t} \times (CliqueSum_{k,j,t})^2 \quad (3)$$

where λ is an indicator variable equal to one if institution i belongs to a clique j at time t . N represents the total number of institutional investors within the investor group in year t . $Cliqueinst_{i,j,t}$ represents institution i 's ownership percentage in firm j at time t . $CliqueSum_{k,i,t}$ is the total proportion of tradable shares of stock i held by group k in year t . $CliqueHold_{i,t}$ is the total fraction of the firm owned by all cliques. $CliqueTop_{i,t}$ represents the total ownership

share held by the single clique with the largest stake in firm i . $CliqueHerf_{i,t}$ is calculated as the sum of the squared ownership fractions of all cliques with holdings in firm i , reflecting the concentration of ownership across cliques.

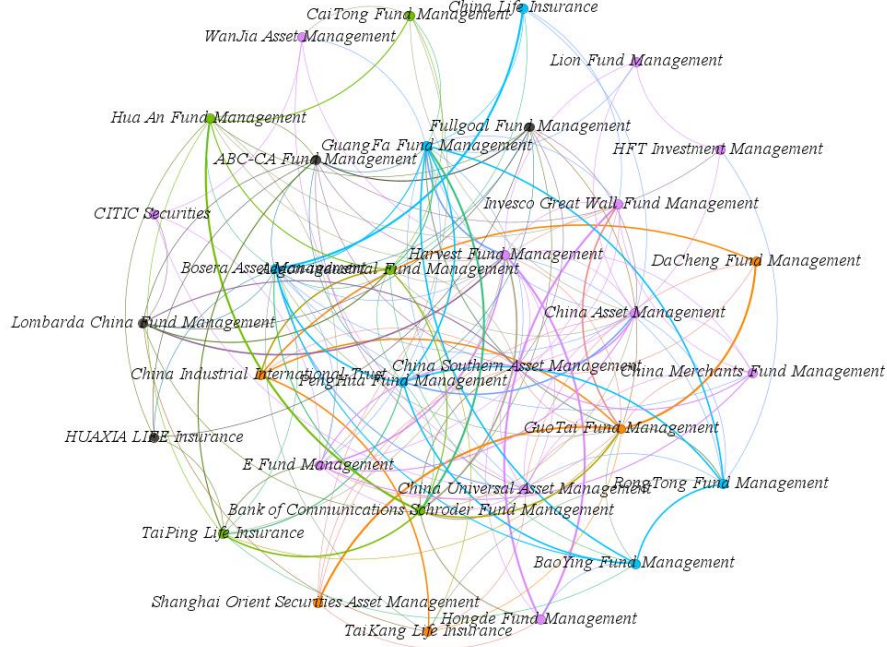


Figure 2. Example of a clique in 2019.

Notes: This figure presents the network of domestic institutional investors in 2019 based on the institutional shareholdings data. Each dot of the same color represents institutional investors belonging to the same clique.

3.2.2. The hotspot stock

Extant literature indicates that the media sentiment triggers investor overreactions [49] and significantly impact financial markets through individual investors' stock trading [50]. We propose that stocks with more media coverage tend to attract higher investors' attention. The development of the Internet has made online media an important channel for investors to obtain information. Therefore, we use the natural logarithm of 1 plus the total number of positive news reports on company i by online media in the year t to measure its hotspot degree.

$$Hotspot_{i,t} = \ln(1 + PositiveNews_{i,t}) \quad (4)$$

3.2.3. The value stock

We measure a firm's fundamental value using the moving weighted average of its return on equity (ROE) over the past three years (Value). ROE is defined as net profit margin $\times$ total asset turnover $\times$ equity multiplier, capturing a firm's profitability, asset efficiency, and financial

leverage more comprehensively than single financial indicators. However, prior research suggests that ROE can be manipulated by management [51], and relying solely on a single-year measure may introduce distortions. Thus, we adopt a three-year moving weighted average to provide a more stable assessment of a firm's fundamental value.

3.2.4. LLM applications in variable validation and behavioral classification

To enhance the construct validity of core variables and uncover the micro-level information transmission mechanism of stock market liberalization, we introduce large language model (LLM) techniques as a methodological complement to traditional quantitative analysis. We apply LLM in two logically connected dimensions: (1) cross-validating the accuracy of Louvain-based institutional clique identification; (2) constructing a cross-border thematic similarity index to distinguish between value and hotspot clustering, and to serve as a proxy for the information spillover channel.

3.2.4.1. Validity check of institutional clique identification.

Traditional community detection based on common shareholdings identifies co-holding groups but cannot verify whether these groups reflect actual institutional coordination with consistent investment logics. To address this, we adopt a two-stage LLM framework. In Stage 1, we extract structured attributes from 12,743 institutional quarterly reports, holding announcements, and investor meeting minutes. Using GPT-4 (temperature = 0.3), we parse each institution's disclosures to identify three core attributes: stated investment style, sector allocation preferences, and documented co-investment patterns. In Stage 2, we feed these vectors into the model for unsupervised community division, with the model blinded to the original network results. We then calculate the Jaccard similarity between LLM-identified groups and Louvain-derived cliques. The average annual similarity reaches 0.782 and is significant at the 1% level, confirming that our network-based measurement accurately captures real-world institutional coordination.

3.2.4.2. Construction of cross-border thematic similarity index.

Grounded in resource dependence theory [52], institutions engaging in value clustering should exhibit narrative convergence with foreign investors, whereas those engaging in hotspot clustering should show divergence. We construct a cross-border thematic similarity index via LLM-based text analysis to verify this theoretical distinction.

(1) Sample and data

We focus on the top 30 stocks with the largest northbound holdings each quarter and collect

two matched sets of research reports at the stock-quarter level: quarterly investment commentaries from 17 major global asset managers (e.g., BlackRock, Vanguard) and quarterly research reports from the top 50 Chinese mutual fund companies. Only report pairs covering the same stock in the same quarter are retained to ensure contemporaneous information sets.

(2) LLM-based analytical procedure

Table 1. LLM Applications in Variable Measurement and Validation

<i>Application Module</i>	<i>Specific Procedure</i>	<i>Model Configuration</i>	<i>Core Purpose and Output</i>
1. Institutional Clique Validity Check	Stage 1: Structured attribute extraction	GPT-4, temperature = 0.3	Extract 3 investment attributes from institutional disclosures: investment style, sector preference, co-investment patterns
	Stage 2: Unsupervised grouping & cross-validation	GPT-4, temperature = 0.3	Conduct blind independent community division; calculate Jaccard similarity with Louvain cliques to verify construct validity
2. Cross-Border Thematic Similarity Construction	Step 1: Fundamental rationale extraction	GPT-4o, temperature = 0.3	Extract top 3 core fundamental investment reasons from each report, filtering out sentiment and technical analysis
	Step 2: Semantic vectorization	1536-dimension LLM embedding model	Convert textual rationales into standardized semantic vectors; compute average vectors for domestic and foreign samples by stock-quarter
	Step 3: Similarity quantification	Cosine similarity algorithm	Calculate cosine similarity between averaged domestic and foreign vectors as the final thematic similarity score
	Step 4: Regime classification	ROC curve method	Determine optimal threshold of 0.62; classify samples into high-alignment (value clustering) and low-alignment (hotspot clustering) regimes

We adopt a three-step framework based on GPT-4o. In Step 1, we extract the top 3 core fundamental investment rationales from each report via a standardized prompt, explicitly excluding market sentiment, technical analysis, and market timing. The prompt is specified as:

"You are a professional financial analyst. Extract the top 3 core fundamental reasons for recommending or holding this stock from the following research report. Focus only on company-specific fundamentals (e.g., earnings growth, market share, technological innovation, governance quality) and industry-level factors. Do not include technical analysis, market timing, or sentiment-based arguments. Output the results in a numbered list format."

In Step 2, we convert each extracted rationale set into a 1536-dimensional semantic vector

via LLM embedding. We compute separate average vectors for foreign and domestic reports by stock-quarter. In Step 3, we measure thematic similarity as the cosine similarity between these two averaged vectors for each stock-quarter observation.

(3) Threshold determination and regime classification

We identify the optimal similarity threshold of 0.62 via the receiver operating characteristic (ROC) curve method, which achieves the strongest discriminant validity between the two clique types. Observations above 0.62 are classified into the high-alignment regime, corresponding to value clustering where domestic narratives align closely with the fundamental-focused investment logic of foreign institutions. Observations at or below 0.62 are classified into the low-alignment regime, corresponding to hotspot clustering where domestic institutions tend to prioritize non-fundamental market factors. This data-driven threshold provides an objective, empirically grounded boundary for distinguishing value clustering from hotspot clustering.

Table 1 summarizes all LLM-based procedures, model configurations, and corresponding outputs across the two application modules.

3.2.5. Stock market liberalization

First, we define the dummy variable *Treat*, which equals 1 if the firm is listed on the SH-HK or SZ-HK Stock Connect, and 0 otherwise. We then create the dummy variable *Post*, which equals 1 if the firm is included in either program in the current or any subsequent year, and 0 otherwise. Finally, we construct the interaction term *HSSC* by multiplying the *Treat* and *Post* variables.

3.3. Methodology

To test hypothesis, we examine the impact of the stock market liberalization on domestic institutional investors' clique behavior. Namely, whether stock market liberalization stimulates domestic institutional investors to engage in hotspot clustering or value clustering.

$$\begin{aligned} Clique_{i,t} = & \alpha_0 + \alpha_1 HSSC_{i,t} + \alpha_2 Hotspot_{i,t} + \alpha_3 HSSC_{i,t} \times Hotspot_{i,t} + \alpha_4 \sum Control_{i,t} \\ & + \gamma_l + \eta_l + \varepsilon_{l,l} \end{aligned} \quad (5)$$

$$\begin{aligned} Clique_{i,t} = & \beta_0 + \beta_1 HSSC_{i,t} + \beta_2 Value_{i,t} + \beta_3 HSSC_{i,t} \times Value_{i,t} + \beta_4 \sum Control_{i,t} \\ & + \gamma_2 + \eta_2 + E_{2,2} \end{aligned} \quad (6)$$

Where $Clique_{i,t}$ represents the institutional investor clique and is measured by $CliqueHold_{i,t}$, $CliqueTop_{i,t}$, and $CliqueHerf_{i,t}$, respectively. $HSSC_{i,t}$ is a dummy variable equal to 1 if the listed firm is eligible for trading after stock market liberalization, and 0 otherwise. $Hotspot_{i,t}$ is the natural logarithm of 1 plus the total number of positive news reports on company i by online

media in the year t , capturing the stock's hotspot degree. $Value_{i,t}$ is the moving weighted average of return on equity of stock i over the previous three years, capturing the stock's fundamental value. The interaction term of $HSSC \times Hotspot$ assesses the impact of liberalization on clique behavior toward hotspot stocks, while the interaction term of $HSSC \times Value$ captures the impact toward value stocks. We control for size, age, leverage ratio, revenue growth, turnover, largest shareholding ratio, management ownership, board independence, CEO duality, and state ownership. We also include firm fixed effects and year fixed effects. Variable definitions refer to Table 2.

Table 2. Variable Definition

<i>Variable</i>	<i>Definition</i>
<i>Variables used in the baseline regression and robustness checks</i>	
<i>Clique</i>	The institutional investor clique that is represented by $CliqueHold_{i,t}$, $CliqueTop_{i,t}$, $CliqueHerf_{i,t}$ separately. $CliqueHold_{i,t}$ is the total fraction of the firm owned by all cliques. $CliqueTop_{i,t}$ represents the total ownership share held by the single clique with the largest stake in the firm. $CliqueHerf_{i,t}$ is the sum of the squared ownership fractions of all cliques with holdings in the firm.
<i>HSSC</i>	A dummy variable equal to one 1 if the listed firm is eligible for trading under either the SH-HK or SZ-HK Stock Connect Program, and 0 otherwise.
<i>Hotspot</i>	The natural logarithm of 1 plus the total positive news about firms.
<i>Value</i>	The moving weighted average of return on equity for the previous three years.
<i>Size</i>	The natural logarithm of a company's total asset at last fiscal year end.
<i>Lev</i>	Company's leverage, calculated as the ratio of total liabilities to total assets at last fiscal year end.
<i>Growth</i>	The growth rate of total sales at the end of last fiscal year.
<i>Age</i>	The natural logarithm of the number of days since the company has listed plus 1.
<i>Turnover</i>	The annual average monthly turnover rate of the firm.
<i>Mhold</i>	The percentage of outstanding shares held by the CEO at the end of last fiscal year.
<i>Indd</i>	The percentage of independent directors in the board at the end of last fiscal year.
<i>Dual</i>	A dummy variable equal to one if the CEO is also chairman of the board and zero otherwise.
<i>Top1</i>	The number of shares held by the largest shareholder divided by the total shares.

SOE	A dummy variable equals one if the firm is the state-owned company and zero otherwise.
Hh_Hv	Hh_Hv equals 1 if the stock belongs to both a high hotspot and high value group and zero otherwise.
Hh_Lv	Hh_Lv equals 1 if the stock belongs to both a high hotspot and low value group and zero otherwise.
Lh_Hv	Lh_Hv equals 1 if the stock belongs to both a low hotspot and high value group and zero otherwise.
WebSearch	The natural logarithm of 1 plus the aggregated annual internet search index. The search key terms include the stock code, company abbreviation, and full company name of listed companies in China.
Tbq	Tobin's Q value equals market capitalization divided by total assets.
Other variables used in the additional analyses	
Bubble	We rank all sample stocks in each period based on their price-to-earnings ratio (PE), price-to-book ratio (PB), and price-to-sales ratio (PS) in ascending order. Bubble equals the average rank of each stock across these three variables to determine its level of speculation.
Indleader	We sort all sample stocks in each period based on their industry proportions of market capitalization (MV), revenue (Rev), and net commercial credit from suppliers (Credit) in ascending order. Indleader equals the average rank of each stock across these three variables to measure its industry status.
Misprice	The Amihud illiquidity captures the degree of stock mispricing.
InstNum	The natural logarithm of the number of institutional investor network groups at the firm level plus 1, which captures the collusion motive among institutional investors.
SYN	The stock price synchronization measures the information content of the stock.
Holdual	The weighted average of the duration (Unit: Year) that institutional investor clique member <i>i</i> holds stock <i>j</i> continuously.
CAR5_5	The cumulative excess returns calculated around days [-5,5] near the announcement date of the annual report.
BHAR	The annual buy-and-hold abnormal returns.

4. Empirical analyses

4.1. Descriptive statistics

Table 3 presents the descriptive statistics. HSSC has a mean of 0.250, indicating that Stock

Connect-eligible firms account for 25% of the sample. The mean values of *CliqueHold*, *CliqueTop*, and *CliqueHerf* are 4.1%, 2.5%, and 0.2%, respectively. These relatively low shareholding proportions suggest limited direct voting power for cliques. Consequently, they may rely on trading activities to fulfill their governance role.

Table 3. Descriptive Statistics

<i>Variable</i>	<i>N</i>	<i>Mean</i>	<i>S.D.</i>	<i>Min</i>	<i>Median</i>	<i>Max</i>
<i>CliqueHold</i>	17020	0.041	0.060	0.000	0.015	0.528
<i>CliqueTop</i>	17020	0.025	0.037	0.000	0.010	0.299
<i>CliqueHerf</i>	17020	0.002	0.007	0.000	0.000	0.111
<i>HSSC</i>	17020	0.250	0.433	0.000	0.000	1.000
<i>Hotspot</i>	17020	4.359	1.044	1.946	4.344	7.258
<i>Value</i>	17020	0.069	0.084	-0.309	0.073	0.265
<i>Size</i>	17020	22.238	1.347	19.742	22.065	26.060
<i>Lev</i>	17020	0.432	0.207	0.050	0.421	0.916
<i>Growth</i>	17020	0.183	0.449	-0.576	0.106	3.038
<i>Age</i>	17020	7.864	1.080	3.807	8.135	9.146
<i>Turnover</i>	17020	0.560	0.570	0.045	0.369	3.155
<i>Mhold</i>	17020	0.110	0.191	0.000	0.001	0.685
<i>Indd</i>	17020	0.384	0.106	0.000	0.375	0.667
<i>Dual</i>	17020	0.264	0.441	0.000	0.000	1.000
<i>Top1</i>	17020	0.359	0.148	0.088	0.341	0.750
<i>SOE</i>	17020	0.394	0.489	0.000	0.000	1.000

4.2. Baseline results

4.2.1. Hypothesis testing

Table 4 presents the impact of stock market liberalization on the characteristics of domestic institutional investors' clique behavior. In Columns (1)-(3), the coefficients on $HSSC \times Hotspot$ are statistically negative at 5% or lower. In Columns (4)-(6), the coefficients on $HSSC \times Value$ are statistically positive at 5% or lower. These results indicate that stock market liberalization reduces the speculative tendencies of the domestic capital market, encouraging institutional investors to form value-based cliques rather than hotspot-driven ones. This finding aligns with hypothesis H(b).

Table 4. Baseline Results

	(1)	(2)	(3)	(4)	(5)	(6)
	<i>CliqueHold</i>	<i>CliqueTop</i>	<i>CliqueHerf</i>	<i>CliqueHold</i>	<i>CliqueTop</i>	<i>CliqueHerf</i>
<i>HSSC×Hotspot</i>	-0.0089*** (-6.784)	-0.0033*** (-3.917)	-0.0005** (-2.209)			
<i>HSSC×Value</i>				0.0381** (2.397)	0.0354*** (3.742)	0.0054*** (2.830)
<i>HSSC</i>	0.0403*** (6.434)	0.0181*** (4.473)	0.0030*** (2.893)	-0.0050** (-2.516)	-0.0005 (-0.391)	0.0002 (0.909)
<i>Hotspot</i>	0.0148*** (18.283)	0.0068*** (13.772)	0.0010*** (10.920)			
<i>Value</i>				0.0796*** (8.804)	0.0357*** (6.402)	0.0042*** (4.658)
<i>Size</i>	0.0085*** (5.891)	0.0048*** (5.445)	0.0003** (2.186)	0.0080*** (5.297)	0.0043*** (4.726)	0.0003* (1.775)
<i>Lev</i>	-0.0307*** (-5.758)	-0.0203*** (-6.014)	-0.0027*** (-4.579)	-0.0194*** (-3.519)	-0.0144*** (-4.142)	-0.0019*** (-3.290)
<i>Age</i>	0.0006 (0.501)	-0.0009 (-1.053)	-0.0003 (-1.626)	0.0028** (2.155)	0.0000 (0.037)	-0.0001 (-0.770)
<i>Growth</i>	0.0045*** (4.897)	0.0023*** (3.890)	0.0003*** (3.002)	0.0053*** (5.687)	0.0028*** (4.578)	0.0004*** (3.656)
<i>Turnover</i>	-0.0103*** (-8.674)	-0.0056*** (-6.869)	-0.0008*** (-5.444)	-0.0048*** (-4.212)	-0.0029*** (-3.807)	-0.0004*** (-3.090)
<i>Mhold</i>	0.0088* (1.915)	0.0062** (2.041)	0.0009 (1.600)	0.0122*** (2.662)	0.0077** (2.561)	0.0011** (2.061)
<i>Indd</i>	0.0074 (1.617)	0.0040 (1.334)	0.0003 (0.568)	0.0069 (1.489)	0.0037 (1.224)	0.0003 (0.509)
<i>Dual</i>	0.0018 (1.191)	0.0015 (1.516)	0.0001 (0.837)	0.0022 (1.402)	0.0016* (1.672)	0.0002 (0.963)
<i>Top1</i>	0.0125 (1.194)	0.0126* (1.942)	0.0024* (1.931)	0.0063 (0.592)	0.0092 (1.428)	0.0020 (1.584)
<i>SOE</i>	-0.0096 (-1.620)	-0.0043 (-1.328)	0.0000 (0.045)	-0.0097 (-1.607)	-0.0043 (-1.332)	0.0000 (0.020)
<i>_cons</i>	-0.1744*** (-5.401)	-0.0829*** (-4.160)	-0.0045 (-1.417)	-0.1285*** (-3.870)	-0.0572*** (-2.794)	-0.0013 (-0.403)
<i>Firm_FE</i>	Yes	Yes	Yes	Yes	Yes	Yes
<i>Year_FE</i>	Yes	Yes	Yes	Yes	Yes	Yes
<i>N</i>	17020	17020	17020	17020	17020	17020
<i>Adj. R2</i>	0.102	0.068	0.029	0.085	0.061	0.024

Notes: Standard errors are corrected by clustering at firm level and t-statistic reported in the parentheses. ***, **, and * denote statistical significance at 0.01, 0.05, and 0.1, respectively. The same applies to all subsequent tables unless otherwise noted.

4.2.2. Consider both hotspot and value

Certain stocks exhibit characteristics of both hotspot and value. We classify the sample into four groups based on the industry-year median of Hotspot and Value, and define three corresponding dummy variables: Hh_Hv equals 1 for high-hotspot and high-value stocks; Hh_Lv equals 1 for high-hotspot and low-value stocks; Lh_Hv equals 1 for low-hotspot and high-value stocks. The low-hotspot and low-value group (Lh_Lv) serves as the benchmark. We

interact these three dummies with HSSC in Model (7) to examine differential effects across groups.

$$\begin{aligned} \text{Clique}_{i,t} = & \alpha_0 + \alpha_1 \text{HSSC}_{i,t} + \alpha_2 \text{Hh_Hv}_{i,t} + \alpha_3 \text{Hh_Lv}_{i,t} + \alpha_4 \text{Lh_Hv}_{i,t} \\ & + \alpha_5 \text{HSSC}_{i,t} \times \text{Hh_Hv}_{i,t} + \alpha_6 \text{HSSC}_{i,t} \times \text{Hh_Lv}_{i,t} + \alpha_7 \text{HSSC}_{i,t} \times \text{Lh_Hv}_{i,t} \\ & + \alpha_8 \sum \text{Control}_{i,t} + \gamma_i + \eta_t + \varepsilon_{i,t} \end{aligned} \quad (7)$$

In Table 5, the coefficient on $\text{HSSC} \times \text{Hh_Lv}$ is significantly negative, while that on $\text{HSSC} \times \text{Lh_Hv}$ is significantly positive. The coefficient on $\text{HSSC} \times \text{Hh_Hv}$ is significantly negative only in Column (1). These results indicate that stock market liberalization significantly reduces institutional investor clustering in high-hotspot, low-value stocks while promoting clustering in low-hotspot, high-value stocks, yet has no significant effect on high-hotspot, high-value stocks.

Table 5. Further Differentiation in the Main Test

	(1) <i>CliqueHold</i>	(2) <i>CliqueTop</i>	(3) <i>CliqueHerf</i>
<i>HSSC</i> × <i>Hh_Hv</i>	-0.0122*** (-3.432)	-0.0022 (-0.969)	-0.0006 (-1.075)
<i>HSSC</i> × <i>Hh_Lv</i>	-0.0170*** (-6.019)	-0.0087*** (-4.731)	-0.0015*** (-3.840)
<i>HSSC</i> × <i>Lh_Hv</i>	0.0062* (1.838)	0.0040* (1.794)	0.0007* (1.717)
<i>HSSC</i>	0.0066** (2.375)	0.0049*** (2.632)	0.0012*** (2.857)
<i>Hh_Hv</i>	0.0265*** (13.304)	0.0109*** (8.662)	0.0015*** (6.460)
<i>Hh_Lv</i>	0.0096*** (6.713)	0.0045*** (4.738)	0.0006*** (4.105)
<i>Lh_Hv</i>	0.0041** (2.305)	0.0018 (1.561)	-0.0002 (-0.758)
<i>_cons</i>	-0.1362*** (-4.294)	-0.0635*** (-3.223)	-0.0018 (-0.546)
<i>Controls</i>	Yes	Yes	Yes
<i>Firm_FE</i>	Yes	Yes	Yes
<i>Year_FE</i>	Yes	Yes	Yes
<i>N</i>	17020	17020	17020
<i>Adj. R²</i>	0.100	0.067	0.029

Notes: Control variables are included in all specifications but suppressed in the table to conserve space; full results are available from the corresponding author upon request. The same applies to all subsequent tables unless otherwise noted.

4.3. Alleviating endogeneity problems

4.3.1. Propensity score matching

To address potential self-selection bias, we conduct a robustness check using Propensity Score Matching (PSM). Following Chen et al. [53]; Xiong et al. [40], we match each treated firm with a non-treated firm using baseline control variables. Specifically, we apply 1-to-1

nearest neighbor matching with a 0.01 caliper and no replacement. After matching, no significant difference remains and bias falls below 10%, indicating comparable pairs. We then re-estimate the baseline models using the matched sample. Table 6 shows that the coefficients on $HSSC \times Hotspot$ are significantly negative at the 5% level or lower in Columns (1)–(3), while those on $HSSC \times Value$ are significantly positive at the 1% level in Columns (4)–(6). These findings suggest that the baseline results are robust to sample selection bias.

Table 6. Robustness: Baseline Regression Using Propensity Score Matched Sample

	(1) <i>CliqueHold</i>	(2) <i>CliqueTop</i>	(3) <i>CliqueHerf</i>	(4) <i>CliqueHold</i>	(5) <i>CliqueTop</i>	(6) <i>CliqueHerf</i>
<i>HSSC × Hotspot</i>	-0.0056*** (-2.862)	-0.0027** (-2.023)	-0.0005** (-2.169)			
<i>HSSC × Value</i>				0.0711*** (2.993)	0.0552*** (3.499)	0.0106*** (3.628)
<i>HSSC</i>	0.0279*** (3.225)	0.0160*** (2.765)	0.0036*** (3.366)	-0.0017 (-0.552)	0.0004 (0.186)	0.0005 (1.442)
<i>Hotspot</i>	0.0135*** (11.654)	0.0069*** (8.901)	0.0009*** (6.341)			
<i>Value</i>				0.0560*** (4.393)	0.0279*** (3.290)	0.0019 (1.177)
<i>_cons</i>	-0.2187*** (-4.626)	-0.1302*** (-4.129)	-0.0131** (-2.220)	-0.1674*** (-3.395)	-0.1005*** (-3.070)	-0.0100 (-1.640)
<i>Controls</i>	Yes	Yes	Yes	Yes	Yes	Yes
<i>Firm_FE</i>	Yes	Yes	Yes	Yes	Yes	Yes
<i>Year_FE</i>	Yes	Yes	Yes	Yes	Yes	Yes
<i>N</i>	6866	6866	6866	6866	6866	6866
<i>Adj. R²</i>	0.445	0.366	0.319	0.434	0.360	0.316

4.3.2. Parallel trend test

To draw valid causal inferences, the DID model must satisfy the parallel trend assumption. We use the model below to examine the dynamic effects of the stock market liberalization on the domestic institutional investors' clique behavior, focusing on hotspot and value stock clustering separately.

$$\begin{aligned}
 CliqueHold_{i,t} = & \beta_0 + \alpha_1 Pre_{-k} \times Treated \times Hotspot + \alpha_2 Post_k \times Treated \times Hotspot \\
 & + \alpha_3 Pre_{-k} + \alpha_4 Post_k + \alpha_5 Hotspot + \gamma Control_{i,t} \\
 & + \sum Firm + \sum Year + \varepsilon_{i,t}
 \end{aligned} \tag{8}$$

$$\begin{aligned}
 CliqueHold_{i,t} = & \beta_0 + \beta_1 Pre_{-k} \times Treated \times Value + \beta_2 Post_k \times Treated \times Value \\
 & + \beta_3 Pre_{-k} + \beta_4 Post_k + \beta_5 Value + \gamma Control_{i,t} \\
 & + \sum Firm + \sum Year + \varepsilon_{i,t}
 \end{aligned} \tag{9}$$

where $Pre-k/Postk$ ($T = 0, 1, \text{ or } 2$) equals 1 if the year is T year(s) before/after the

implementation year, and 0 otherwise. The estimates of α_1 and β_1 are insignificant, indicating no significant difference in clique behavior between the treated and control groups before the stock market liberalization. Additionally, the estimates of α_3 , α_4 , β_3 , and β_4 suggest the stock market liberalization significantly impacts the clique behavior of domestic institutional investors after the first year of policy implementation. Figure 3 presents the parallel trend test, supporting the validity of the parallel trend assumption.

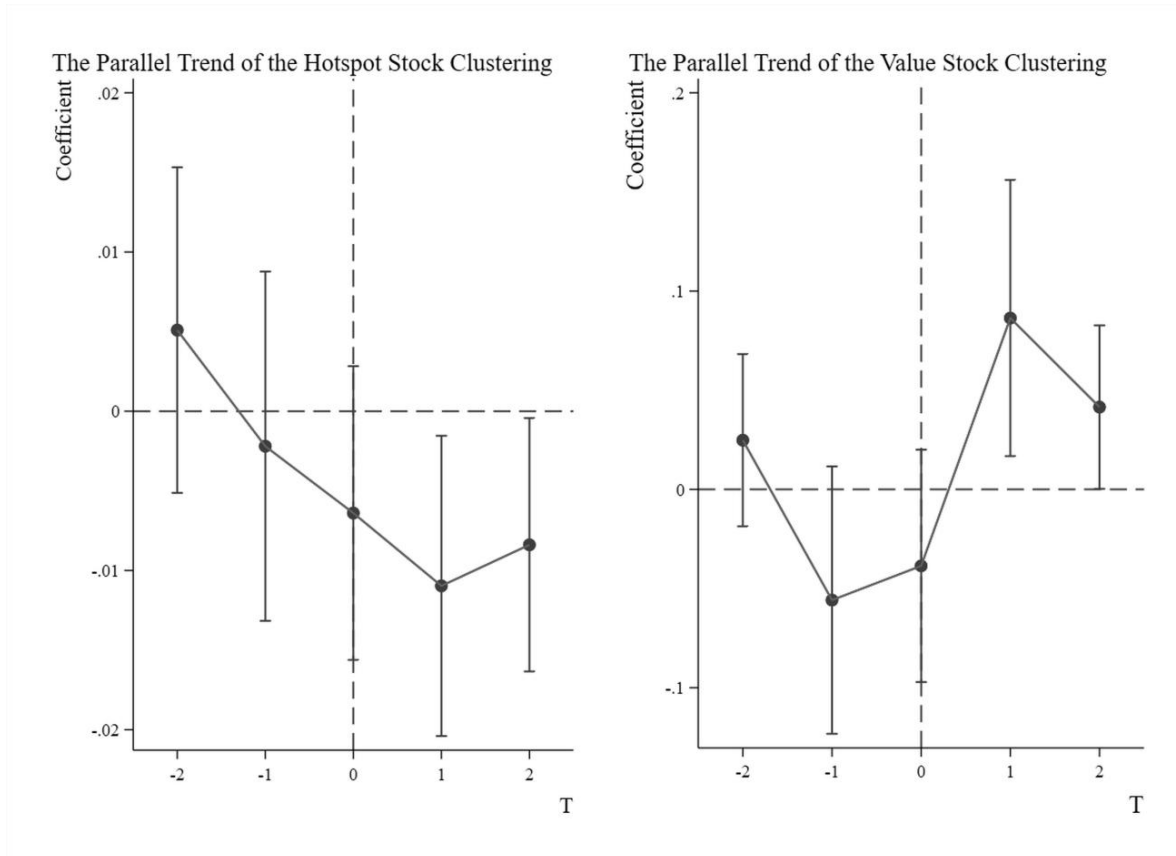


Figure 3. Parallel trend assumption.

Notes: This figure presents the point estimates and 95% confidence intervals for the dynamic effects of the stock market liberalization on the clique behavior of domestic institutional investors concerning hotspot and value stocks individually.

4.4. Other robustness checks

We conduct robustness tests through sample filtering, variable substitution, and network criteria adjustments. First, to address potential pre-existing foreign investor influence, we exclude QFII-held samples and companies with simultaneous 'A+H shares' listings. The results are presented in Columns (1) and (2) of Table 7. Second, we substitute the hotspot and value measures. Hotspot stocks are identified using the aggregated annual internet search index (WebSearch), including the stock code, company abbreviation, and full company name. Firm value is measured by Tobin's Q (Tbq). The results are shown in Columns (3) and (4) of Table

7.

Table 7. Other Robustness Checks: Sample Filtering and Alternative Measures

	<i>(1)CliqueHold</i>	<i>(2)CliqueHold</i>	<i>(3)CliqueHold</i>	<i>(4)CliqueHold</i>
<i>HSSC×Hotspot</i>	-0.0089*** (-5.952)			
<i>HSSC×Value</i>		0.0705*** (7.618)		
<i>HSSC×WebSearch</i>			-0.1462*** (-5.485)	
<i>HSSC×Tbq</i>				0.0070*** (5.241)
<i>HSSC</i>	0.0396*** (5.606)	-0.0063*** (-2.972)	0.3843*** (5.447)	-0.0159*** (-5.444)
<i>Hotspot</i>	0.0144*** (16.927)			
<i>Value</i>		0.0540*** (3.048)		
<i>WebSearch</i>			-0.0282 (-1.475)	
<i>Tbq</i>				0.0104*** (18.774)
<i>_cons</i>	-0.1989*** (-6.004)	-0.1513*** (-4.459)	-0.1241** (-2.352)	-0.3652*** (-10.864)
<i>Controls</i>	Yes	Yes	Yes	Yes
<i>Firm_FE</i>	Yes	Yes	Yes	Yes
<i>Year_FE</i>	Yes	Yes	Yes	Yes
<i>N</i>	15106	15106	16186	16632
<i>Adj. R²</i>	0.105	0.088	0.081	0.130

Third, following Edmans and Holderness [12], we recompute clique ownership by lowering the holdings threshold. To mitigate the impact of network construction criteria, we redefine the network by requiring two institutional investors to each hold at least 3% of the same company's tradable shares. The results are presented in Table 8.

Table 8. Other Robustness Checks: Institutional Investors Clique Connectivity (3%)

	(1) <i>CliqueHold</i>	(2) <i>CliqueTop</i>	(3) <i>CliqueHerf</i>	(4) <i>CliqueHold</i>	(5) <i>CliqueTop</i>	(6) <i>CliqueHerf</i>
<i>HSSC×Hotspot</i>	-0.0090*** (-6.439)	-0.0036*** (-4.294)	-0.0005*** (-2.590)			
<i>HSSC×Value</i>				0.0461*** (2.694)	0.0229** (2.369)	0.0035* (1.808)
<i>HSSC</i>	0.0417*** (6.136)	0.0181*** (4.423)	0.0030*** (2.962)	-0.0051** (-2.337)	-0.0008 (-0.629)	0.0001 (0.186)
<i>Hotspot</i>	0.0160*** (18.186)	0.0072*** (13.533)	0.0011*** (11.478)			
<i>Value</i>				0.0873*** (8.697)	0.0395*** (6.357)	0.0055*** (5.095)
<i>_cons</i>	-0.2205*** (-6.190)	-0.0977*** (-4.583)	-0.0079** (-2.162)	-0.1687*** (-4.624)	-0.0737*** (-3.370)	-0.0047 (-1.235)
<i>Controls</i>	Yes	Yes	Yes	Yes	Yes	Yes
<i>Firm_FE</i>	Yes	Yes	Yes	Yes	Yes	Yes
<i>Year_FE</i>	Yes	Yes	Yes	Yes	Yes	Yes
<i>N</i>	17020	17020	17020	17020	17020	17020
<i>Adj. R2</i>	0.099	0.073	0.035	0.083	0.066	0.029

5. Additional analyses

This section further investigates the heterogeneity and underlying mechanisms through which stock market liberalization affects the clique behavior of domestic institutional investors.

5.1. Heterogeneity test

We further examine the types of clustered stocks that are either reduced or increased following stock market liberalization. Following Dass et al. [54], Bali et al. [55], and Liu et al. [56], we rank all sample stocks based on their price-to-earnings ratio (PE), price-to-book ratio (PB), and price-to-sales ratio (PS) in ascending order for each period. The average rank determines the speculation level (Bubble), with higher values indicating greater speculation, as shown in equation (10). Similarly, we rank stocks by their industry proportions of market capitalization (MV), revenue (Rev), and net commercial credit from suppliers (Credit) to measure industry status (Indleader), as shown in equation (11). Higher values indicate stronger industry leadership. Here, N is the total number of stocks in the period. Group regression is conducted by classifying stocks into bubble and non-bubble groups, as well as industry leaders and non-leaders, based on the annual median values of Bubble and Indleader.

$$Bubble_{i,t} = \frac{1}{3}(PE_N/N + PB_N/N + PS_N/N) \quad (10)$$

$$Indleader_{i,t} = \frac{1}{3}(MV_N/N + Rev_N/N + Credit_N/N) \quad (11)$$

Columns (1) and (2) of Table 9 show that stock market liberalization effectively reduces hotspot clustering for both bubble and non-bubble stocks. The coefficient of HSSC×Hotspot reveals that this weakening effect is more pronounced for bubble stocks, with a significant difference between groups at the 5% level. Columns (3) and (4) indicate that domestic institutional investors exhibit a stronger preference for value clustering with industry-leading companies after stock market liberalization. The coefficient of HSSC×Value shows a significant between-group difference at the 5% level. This suggests that foreign capital entry leads domestic institutional investors to focus more on core assets such as industry leaders.

Table 9. Additional Test: The Clustered Stock Types

	(1) <i>Bubble CliqueHold</i>	(2) <i>Non-bubble CliqueHold</i>	(3) <i>Indleader CliqueHold</i>	(4) <i>Non-Indleader CliqueHold</i>
<i>HSSC×Hotspot</i>	-0.0133*** (-5.342)	-0.0074*** (-4.657)		
<i>HSSC×Value</i>			0.0611** (2.161)	0.0015 (0.076)
<i>HSSC</i>	0.0544*** (4.576)	0.0311*** (3.897)	-0.0033 (-0.946)	-0.0009 (-0.315)
<i>Hotspot</i>	0.0106*** (9.892)	0.0158*** (12.548)		
<i>Value</i>			0.0372*** (3.952)	0.1190*** (7.358)
<i>_cons</i>	-0.1554*** (-3.419)	-0.2499*** (-4.405)	-0.1380*** (-2.982)	-0.0039 (-0.056)
<i>Controls</i>	Yes	Yes	Yes	Yes
<i>Firm_FE</i>	Yes	Yes	Yes	Yes
<i>Year_FE</i>	Yes	Yes	Yes	Yes
<i>Diff (Empirical P-value)</i>		0.020**		0.040**
<i>N</i>	8434	8586	8817	8203
<i>Adj. R²</i>	0.093	0.115	0.089	0.119

5.2. Potential mechanisms

5.2.1. Mechanism on hotspot clustering

By evaluating long-term growth potential and selling overvalued stocks, foreign investors integrate private information into prices, driving them toward intrinsic value [53]. We argue that foreign capital inflows reduce mispricing in hotspot stocks, thereby curbing domestic institutional investors' hotspot clustering behavior. Using the Amihud illiquidity measure [60] to capture market mispricing (Misprice), Column (1) of Table 10 shows a significantly negative coefficient on HSSC×Hotspot. This indicates that stock liberalization reduces mispricing in hotspot stocks and suppresses hotspot clustering.

Previous literature suggests that medium- to large-sized shareholders may collude to

expropriate other shareholders' interests [57]. When institutional investors' holdings closely align with those of controlling shareholders, collusion likelihood increases [58]. Higher concentration of institutional networks at the firm level strengthens their price influence and enhances collusion potential with major shareholders, thereby increasing the likelihood of stock price bubbles. Following Attig et al. [59], we calculate the number of institutional network groups at the firm level to capture collusion potential (InstNum). Column (2) of Table 10 reveals a significant negative coefficient on $HSSC \times Hotspot$. This suggests that liberalization reduces hotspot clustering by mitigating collusion incentives within institutional investor networks.

5.2.2. Mechanism on value clustering

Following foreign capital entry, experienced foreign investors with standardized approaches effectively assess stock values. This improves price information integration and promotes long-term holding of value stocks.

Table 10. Mechanism Test

	(1) <i>Misprice</i>	(2) <i>InstNum</i>	(3) <i>SYN</i>	(4) <i>Holdual</i>
<i>HSSC</i> × <i>Hotspot</i>	-0.1011*** (-3.768)	-0.0416*** (-6.087)		
<i>HSSC</i> × <i>Value</i>			0.5209** (2.107)	0.0115*** (2.985)
<i>HSSC</i>	0.4199*** (2.970)	0.2116*** (5.802)	-0.0183 (-0.521)	0.0037*** (5.470)
<i>Hotspot</i>	0.1695*** (6.616)	0.0944*** (14.607)		
<i>Value</i>			-0.8003*** (-5.289)	-0.0012 (-0.445)
<i>_cons</i>	23.8442*** (21.628)	-4.5200*** (-18.069)	1.1291** (2.193)	0.0287*** (2.976)
<i>Controls</i>	Yes	Yes	Yes	Yes
<i>Firm_FE</i>	Yes	Yes	Yes	Yes
<i>Year_FE</i>	Yes	Yes	Yes	Yes
<i>N</i>	17020	17020	17020	17020
<i>Adj. R2</i>	0.440	0.174	0.236	0.027

Following Morck et al. [61], we use stock price synchronization (SYN) to measure information content, where a higher value indicates lower synchronization and greater information content. Additionally, following Cremers et al. [62], we calculate the weighted average holding duration of clique members (Holdual). In Column (3) and Column (4) of Table 10, the coefficients on $HSSC \times Value$ are significantly positive. This suggests that stock market liberalization enhances the information quality of value stocks and extends holding periods, both of which stimulate value clustering among domestic institutional investors.

5.3. The returns from clique behavior

We further examine whether clique behavior improves portfolio returns through reduced hotspot clustering and increased value clustering after stock market liberalization. We measure short-term returns using cumulative abnormal returns (CAR5_5), calculated as excess returns over the [-5, 5] window around the annual report announcement. For long-term returns, we employ annual buy-and-hold abnormal returns (BHAR). Table 11 shows that the coefficients on $HSSC \times CliqueHold$ are significantly positive, indicating that stocks with higher clique ownership yield higher returns following liberalization. The reduction in hotspot clustering and the increase in value clustering after stock liberalization contribute to these higher returns.

Table 11. The Returns from Clique Behavior

	(1)CAR5_5	(2)BHAR
<i>HSSC×CliqueHold</i>	0.1467*** (3.232)	0.3877** (2.166)
<i>HSSC</i>	-0.0035 (-0.784)	0.0067 (0.458)
<i>CliqueHold</i>	0.0394 (1.382)	1.5744*** (14.378)
<i>_cons</i>	-0.2155** (-2.441)	0.7702** (2.373)
<i>Controls</i>	Yes	Yes
<i>Firm_FE</i>	Yes	Yes
<i>Year_FE</i>	Yes	Yes
<i>N</i>	17020	17020
<i>Adj. R²</i>	0.061	0.429

6. Conclusions and implications

This study examines how stock market liberalization shapes the clique behavior of domestic institutional investors, distinguishing between hotspot and value clustering. To enhance construct validity, we complement network-based clique identification with large language model (LLM)-based textual analysis of institutional disclosures. We further construct a cross-border thematic similarity index from matched domestic and foreign research reports to proxy information spillover channels. Using a staggered difference-in-differences model and Chinese listed firms from 2011 to 2019, we find that stock market liberalization reduces hotspot clustering while promoting value clustering. Specifically, stock market liberalization significantly reduces clustering in high-hotspot, low-value stocks and promotes clustering in low-hotspot, high-value stocks, yet has no significant effect on high-hotspot, high-value stocks. In terms of stock characteristics, stock market liberalization mitigates hotspot clustering in speculative stocks while fostering value clustering in industry leaders. These effects are driven by improvements in market efficiency and institutional holding structures. Domestic

institutional investors achieve enhanced returns as they shift their investment strategy.

This paper presents three implications. First, stock market liberalization stimulates information exchange and transactional competition between Chinese and foreign investors. Foreign capital's value discovery and trading strategies may pressure domestic institutional investors to improve their capabilities and shift toward fundamental-based trading. The LLM-based textual evidence of narrative convergence documented in this study corroborates such cross-border social learning. Second, efforts should focus on enhancing institutional holding structures within China's capital market. Strengthening institutional investors' influence can help mitigate speculative tendencies, such as over-concentration in high-hotspot stocks, and promote more rational, value-oriented practices. Third, given that liberalization promotes value clustering in industry-leading firms, policymakers should consider targeted incentives to support innovation-driven companies. This could further attract domestic and foreign institutional investors to long-term value investments, which are crucial for China's economic transformation and technological advancement.

References

- [1] Gupta, N., & Yuan, K. (2009). On the growth effect of stock market liberalizations. *The Review of Financial Studies*, 22(11), 4715-4752.
- [2] Mitton, T. (2006). Stock market liberalization and operating performance at the firm level. *Journal of Financial Economics*, 81(3), 625-647.
- [3] Chen, Y., Huang, J., Li, X., & Yuan, Q. (2022). Does stock market liberalization improve stock price efficiency? Evidence from China. *Journal of Business Finance and Accounting*, 49(7-8), 1175-1210.
- [4] Yoon, A. S. (2021). The role of private disclosures in markets with weak institutions: Evidence from market liberalization in China. *The Accounting Review*, 96(4), 433-455.
- [5] Zhang, L., Li, Z., Liao, Y., Wang, Y., & Hu, N. (2023). Foreign investment and information quality: A quasi-experiment from China. *International Review of Financial Analysis*, 90, 102796.
- [6] Chan, M. K., & Kwok, S. (2017). Risk-sharing, market imperfections, asset prices: Evidence from China's stock market liberalization. *Journal of Banking and Finance*, 84, 166-187.
- [7] Huang, J., Chen, Y., Dai, X., & Ni, X. (2022). Stock market liberalization and corporate cash holdings: Evidence from China. *Accounting and Finance*, 62, 1925-1955.
- [8] Moshirian, F., Tian, X., Zhang, B., & Zhang, W. (2021). Stock market liberalization and innovation. *Journal of Financial Economics*, 139(3), 985-1014.
- [9] Meng, Y., Xiong, L., Xiao, L., & Bai, M. (2023). The effect of overseas investors on local market efficiency: Evidence from the Shanghai/Shenzhen–Hong Kong Stock Connect. *Financial Innovation*, 9(1), 1-32.
- [10] Choi, N., & Skiba, H. (2015). Institutional herding in international markets. *Journal of Banking and Finance*, 55, 246-259.
- [11] Tan, L., Chiang, T. C., Mason, J. R., & Nelling, E. (2008). Herding behavior in Chinese stock markets: An examination of A and B shares. *Pacific-Basin Finance Journal*, 16(1-

- 2), 61-77.
- [12] Edmans, A., & Holderness, C. G. (2017). Blockholders: A survey of theory and evidence. *The Handbook of the Economics of Corporate Governance*, 1, 541-636.
- [13] McCahery, J. A., Sautner, Z., & Starks, L. T. (2016). Behind the scenes: The corporate governance preferences of institutional investors. *The Journal of Finance*, 71(6), 2905-2932.
- [14] Chen, X., Li, Q., Cui, W., & Hu, Y. (2023). Institutional investor network embedding and firms' total factor productivity. *Emerging Markets Finance and Trade*, 59(6), 1896-1918.
- [15] Rossi, A. G., Blake, D., Timmermann, A., Tonks, I., & Wermers, R. (2018). Network centrality and delegated investment performance. *Journal of Financial Economics*, 128(1), 183-206.
- [16] Li, J., Zhou, Z. Q., Zhang, Y., & Xiong, X. (2023). Information interaction among institutional investors and stock price crash risk based on multiplex networks. *International Review of Financial Analysis*, 89, 102780.
- [17] Park, J., Sani, J., Shroff, N., & White, H. (2019). Disclosure incentives when competing firms have common ownership. *Journal of Accounting and Economics*, 67(2-3), 387-415.
- [18] Baik, B., Kang, J. K., & Kim, J. M. (2010). Local institutional investors, information asymmetries, and equity returns. *Journal of Financial Economics*, 97(1), 81-106.
- [19] Shue, K. (2013). Executive networks and firm policies: Evidence from the random assignment of MBA peers. *The Review of Financial Studies*, 26(6), 1401-1442.
- [20] Faleye, O., Kovacs, T., & Venkateswaran, A. (2014). Do better-connected CEOs innovate more? *Journal of Financial and Quantitative Analysis*, 49(5-6), 1201-1225.
- [21] Granovetter, M. (1985). Economic action and social structure: The problem of embeddedness. *American Journal of Sociology*, 91(3), 481-510.
- [22] David, P., Duru, A., Lobo, G. J., Maharjan, J., & Zhao, Y. (2022). Threat of exit by non-blockholders and income smoothing: Evidence from foreign institutional investors in Japan. *Contemporary Accounting Research*, 39(2), 1358-1388.
- [23] Yang, B., Guo, C., & Fan, Y. (2024). Institutional investor networks and ESG performance: Evidence from China. *Emerging Markets Finance and Trade*, 60(1), 113-137.
- [24] Pan, N., Xu, Q., & Zhu, H. (2021). The impact of investor structure on stock price crash sensitivity: Evidence from China's stock market. *Journal of Management Science and Engineering*, 6(3), 312-323.
- [25] Barber, B. M., & Odean, T. (2008). All that glitters: The effect of attention and news on the buying behavior of individual and institutional investors. *The Review of Financial Studies*, 21(2), 785-818.
- [26] Li, W., Rhee, G., & Wang, S. S. (2017). Differences in herding: Individual vs. institutional investors. *Pacific-Basin Finance Journal*, 45, 174-185.
- [27] Mei, J., Scheinkman, J. A., & Xiong, W. (2005). Speculative trading and stock prices: Evidence from Chinese AB share premia. *Annals of Economics and Finance*, 10(2), 225-255.
- [28] Abreu, D., & Brunnermeier, M. K. (2002). Synchronization risk and delayed arbitrage. *Journal of Financial Economics*, 66(2-3), 341-360.
- [29] Abreu, D., & Brunnermeier, M. K. (2003). Bubbles and crashes. *Econometrica*, 71(1), 173-204.
- [30] Jang, J., & Kang, J. (2019). Probability of price crashes, rational speculative bubbles, and the cross-section of stock returns. *Journal of Financial Economics*, 132(1), 222-247.
- [31] Marcoux, M., & Lusseau, D. (2013). Network modularity promotes cooperation. *Journal of Theoretical Biology*, 324, 103-108.
- [32] Liao, C., Tang, G., & Xu, X. (2021). Smart money or chasing stars: Evidence from

- northbound trading in China. *International Journal of Finance and Economics*, 29(2), 1781-1803.
- [33] Gu, M., Kang, W., & Xu, B. (2018). Limits of arbitrage and idiosyncratic volatility: Evidence from China stock market. *Journal of Banking and Finance*, 86, 240-258.
- [34] Ng, L., & Wu, F. (2007). The trading behavior of institutions and individuals in Chinese equity markets. *Journal of Banking and Finance*, 31(9), 2695-2710.
- [35] Schuppli, M., & Bohl, M. T. (2010). Do foreign institutional investors destabilize China's A-share markets? *Journal of International Financial Markets, Institutions and Money*, 20(1), 36-50.
- [36] Edelen, R. M., Ince, O. S., & Kadlec, G. B. (2016). Institutional investors and stock return anomalies. *Journal of Financial Economics*, 119(3), 472-488.
- [37] Ramalingegowda, S., Utke, S., & Yu, Y. (2021). Common institutional ownership and earnings management. *Contemporary Accounting Research*, 38(1), 208-241.
- [38] Scharfstein, D. S., & Stein, J. C. (1990). Herd behavior and investment. *The American Economic Review*, 80(1), 465-479.
- [39] Alldredge, D. M. (2020). Institutional trading, investor sentiment, and lottery-like stock preferences. *Financial Review*, 55(4), 603-624.
- [40] Xiong, L., Deng, H., & Xiao, L. (2021). Does stock market liberalization mitigate litigation risk? Evidence from Stock Connect in China. *Economic Modelling*, 102, 105581.
- [41] Li, K. (2020). Does information asymmetry impede market efficiency? Evidence from analyst coverage. *Journal of Banking and Finance*, 118, 105856.
- [42] Aggarwal, R., Erel, I., Ferreira, M., & Matos, P. (2011). Does governance travel around the world? Evidence from institutional investors. *Journal of Financial Economics*, 100(1), 154-181.
- [43] Huo, R., & Ahmed, A. D. (2017). Return and volatility spillovers effects: Evaluating the impact of Shanghai-Hong Kong Stock Connect. *Economic Modelling*, 61, 260-272.
- [44] Chen, Z., Du, J., Li, D., & Ouyang, R. (2013). Does foreign institutional ownership increase return volatility? Evidence from China. *Journal of Banking and Finance*, 37(2), 660-669.
- [45] Chuang, Y.-W., Lin, Y.-F., & Weng, P.-S. (2019). Why and how do foreign institutional investors outperform domestic investors in futures trading: Evidence from Taiwan. *Journal of Futures Markets*, 39(3), 279-301.
- [46] Bajo, E., Chemmanur, T. J., Simonyan, K., & Tehranian, H. (2016). Underwriter networks, investor attention, and initial public offerings. *Journal of Financial Economics*, 122(2), 376-408.
- [47] Crane, A. D., Koch, A., & Michenaud, S. (2019). Institutional investor cliques and governance. *Journal of Financial Economics*, 133(1), 175-197.
- [48] Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10), P10008.
- [49] Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance*, 62(3), 1139-1168.
- [50] Mavruk, T. (2021). Cross-regional differences in news tone and local stock ownership. *Management Science*, 67(5), 3276-3298.
- [51] Gao, K., Pittman, J., Wang, X., & Wang, Z. T. (2023). Stock market liberalization and earnings management: Evidence from a quasi-natural experiment in China. *Contemporary Accounting Research*, 40(4), 2547-2576.
- [52] Pfeffer, J., & Salancik, G. R. (2003). *The external control of organizations: A resource dependence perspective*. Stanford University Press.
- [53] Chen, Y., Han, H., Shen, X., & Zeng, J. (2023). The effect of stock market liberalization

- on corporate cash savings sensitivity: Evidence from China. *International Review of Economics and Finance*, 88, 1247-1271.
- [54] Dass, N., Massa, M., & Patgiri, R. (2008). Mutual funds and bubbles: The surprising role of contractual incentives. *The Review of Financial Studies*, 21(1), 51-99.
- [55] Bali, T. G., Cakici, N., & Whitelaw, R. F. (2011). Maxing out: Stocks as lotteries and the cross-section of expected returns. *Journal of Financial Economics*, 99(2), 427-446.
- [56] Liu, B., Wang, H., Yu, J., & Zhao, S. (2020). Time-varying demand for lottery: Speculation ahead of earnings announcements. *Journal of Financial Economics*, 138(3), 789-817.
- [57] Zwiebel, J. (1995). Block investment and partial benefits of corporate control. *The Review of Economic Studies*, 62(2), 161-185.
- [58] Cai, C. X., Hillier, D., & Wang, J. (2016). The cost of multiple large shareholders. *Financial Management*, 45(2), 401-430.
- [59] Attig, N., El Ghouli, S., & Guedhami, O. (2009). Do multiple large shareholders play a corporate governance role? Evidence from East Asia. *Journal of Financial Research*, 32(4), 395-422.
- [60] Amihud, Y. (2002). Illiquidity and stock returns: Cross-section and time-series effects. *Journal of Financial Markets*, 5(1), 31-56.
- [61] Morck, R., Yeung, B., & Yu, W. (2000). The information content of stock markets: Why do emerging markets have synchronous stock price movements? *Journal of Financial Economics*, 58(1-2), 215-260.
- [62] Cremers, M., Pareek, A., & Sautner, Z. (2020). Short-term investors, long-term investments, and firm value: Evidence from Russell 2000 index inclusions. *Management Science*, 66(10), 4535-4551.

Where Vision Meets Language: A Coupling-Locus Taxonomy of Multimodal Large Language Models

Haosong Hu,Xinyu Song*

Geely University of China, ChengDu, China, 641423

Received: Sept 23, 2026

Revised: Sept 25, 2026

Accepted: Sept 28, 2026

Published online: September 28, 2026

To appear in: *International Journal of Advanced AI Applications*, Vol. 2, No. 10 (October 2026)

* Corresponding Author: Xinyu Song (songxinyu@guc.edu.cn)

Abstract. Multimodal large language models have moved from research prototypes to the default interface of frontier AI systems, and their architectural literature has grown faster than its syntheses. This survey organizes that literature around a single design question every system answers implicitly: at what point does visual information bind to the language model? The answer defines four families—input-side connectors joining frozen encoders to language models, token-space coupling that discretizes images into the model's own vocabulary, weight-space early fusion inside a jointly trained backbone, and output-side orchestration of generative components—each with a characteristic trade-off between modularity, binding fidelity, and data cost. Building on this coupling-locus taxonomy, the survey traces the field from 2020 to 2025 across training recipes, resolution and token engineering, mixture-of-experts scaling, and the benchmark landscape. The synthesis is a mismatch the taxonomy exposes: integration has matured faster than perception, and documented failure modes—chart misreading, counting and geometry failures, hallucinated objects—persist across all four loci because they originate upstream of the binding point. The survey closes with consequences for evaluation practice and the open question of whether native multimodal training or engineered perception will close the gap.

Keywords: *multimodal large language models; vision-language alignment; coupling locus; early fusion; visual instruction tuning; benchmark evaluation*

1. Introduction

1.1. The Integration Paradox

By late 2024, the flagship multimodal systems could take a screenshot of a technical report,

answer questions about its contents, and produce both prose and images in a single interleaved response [1-3]. The integration agenda—the question of whether a model can accept images, video, and audio as input and emit more than one modality as output—has, by any reasonable reading, been settled: every frontier lab ships a model that does so, and the open-source ecosystem reproduces the recipe at smaller scale. Yet the same systems, evaluated carefully, fail at perception questions that most humans answer without deliberation. A diagnostic suite of seven visual skills—counting line intersections, measuring relative angles, identifying nested shapes—finds that frontier VLMs average 58.07%, with the strongest at 77.84%, on tasks humans solve almost perfectly [4]. On realistic chart questions that require reading values out of an axis rather than recognizing chart style, the strongest proprietary model trails non-expert humans by roughly thirty points on the reasoning split (47.1% versus 80.5%) [5]. A benchmark of image pairs that CLIP-style encoders cannot distinguish reveals models that answer fluently, explain themselves plausibly, and perform at or below chance on the underlying perceptual discrimination [6].

We call the coexistence of these two facts the integration paradox: the modality interface has been engineered to maturity faster than the visual competence behind it, so the flagship systems of 2025 are simultaneously the most integrated and among the least perceptually reliable systems ever fielded at scale. The paradox matters because the two facts are routinely conflated in system comparisons—integration benchmarks are read as evidence of perception—which in turn shapes where the community spends its effort. This survey's purpose is to organize the architectural literature so that the two claims can be evaluated separately.

1.2. Scope and Method

We focus on architecture and its evidence: how multimodal systems are built, what each design commits to, and what the published measurements say those commitments deliver. Training data, alignment, and evaluation are covered where they bear on architecture, but the survey is not a catalogue of models. We searched arXiv and the proceedings of the major machine-learning and computer-vision venues for 2020–2025, prioritizing (i) papers that introduce a system or an architectural mechanism with published evaluations, (ii) papers that introduce or reinterpret benchmarks measuring those systems, and (iii) surveys against which this one differentiates itself. System claims are reported from the system's own paper; benchmark figures are reported from the benchmark's paper. Approximate figures are marked as approximate, and no number in this survey is quoted from a secondary summary. Where a design's behavior is contested across papers—for example, whether Q-Former-style bottlenecks

discard or merely compress information—we report both positions and mark the question open. This discipline produces a bibliography weighted toward primary sources: technical reports for industry systems, venue proceedings for academic systems, and benchmark papers for all quantitative comparisons.

1.3. Positioning and the Coupling-Locus Framework

The existing surveys organize the field along two other axes. Yin et al. organize by capability: what models can do—visual understanding, visual generation, multimodal agents—and list representative systems per capability [7]. Li et al. organize by component and pipeline: what goes into a system—encoders, connectors, language models, training stages—and enumerate design choices within each slot [8]. Both are inventories; neither makes a claim about why the architecture space looks the way it does or where its failures concentrate. Table 1 contrasts the three framings.

Table 1. Coverage comparison between recent surveys and this work.

Dimension	Yin et al. [7]	Li et al. [8]	This survey
Organizing axis	Capabilities (understanding, generation, agents)	Pipeline components (encoder, connector, LLM)	Locus of cross-modal binding
Temporal structure	Snapshot	Snapshot	Four generations, 2020–2025
Analytical claim	None (inventory)	None (inventory)	Modularity–fidelity trade-off; coupling $\neq$ perception
Failure evidence	Examples	Examples	Perception benchmarks across all loci

Our organizing question is the one every design answers implicitly: at what point, and in what representation, does visual information bind to the language model? We call this the coupling locus, and it takes four values. In input-side coupling, a vision encoder and a language model are trained separately and joined by a trainable bridge at the input interface—the connector era. In token-space coupling, visual content is discretized into tokens of the same vocabulary the language model uses, so binding happens in the shared sequence. In weight-space coupling, modalities are fused inside a backbone trained jointly from the start—early fusion. In output-side coupling, the language model binds to other modalities not as an input interface but as a controller that invokes generative components. The four loci form a historical sequence—each generation migrated the binding point deeper into the model—but they also form a trade-off space, and Section 3 argues the trade-off is real: modularity at the input buys data efficiency and reuse at the cost of binding fidelity; fusion buys fidelity at the cost of data

and compute; orchestration buys interface flexibility at the cost of error compounding.

The taxonomy is not merely descriptive, because it buys two analytical results. First, it converts the field's headline progress into a measurable trajectory: the binding point has moved from a bolted-on projection (2021–2022) through the sequence (2023) into the weights (2024–), and each move was motivated by a specific failure of the previous locus, which Section 3 traces. Second, and unexpectedly, the taxonomy exposes what has not changed: the perception failures documented in 2024–2025—chart misreading, counting, geometry, hallucinated objects—appear in systems from all four loci, because they originate in the encoder and the pretraining data, upstream of wherever the binding happens. Integration architecture, however deep, does not repair a representation that lost the information before binding. This is the survey's synthesis: the coupling locus explains integration maturity; it does not explain, and does not fix, perception fidelity. Sections 4–7 cover the machinery that surrounds the locus—training recipes, resolution and token engineering, efficient scaling, and the benchmark landscape—and Section 8 returns to the synthesis.

2. Two Lineages and the Coupling Problem

2.1. The Vision Lineage

The modern visual encoder descends from the Vision Transformer, which established that pure transformer attention over image patches matches convolutional networks at scale and, crucially, emits a fixed set of patch embeddings consumable by any other transformer [9]. Contrastive image-text pretraining then supplied the semantics: CLIP trained a dual encoder on hundreds of millions of noisy web pairs so that images and their captions share an embedding space, yielding zero-shot classification and, incidentally, the visual backbone most MLLMs still use [10]. ALIGN confirmed the recipe scales with even noisier data at larger volume [11]. A complementary line asked what self-supervised objectives could learn without captions: masked image modeling reconstructs patches from visible ones [12], and DINOv2 distilled large-scale self-supervision into features that rival weakly supervised ones for dense tasks [13]. By 2022 the field possessed high-quality frozen vision encoders and, with large caption datasets such as LAION-5B and curated ones such as DataComp, the data to train them [14,15].

Two properties of this lineage shaped everything downstream. First, the encoders were global: CLIP's contrastive objective optimizes a single vector per image, which is efficient for retrieval and demonstrably lossy for fine detail—a property later work would trace directly into MLLM perception failures (Section 7). Second, the encoders were frozen by default: their quality was

expensive to produce, so the design question became how to reuse them rather than retrain them, which is precisely what set the coupling agenda.

A third property was inherited less consciously: the data regime. CLIP's supervision was web alt-text—short, noisy, and optimized for the gist of a photograph—and the same regime propagated into the caption and interleaved corpora that would later feed MLLM pretraining. What the objective and the data jointly optimized was co-occurrence between a scene and its description, not the extraction of values, glyphs, or spatial relations from structured content. The blind-spot analyses of Section 7 are, in an important sense, audits of this inheritance: image pairs that differ in exactly the details an alt-text regime never needed to encode [6]. The vision lineage thus delivered to the coupling problem a mature, frozen, semantically competent encoder—whose specific losses became the new field's specific failures.

2.2. The Language Lineage

On the text side, the decisive facts were scale and instruction-following. GPT-3 demonstrated that a decoder-only transformer scaled up exhibits few-shot in-context learning, making the pretrained language model a substrate that adapts to new tasks without weight updates [16]. FLAN showed that instruction tuning—fine-tuning on tasks phrased as instructions—made that adaptation far more reliable [17], and InstructGPT showed that reinforcement learning from human feedback aligned the resulting behavior with user intent [18]. The consequence for multimodality was an economic one: by 2022 there existed expensive, general, instruction-following language models, and the obvious question was whether visual capability could be grafted onto them rather than trained from scratch. The same in-context learning that made the language model adaptable made it plausibly grafted-able—if a model can learn a task from examples in its context, perhaps it can learn a modality there too. Flamingo would test exactly that hypothesis [19].

2.3. The Coupling Problem

Between the two lineages stands a representational obstacle. Liang et al. analyzed the embedding spaces of aligned multimodal encoders and found that, despite their semantic alignment, image and text embeddings occupy two disjoint cones in the shared space—the modality gap—so that naively concatenating representations puts the modalities in regions neither was trained to process [20]. The gap formalizes why coupling is a design problem rather than a wiring exercise: what is needed is not an interface but a translation, learned somewhere, and every MLLM since 2021 is an answer to the question of where to learn it. Three sub-

decisions parameterize the answer. Where does the translation happen (input interface, shared vocabulary, or shared weights)? What stays frozen during translation (the encoder, the language model, both, neither)? And what data supervises it (paired captions, interleaved documents, or task instructions)? The four loci of Section 3 are the four characteristic answers; the generations of the field are the record of how the answers migrated under evidence.

3. The Coupling-Locus Taxonomy

Table 2 summarizes the four loci before the subsections elaborate them; Figure 1 places their representative systems on a 2020–2025 timeline, making the field-scale migration of the binding point from input connector toward jointly trained backbones visible generation by generation. The table's last two columns preview the synthesis of Section 3.5: each locus buys something specific and pays something specific, and the payments—unlike the gains—tend to survive the next generation.

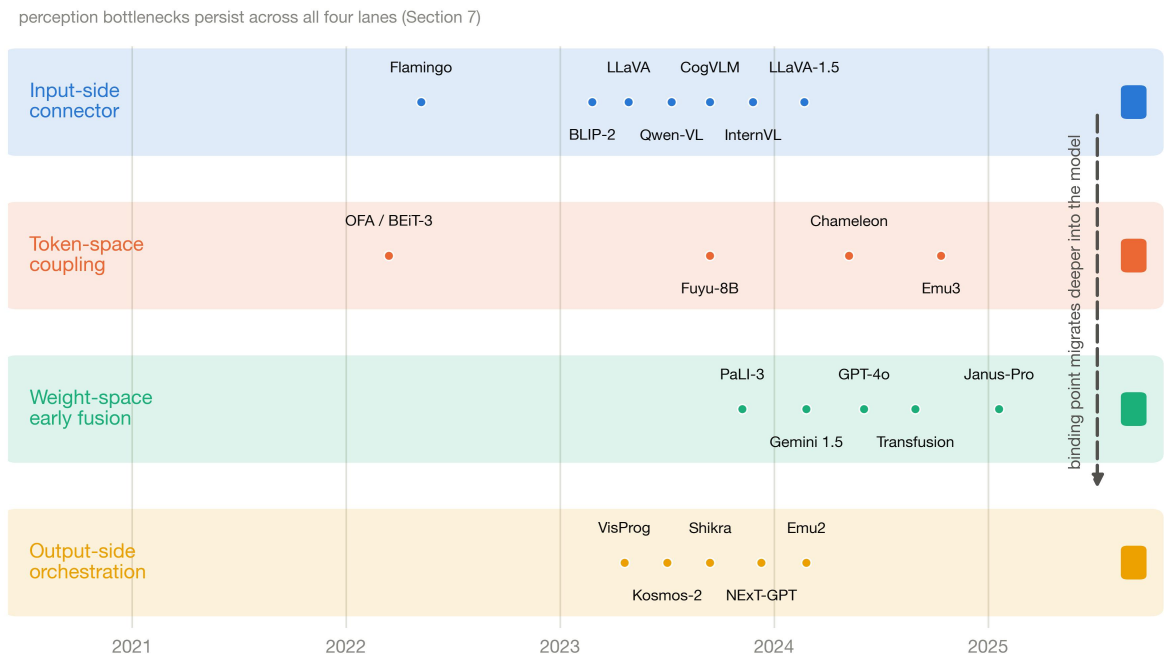


Figure 1. The coupling-locus taxonomy as a timeline (2020–2025). Each lane collects representative systems for one locus; over the period the binding point migrates from input-side connectors toward weight-space early fusion, while the perception benchmarks of Section 7 fail systems in all four lanes alike.

Table 2. The coupling-locus taxonomy: four answers to where vision binds to language.

Locus	Binding mechanism	Representative systems	Typical frozen components	Characteristic strength	Characteristic cost
Input-side connector	Trainable bridge at the LM input	Flamingo, BLIP-2, LLaVA, InstructBLIP, Qwen-VL, CogVLM	Encoder and/or LM (varies)	Data efficiency; LLM reuse	Resolution ceiling; modality gap; weak visual grounding
Token-space coupling	Discrete image tokens in the LM vocabulary	OFA, BEiT-3, Fuyu-8B, Chameleon, Emu3	Varies; often none after pretraining	Unified sequence; native generation	Token count explosion; discretization information loss
Weight-space coupling	Joint training of a multimodal backbone	PaLI/PaLI-3, Gemini, Transfusion, Show-o, Janus-Pro	Nothing (or late-stage only)	Binding fidelity; omni-interaction	Data and compute cost; capability entanglement
Output-side orchestration	LM as controller over generators	Emu2 pipelines, NExT-GPT, VisProg	LM and components both	Interface flexibility; any-to-any	Error compounding across components

3.1. Input-Side Coupling: The Connector Era

The connector era's founding design is Flamingo: freeze a powerful language model, freeze a vision encoder, and insert gated cross-attention layers into the language model so that it can attend to visual features sampled by a Perceiver Resampler, training only the new layers and the resampler on interleaved web data [19]. The gates initialize at zero so the model starts as pure text and grows into vision, and the frozen language model retains its few-shot ability—Flamingo demonstrated multimodal few-shot learning on interleaved sequences. The cost profile was set by the era's ambition: an 80-billion-parameter system whose trainable fraction was small, trained on web-scale interleaved data available to few.

BLIP-2 attacked the data problem with an explicit bridge module, the Q-Former, which distills encoder features into a fixed small set of learned query embeddings and is trained in two stages—first to align with the frozen encoder on captioning objectives, then to connect with the frozen language model—so that instruction-tuned chat behavior emerged at roughly two orders of magnitude less trainable capacity than Flamingo [21]. LLaVA attacked from the opposite direction with radical minimalism: a single linear projection mapping encoder features into the language model's embedding space, with the language model unfrozen, and—its real contribution—a data method that used GPT-4 itself to convert caption pairs into instruction-following conversations, a recipe reproducible by any lab with a projection layer [22]. Its

successor fixed the obvious deficiencies (higher resolution, academic-task data, better prompts) and showed the recipe competitive with far larger systems at 13B parameters [23]. InstructBLIP industrialized the instruction-tuning stage across 26 datasets [24], and MiniGPT-4 and mPLUG-Owl spread the pattern through the open-source community [25,26].

The era's most consequential export, however, was not any single system but the recipe's portability. Because LLaVA-style coupling requires only a projection layer, an open encoder, and an open language model, the recipe diffused through the open-source ecosystem within a year—MinGPT-scale reproductions, domain variants, and multilingual derivatives multiplied—establishing "LLaVA-style" as the default substrate of academic multimodal research. This democratization had an architectural consequence the frontier literature rarely states: the overwhelming majority of published multimodal results, including most benchmark numbers cited in this survey's Section 7, were produced on connector-era substrates, so the field's collective self-image—its sense of what works and what fails—is disproportionately a portrait of the input-side locus. Claims that a failure is general to MLLMs should be read with that provenance in mind, and Section 7's cross-locus comparisons are an attempt to supply exactly that reading.

The era then consumed its own budget of frozenness. Qwen-VL made the visual resampler trainable and supported variable resolutions [27]; CogVLM unfroze the language model's deeper layers with trainable "visual expert" modules inserted into them, arguing that shallow adapters leave the language model unable to use visual detail [28]; InternVL scaled the encoder itself to 6 billion parameters and aligned it tightly with the language model, treating encoder capacity as the binding constraint [29]. Read as a sequence, the connector era's history is a steady retreat from its founding economy: each system trained more of the stack because each frozen layer was measured to be a bottleneck somewhere downstream.

The connector locus's structural costs explain both the retreat and what came next. The first is a resolution ceiling: connectors consume fixed-size encoder outputs, so a 336×336 encoder input—adequate for "what is in this photo"—destroys the fine structure of documents, charts, and screenshots, and Section 5 describes the engineering arms race this triggered. The second is the modality gap inherited from the encoder: contrastively pretrained embeddings sit in their own cone [20], and a projection can move them but not regenerate what the encoder discarded. The third cost is empirical and the most consequential for later chapters: connector-era models proved systematically prone to object hallucination—describing objects that are not present—when probed adversarially, a failure the POPE protocol tied to language priors dominating

visual evidence [30]. None of these costs is a bug in a particular system; all three are properties of learning the translation after and outside the representations being translated.

3.2. Token-Space Coupling: Images in the Vocabulary

The second locus answers the modality gap by removing the interface altogether: represent visual content as tokens of the same discrete vocabulary the language model already uses, so that binding happens inside the ordinary sequence. The lineage begins in the encoder-decoder generation: OFA and Unified-IO cast image understanding, generation, and text tasks as sequence-to-sequence problems over unified vocabularies [31,32], and BEiT-3 consolidated the program under the slogan "image as a foreign language," training a single backbone over masked data in both modalities [33]. Fuyu-8B pushed the simplification to its limit for the interface: rather than an encoder, raw image patches are linearly projected straight into the decoder, making resolution a configuration choice rather than an architectural constraint and suiting the model to UI screenshots and diagrams [34].

The contemporary wave couples tokenization with autoregressive generation over discrete image tokens. Chameleon trains a single transformer on interleaved text (BPE tokens) and images (VQ tokens from a large codebook) with an early-fusion recipe—everything mixed from the start—demonstrating an open model that generates and understands in one vocabulary [3]. The Emu series traces the same locus from the other side: Emu established generative pretraining in multimodality [35], Emu2 scaled continuous embeddings as in-context multimodal tokens to emergent composition [36], and Emu3 closed the loop with pure next-token prediction over discrete visual and textual tokens, arguing that diffusion is not required for visual generation [37].

The locus's costs are the mirror of its elegance. Token count is the first: codebook representations of a single image consume hundreds to thousands of tokens—Chameleon's 1024 tokens per image against tens for a sentence [3]—so context windows and attention budgets fill with visual content, and generation quality is bounded by codebook size against reconstruction fidelity. Token fidelity is the second and deeper cost: discretization is lossy precisely where MLLMs are weakest, in fine detail, small text, and chart-like structure, so the locus trades away exactly the information its hardest benchmarks measure; quantization artifacts in generated images are the visible residue of the same loss. The third cost is economic: unified-vocabulary models cannot inherit a pretrained language model without inheriting its vocabulary conflicts, so most were trained from substantial scratch, paying in data what the connector era paid in capability reuse. The locus thus sells binding depth at a price the field's smaller labs could not

pay—which is why, in practice, it appears at the frontier and in specialized document models rather than as the community default. The document niche deserves its own sentence, because it is where the locus's discretization trade-off was met head-on: chart- and document-understanding models such as ChartVLM and TinyChart couple the unified-vocabulary design with explicit structural intermediates—chart-to-table extraction, guard tokens that check whether a number was actually grounded in the plot, or program-of-thoughts traces that force values to be read before they are computed over [38,39]. That these mitigation mechanisms are needed at all is evidence about the locus: when images are tokens, the model's fluency over them does not guarantee it has read them, and the interface between perception and computation must be re-engineered by hand.

3.3. Weight-Space Coupling: Early Fusion in the Backbone

The third locus moves the binding into the weights: modalities are fused inside a backbone trained jointly, with no interface to learn afterward because there was never a time the modalities were separate. The PaLI lineage supplies the controlled version—joint pretraining over captioning and VQA at multiple scales, with PaLI-3 showing that a compact model over a SigLIP encoder can match far larger predecessors [40]—while the frontier labs moved to native multimodality: Gemini was trained multimodally from the start across text, image, audio, and video [2], and GPT-4o extended the design to real-time audio-visual interaction at conversational latencies [41].

Two research systems make the locus's design space explicit. Transfusion trains a single transformer with two objectives at once—next-token prediction on text and diffusion on image latents—keeping modalities separate in objective but not in parameters [42]. Show-o mixes the two within attention itself, combining autoregressive modeling for text with discrete diffusion for images in one masking-scheme-aware sequence [43]. Janus-Pro addresses the locus's characteristic conflict—understanding wants discriminative features while generation wants reconstruction-fidelity features—by decoupling the encoders (separate visual pathways for the two purposes) while sharing the transformer backbone and autoregressive objective, improving both sides at once [44]. The design illustrates the locus's frontier question: not whether to fuse, but what to share, with evidence accumulating that shallow sharing (encoders) with deep fusion (backbone) is the productive middle.

The locus buys binding fidelity and emergent cross-modal behavior—real-time interleaved conversation, in-context image composition—and pays in the currencies the connector era conserved: data, compute, and interpretability. It also pays in a subtler currency that matters for

the survey's synthesis: with everything trained jointly, attribution becomes harder. When a weight-space model misreads a chart, the failure is not localizable to a bridge or a frozen encoder; the multimodal competence is smeared across the network, and no parameter subset corresponds to "the visual channel" the way a projection layer does. The contrast with Table 3's failure-localization row is deliberate: localization is an engineering property that the connector era provided for free and the fusion era deliberately traded away, and as later sections show, the field's newest failure modes are the ones it can least afford to have made opaque. Section 7 will show the perceptual failures are no smaller here than elsewhere, and the loci's most important common finding is that none of them—not even native fusion—has escaped the encoder-inherited perception bottleneck.

3.4. Output-Side Coupling: The Model as Controller

The fourth locus is defined by direction: visual binding happens at the output interface, with the language model as a controller that plans, delegates to generative components, and reads back their results. NExT-GPT wires an MLLM core to encoders and decoders across text, image, audio, and video with instruction-tuned routing layers between them, demonstrating any-to-any transfer with off-the-shelf components [45]. Visual programming formalizes the controller role as code: the model writes a visual program whose API calls invoke detection, OCR, and VQA modules, executing perception as procedure rather than as generation [46], in the same spirit as tool-orchestration systems built over language models [47]. The locus is also where grounding instruments live: Kosmos-2 binds referring expressions to boxes in the input interface [48], while Shikra and Ferret let the model reason over and emit coordinates in natural language [49,50]—capability work that later becomes mitigation machinery in perception-critical deployments. Visual programming's decomposition idea has proved especially durable on structured content: a chart question answered by a program that first calls an extraction routine and then computes over the extracted values cannot silently substitute a plausible number for a read one, because the substitution point is explicit. The orchestration locus is thus where the field first encountered, in tool form, the principle that perception and computation should be separated at an auditable interface—a principle Section 7 shows the end-to-end models still lack, and whose training-time analogue is developed in the companion survey.

The locus's trade-off is architectural honesty: it does not pretend the language model perceives, and instead arranges for perception to happen where it is reliable. Its costs are the flip side. Component pipelines compound errors—a misread by the OCR module propagates into a confident downstream plan—and the controller's own perception is still that of whatever

core it embeds, so orchestration relocates rather than removes the bottleneck. The historical pattern, then, is not that orchestration lost to fusion; it is that orchestration's components became fusion-era models, and the locus persists as the systems layer around whatever the inner model is.

3.5. Synthesis: The Modularity–Fidelity Trade-off

Read across loci, the field's trajectory is a single variable moving: the depth at which the translation is learned. The connector era learned it at the input interface while conserving the investments on both sides; token-space coupling learned it in the vocabulary while paying in token budgets and discretization loss; weight-space coupling learned it in the weights while paying in data and compute. The movement was evidence-driven at each step—frozen-layer bottlenecks [28], resolution ceilings [23], understanding-generation conflicts [44]—and Table 3 summarizes the accounting.

Table 3. The trade-off accounting across coupling loci.

Property	Input connector	Token-space	Weight-space	Output orchestration
Data efficiency	High (reuses pretrained parts)	Low (usually from scratch)	Lowest (joint from start)	High (all parts pretrained)
Binding fidelity	Limited by frozen representations	Limited by discretization	Highest	Limited by controller's core
Perception ceiling	Encoder-bounded	Encoder/tokenizer-bounded	Encoder/data-bounded	Component-bounded
Generation ability	Added via decoders	Native (same tokens)	Native	Native (by delegation)
Failure localization	Easy (bridge)	Medium	Hard	Per-component
Deployability (2025 practice)	Dominant open-source default	Frontier + document niche	Frontier omni systems	Systems layer everywhere

Two readings of the table matter for what follows. The first is that the loci are converging functionally even while differing architecturally: connector-era leaders unfroze their stacks until the difference from weight-space models is a matter of initialization history, and weight-space systems ship connector-style adapters for deployment efficiency—the locus is becoming a training-recipe variable rather than a species boundary. The second reading is the survey's synthesis: the row that refuses to improve across loci is the perception ceiling. Every locus's ceiling is set by an encoder or a controller core trained with the same contrastive-or-caption data regime, and Section 7 shows the 2024–2025 perception benchmarks measuring precisely that inheritance. The binding point is no longer where the field's failures live; the representation being bound is.

4. Training Recipes Across Loci

Whatever the locus, training has consolidated into a three-stage pipeline: alignment pretraining, instruction tuning, and preference alignment. Alignment pretraining couples the modalities on web-scale data—paired captions (LAION-5B, DataComp) and interleaved documents (OBELISC) supply the two regimes [14,15,51]—with the locus determining what trains: bridge modules in the connector era, everything in the fusion era. Instruction tuning then converts capability into behavior on curated conversational data, the recipe LLaVA demonstrated at small scale and InstructBLIP industrialized across task datasets [22,24]. Preference alignment, the third stage, arrived later and with a distinctive multimodal twist: DPO-style optimization over preference pairs transfers directly [52], but the field's most-cited variants target where in the response the preference attaches. RLHF-V collects human feedback at the level of individual claims—human annotators correct specific hallucinated statements rather than ranking whole responses—and produces substantially more trustworthy responses than response-level preference tuning [53]; RLAI-F-V scales the same claim-level structure with open AI feedback instead of human annotators [54]. The pattern is worth naming because it recurs in Section 7: multimodal alignment work keeps discovering that response-level signals are too coarse for perception errors, and keeps re-inventing finer-grained ones.

A fourth stage is emerging and belongs to this survey only as a pointer: reinforcement learning with verifiable outcome rewards, adapted from text reasoning to multimodal tasks, now trains MLLMs to produce reasoning chains scored by final-answer correctness. It is powerful and, as companion work on reasoning faithfulness argues at length, structurally blind to the difference between right answers reached through correct perception and right answers reached through compensated errors—the subject of the companion survey to this one. For architecture purposes the relevant fact is that the recipe's stages are largely orthogonal to the locus: the same three stages run over connectors, token vocabularies, and fused backbones, which is why Section 3.5 could read the loci as a shared substrate with different binding depths.

The recipe's stages also differ in how much of their effect is attributable to data rather than design, and the honest summary is: mostly data. Instruction-tuning corpora built by prompting a strong teacher model to rephrase captions into conversations (the LLaVA method) made small systems look qualitatively new [22]; subsequent ablations across the connector literature repeatedly find that swapping the instruction corpus moves benchmark scores as much as swapping the coupling mechanism. The pattern generalizes a lesson from the contrastive era—DataComp's systematic search showed dataset curation outweighing architectural variation in

CLIP-style training [15]—and it cautions against over-reading architecture comparisons in this survey's own tables: much of what a locus buys or costs is entangled with the data its recipes happen to ship with. The coupling-locus framework organizes where binding happens; what binds is still decided in the data.

5. Engineering the Interface: Resolution and Token Economics

Between the locus and the recipe sits a layer of engineering the taxonomy explains but does not exhaust: how many visual tokens, at what resolution, enter the model. The connector era's fixed 224–336 pixel inputs were adequate for classification-style understanding and demonstrably inadequate for documents and charts, and three families of fixes emerged. Tiling approaches decompose the image into overlapping high-resolution crops processed independently—LLaVA-NeXT's AnyRes and InternVL 1.5's dynamic aspect-ratio tiling are the reference designs—preserving detail at the cost of token count [55,56]. Native-dynamic-resolution approaches instead make the encoder itself resolution-free: Qwen2-VL accepts arbitrary resolutions and encodes images as variable-length patch sequences with multimodal rotary position embeddings, spending tokens where the image is large rather than where a fixed grid dictates [57], an option NaViT had opened on the encoder side by packing native-resolution patches into training sequences [58]. Sequence-level approaches manage the budget directly: LLaVA-UHD balances compressed global context with spatially divided local detail [59].

The token budget's other side is compression. FastV observed that visual tokens receive disproportionately little attention after the earliest layers—attention to them collapses by the second block—so nearly half can be pruned at inference with negligible quality loss [60]; LLaVA-PruMerge combines pruning with merging of retained tokens [61]; TokenPacker compresses region-level detail into coarse queries with point guidance [62]. The engineering consensus that emerged—dynamic high-resolution input, aggressive downstream compression—reflects a real asymmetry the taxonomy makes visible: input resolution is a perception constraint (the information must be present before binding), while token count is a compute constraint (it only must be present where attention will use it). The asymmetric fix—spend tokens at input, prune them deeper—is thus not a hack but the correct response to where information is destroyed versus merely discounted.

The document-centric branch of the field makes the stakes concrete, and predates the connector era's reckoning with them. OCR-free document understanding began as its own architecture question—Donut trained an encoder-decoder directly on document images to emit

structured text without a recognition front-end [63]—and the pattern generalizes: wherever the content is structured (documents, charts, forms, screenshots), the field has repeatedly found that the interface, not the model, is the bottleneck, and has built the interface back by hand (tiling, dynamic resolution, structural intermediates). OCR-heavy benchmarks (TextVQA, DocVQA, OCRBench) and chart benchmarks (Section 7) are where fixed-resolution connector models fail categorically and where the engineering families above were concentrated; the correlation between resolution engineering and document performance across system generations is one of the cleaner architecture–capability mappings in the literature. What the engineering cannot fix is upstream loss in the encoder: tiling and dynamic resolution preserve pixels, but a contrastive encoder that never learned to represent small glyphs or thin line crossings discards them after they arrive. The next section's scaling question and Section 7's perception evidence meet at exactly this boundary. OCR-heavy benchmarks (TextVQA, DocVQA, OCRBench) and chart benchmarks (Section 7) are where fixed-resolution connector models fail categorically and where the engineering families above were concentrated; the correlation between resolution engineering and document performance across system generations is one of the cleaner architecture–capability mappings in the literature. What the engineering cannot fix is upstream loss in the encoder: tiling and dynamic resolution preserve pixels, but a contrastive encoder that never learned to represent small glyphs or thin line crossings discards them after they arrive. The next section's scaling question and Section 7's perception evidence meet at exactly this boundary.

6. Scaling Efficiency: Mixture-of-Experts and Compact Models

Two scaling stories run beneath the architecture narrative. The first is sparsification: mixture-of-experts transformers activate a fraction of parameters per token, and the MLLM variants—MoE-LLaVA at small scale [64], DeepSeek-VL2 with fine-grained experts and high routing granularity [65], Aria with a MoE language core over a lightweight visual encoder [66], and Kimi-VL activating a few billion of tens of billions of parameters [67]—demonstrate that the coupling recipes transfer to sparse backbones with little re-engineering. MoE matters to the taxonomy in a specific way: by decoupling total capacity from per-token cost, it makes weight-space coupling affordable at scales where dense training would be prohibitive, so sparsification is quietly dissolving the fusion era's principal economic objection.

The second story is compression downward. Phi-3-vision showed a 4B-class model integrating text and vision competitively [68], and MobileVLM brought the full connector recipe to mobile-class compute [69]. The compressed models inherit the connector locus's costs

(resolution ceilings, encoder inheritance) and add a new one worth naming: at small scale, the language model itself is a weaker reasoning substrate, so perception errors that a larger core might repair in-context propagate instead. The efficiency frontier and the perception frontier are therefore not independent: Section 7's evidence about perception failures should be read as lower bounds for the deployed small-model tier that serves most real traffic.

7. Evaluation: Benchmarks and the Perception Bottleneck

7.1. The Benchmark Landscape

Table 4 organizes the evaluation space into families by what they measure. The historical arc across families is consistent: recognition-style benchmarks saturate, integration-style benchmarks saturate at the frontier, and perception-style benchmarks retain large gaps.

Table 4. Benchmark families and the state of their gaps (2020–2025).

Family	Representative benchmarks	What they probe	State (2025)
Recognition VQA	VQAv2	Object/attribute QA over photos	Saturated; no longer discriminative [70] Frontier ~62–78% vs expert humans ~89%;
Knowledge & exams	MMMU, MMMU-Pro	College-level multimodal reasoning	vision-only variants reveal text-shortcut inflation [71,72] ChartQA near-saturated; reasoning splits retain gaps [73,74]
Mathematical/structured	MathVista, ChartQA	Numeric and geometric reasoning over visual structure	
Realistic chart reasoning	CharXiv	Value extraction + compositional reasoning on hard charts	Largest gap in evaluation space: best systems roughly half of non-expert human accuracy [5] Frontier averages 58.07% (best 77.84%) where humans score near 100%; some pairs at chance [4,6]
Perception primitives	BlindTest, MMVP	Counting, geometry, orientation; CLIP-blind pairs	Persistent across model families; improves with alignment but not eliminated [30,75,76] Active frontier; long-video reasoning weakest [77,78]
Hallucination	POPE, HallusionBench, AMBER	Objects absent from image; language-prior overrides	
Video	MVBench, Video-MME	Temporal reasoning over video	

Three findings organize the table. First, saturation is real but informative: VQAv2 and ChartQA-style recognition saturate because they can be solved by the language prior plus

coarse appearance, so their saturation measures the language prior, not perception. Second, the gap ordering is diagnostic: the benchmarks with the largest remaining gaps are precisely those that isolate perception from language priors—BlindTest uses abstract geometry no caption corpus teaches; MMVP builds pairs that CLIP embeddings cannot separate; CharXiv's descriptive-versus-reasoning split separates value extraction from inference over extracted values; HallusionBench repeats questions across image edits to detect prior-driven answering. Third, benchmark design now targets the shortcut: MMMU-Pro forces vision-only input by rendering questions into images, and scores drop substantially once the text shortcut is closed [72]—direct evidence that part of headline multimodal performance was language performance wearing an image.

7.2. The Perception Bottleneck Is Upstream of the Coupling

The survey's synthesis rests on a cross-cutting observation: the perception-style gaps in Table 4 do not sort by coupling locus. Systems with input-side connectors, token-space models, and weight-space frontier models all fail the same primitive-perception suites at qualitatively similar rates, because the failures originate in a component all loci share: the visual encoder trained with contrastive-or-caption objectives on web data. Eyes Wide Shut supplies the mechanism with its CLIP-blind pair analysis—image pairs that are trivially different to humans yet produce near-identical CLIP embeddings, on which MLLMs answer at or below chance while producing fluent explanations [6]; the BlindTest suite generalizes the observation across frontier systems [4]; and "MLLMs know where to look" sharpens it further, showing that in many failure cases the correct visual information is present in the model's internal attention but is not used in the emitted answer, so that simple training-free attention-guided decoding recovers a large share of the errors [79]. The bottleneck, in other words, is not only in what the encoder encodes but in what the coupled system uses—and neither mechanism is touched by moving the binding point deeper.

For the architecture narrative this reframes the generations of Section 3: the migration of the coupling locus solved integration (accepting, mixing, producing modalities), and its remaining agenda—perception—lies elsewhere, in encoder objectives, data regimes, and decoding. The quantitative arc makes the divergence stark. On knowledge-style integration benchmarks, the frontier moved from GPT-4V's 56.8% on MMMU [71] to frontier systems above 75% within roughly a year and a half—a steep, monotonic climb. On the perception suites, the same period produced no comparable movement: BlindTest's geometric primitives and CharXiv's chart-reasoning split were published because the then-frontier failed them, and successor systems

have improved only marginally, with several remaining near chance on the hardest discriminative pairs [4-6]. One curve climbed because integration compounds—the language core underneath grew stronger every quarter; the other stayed flat because perception does not inherit from the language core at all. That asymmetry, not any single number, is the empirical core of this survey. The reframing matters practically because the field's biggest benchmark gaps and its biggest deployment risks (chart misreading in financial and scientific document workflows; hallucinated objects in agentic perception) concentrate exactly where integration maturity provides no protection. It matters intellectually because it separates two claims often made together: that deeper coupling is the path to stronger multimodality (supported, for integration), and that deeper coupling will fix perception (unsupported by the evidence across loci).

8. Beyond Static Images: Video, Audio, and the Omni Horizon

The video line extends the connector recipe to the temporal axis—frame sampling into existing connectors (Video-LLaVA-style designs) plus dedicated temporal benchmarks (MVBench for temporal skills, Video-MME and long-video suites for duration scaling) [77,78]—and inherits every static-image cost plus a new one: token budgets that were tight for one image are prohibitive for thousands of frames, making video the domain where compression engineering (Section 5) is least optional. Generation on the same axis developed separately but converged on the same lesson: Sora's diffusion-transformer treatise on video generation argued that scale over spatiotemporal patches yields emergent simulation capacity, a claim whose evaluation remains contested but whose architectural bet—shared backbone, raw patches, no modality-specific machinery—is the weight-space locus applied to synthesis [80]. The audio line ran the connector playbook over speech and sound (SALMONN) [81]. The omni systems—GPT-4o's real-time audio-visual interaction [41] and Qwen2.5-Omni's Thinker-Talker architecture separating a multimodal "thinker" backbone from streaming speech "talker" heads [82]—represent the weight-space locus extended across modalities, and their native interactivity is the clearest capability dividend fusion has paid. On the survey's central axis, though, the extensions are confirmations: omni models bind more modalities deeper than ever, while the perception primitives of Section 7 remain unsolved upstream—a reminder that the coupling axis and the perception axis of this survey's title are, in the current evidence, independent variables.

9. Open Challenges and Future Directions

The perception bottleneck as the field's core problem. The synthesis of Section 7 sets the agenda: encoder objectives that preserve fine-grained structure (the DINOv2-style dense-feature line is a starting point [13]), data regimes that supervise perception rather than co-occurrence, and decoding that uses what the model encodes [79]. The evidence that errors survive despite present internal information [79] makes this the rare architecture problem where a large fraction of the gap may be recoverable without new pretraining.

Faithfulness of multimodal reasoning. As MLLMs are trained to emit reasoning chains, new failure modes appear at the intersection of perception and generation: fluent chains that cite values never extracted, or justify answers with invented image content. This is the subject of the companion survey; the architectural fact relevant here is that the failure is locus-independent, so mitigation will live in training signals and verification rather than in coupling depth.

Evaluation validity. The shortcut findings (MMMU-Pro's vision-only collapse [72], CharXiv's descriptive-reasoning decomposition [5]) imply headline scores systematically overstate multimodal competence. The open problem is institutional: dynamic, contamination-resistant, process-evaluating benchmarks are more expensive than static suites, and the field's incentive gradients still reward the latter.

The endgame question: unified or modular. Weight-space coupling is winning the integration argument and the frontier, while orchestration persists as the systems layer and connector recipes dominate open-source practice. The evidence of this survey does not settle the endgame; it constrains it: whatever the coupling locus of future systems, perception quality will be set upstream of it, and the field's marginal return on integration depth now appears lower than its marginal return on perception.

Interpretability for multimodal models. Localization of failures (bridge? encoder? decoder? attention?) is tractable in connector-era models and opaque in fused ones; the finding that internal representations often contain the right answer [79] makes multimodal interpretability unusually applied-relevant: the models fail using information they have.

Safety across the interface. The modality interface is also an attack surface: instructions smuggled inside images bypass text-side safeguards, and the system cards that accompany flagship multimodal releases document adversarial-image evaluations as a first-class risk category rather than an afterthought [1]. The taxonomy adds a structural note to the safety discussion: each locus relocates, rather than removes, the trust boundary—connector-era safety lives at the projection between a frozen encoder and an adaptable LM; orchestration spreads it

across components, each with its own failure modes; fusion concentrates it in a network whose failures are, per the interpretability challenge above, hardest to localize. Safety engineering for MLLMs will need to be re-derived per locus, not ported from the text-only playbook.

10. Conclusion

This survey organized multimodal large language models by the design question their literature answers implicitly—where does vision bind to language—and found the answer's history is the field's history: a bolted-on interface (2021), a bridge economy (2022), a shared vocabulary (2023), and fused weights (2024–), each generation buying binding depth with data and compute, each leaving a characteristic residue in what fails and what can be localized when it does. The taxonomy's second yield is the one this survey would most like to outlive it: across all four loci, the perception primitives—reading a value off an axis, counting, discriminating nearly identical scenes—remain the documented failure surface, because they are set by encoders and data upstream of any binding point. Integration has been engineered to maturity; perception has not been engineered at all, only inherited. The next generation's architecture question is therefore not where to bind vision to language, but what is left in vision by the time it arrives—and the tools for that question live in encoder objectives, data regimes, and training signals, which is where this survey's companion survey on reasoning-faithful multimodal models begins.

References

- [1] OpenAI. (2023). GPT-4V(ision) system card. OpenAI Technical Report.
- [2] Gemini Team, Google. (2024). Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. arXiv preprint arXiv:2403.05530.
- [3] Chameleon Team. (2024). Chameleon: Mixed-modal early-fusion foundation models. arXiv preprint arXiv:2405.09818.
- [4] Rahmanzadehgervi, P., Bolton, L., Taesiri, M. R., & Nguyen, A. T. (2024). Vision language models are blind: Failing to translate detailed visual features into words. arXiv preprint arXiv:2407.06581.
- [5] Wang, Z., Xia, M., He, L., Chen, H., Liu, Y., et al. (2024). CharXiv: Charting gaps in realistic chart understanding in multimodal LLMs. arXiv preprint arXiv:2406.18521.
- [6] Tong, S., Liu, Z., Zhai, Y., Ma, Y., LeCun, Y., & Xie, S. (2024). Eyes wide shut? Exploring the visual shortcomings of multimodal LLMs. arXiv preprint arXiv:2401.06209.
- [7] Yin, S., Fu, C., Zhao, Y., Li, K., Sun, X., Xu, T., & Li, E. (2023). A survey on multimodal large language models. arXiv preprint arXiv:2306.13549.
- [8] Li, G., Chen, X., Xu, G., Lin, T., Wang, P., Liang, H., & Zhang, S. (2024). Multimodal large language models: A survey. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW).
- [9] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021).

- An image is worth 16x16 words: Transformers for image recognition at scale. In International Conference on Learning Representations (ICLR).
- [10] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In Proceedings of the 38th International Conference on Machine Learning (ICML).
 - [11] Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., Le, Q., Sung, Y.-H., Li, Z., & Duerig, T. (2021). Scaling up visual and vision-language representation learning with noisy text supervision. In Proceedings of the 38th International Conference on Machine Learning (ICML).
 - [12] He, K., Chen, X., Xie, S., Li, Y., Dollár, P., & Girshick, R. (2022). Masked autoencoders are scalable vision learners. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).
 - [13] Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., et al. (2024). DINOv2: Learning robust visual features without supervision. In Transactions on Machine Learning Research.
 - [14] Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., et al. (2022). LAION-5B: An open large-scale dataset for training next generation image-text models. In Advances in Neural Information Processing Systems 35.
 - [15] Gadre, S. Y., Ilharco, G., Fang, A., Hayase, J., Smock, G., Oh, S., et al. (2023). DataComp: In search of the next generation of multimodal datasets. In Advances in Neural Information Processing Systems 36.
 - [16] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., et al. (2020). Language models are few-shot learners. In Advances in Neural Information Processing Systems 33.
 - [17] Wei, J., Bosma, M., Zhao, V. Y., Guu, K., Yu, A. W., Lester, B., Du, N., Dai, A. M., & Le, Q. V. (2022). Finetuned language models are zero-shot learners. In International Conference on Learning Representations (ICLR).
 - [18] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., et al. (2022). Training language models to follow instructions with human feedback. In Advances in Neural Information Processing Systems 35.
 - [19] Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., et al. (2022). Flamingo: A visual language model for few-shot learning. In Advances in Neural Information Processing Systems 35.
 - [20] Liang, V. W., Zhang, Y., Kwon, Y., Yeung, S., & Zou, J. (2022). Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. In Advances in Neural Information Processing Systems 35.
 - [21] Li, J., Li, D., Savarese, S., & Hoi, S. (2023). BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In Proceedings of the 40th International Conference on Machine Learning (ICML).
 - [22] Liu, H., Li, C., Wu, Q., & Lee, Y. J. (2023). Visual instruction tuning. In Advances in Neural Information Processing Systems 36.
 - [23] Liu, H., Li, C., Li, Y., & Lee, Y. J. (2024). Improved baselines with visual instruction tuning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).
 - [24] Dai, W., Li, J., Li, D., Tiong, A. M. H., Zhao, J., Wang, W., Li, B., Fung, P., & Hoi, S. (2023). InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In Advances in Neural Information Processing Systems 36.
 - [25] Zhu, D., Chen, J., Shen, X., Li, X., & Elhoseiny, M. (2023). MiniGPT-4: Enhancing vision-language understanding with advanced large language models. arXiv preprint

- p arXiv:2304.10592.
-
- [26] Ye, Q., Xu, H., Xu, G., Ye, J., Yan, M., Zhou, Y., Wang, J., Hu, A., Shi, P., Shi, Y., Li, C., Xu, Y., Chen, H., Tian, B., Zhang, J., Li, Z., Zhao, J., & Huang, F. (2023). mPLUG-Owl: Modularization empowers large language models with multimodality. arXiv preprint arXiv:2304.14178.
- [27] Bai, J., Bai, S., Yang, S., Wang, W., Tan, S., Wang, P., Lin, J., Zhou, C., & Zhou, J. (2023). Qwen-VL: A versatile vision-language model for understanding, localization, text reading, and beyond. arXiv preprint arXiv:2308.12966.
- [28] Wang, W., Lv, Q., Yu, W., Hong, W., Qi, J., Wang, Y., et al. (2023). CogVLM: Visual expert for pretrained language models. arXiv preprint arXiv:2311.03729.
- [29] Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., et al. (2024). InternVL: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).
- [30] Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, W. X., & Wen, J.-R. (2023). Evaluating object hallucination in large vision-language models. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP).
- [31] Wang, P., Yang, A., Men, R., Lin, J., Bai, S., Li, Z., Ma, J., Zhou, C., Zhou, J., & Yang, Z. (2022). OFA: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In Proceedings of the 39th International Conference on Machine Learning (ICML).
- [32] Lu, J., Clark, C., Lee, S., Zhang, Z., Khosla, S., Marten, R., Hoiem, D., & Kembhavi, A. (2022). Unified-IO: A unified model for vision, language, and multi-modal tasks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).
- [33] Bao, H., Ye, Q., Li, W., & Dong, L. (2023). BEiT-3: Image as a foreign language: BEiT pretraining for all vision and vision-language tasks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).
- [34] Adept AI. (2023). Fuyu-8B: A multimodal architecture for AI agents. Adept AI Research Blog.
- [35] Sun, Q., Yu, Q., Cui, Y., Zhang, F., Zhang, X., et al. (2023). Generative pretraining in multimodality. arXiv preprint arXiv:2307.05222.
- [36] Sun, Q., Cui, Y., Zhang, X., Zhang, F., Yu, Q., Luo, Z., et al. (2024). Emu2: Generative multimodal models are in-context learners. arXiv preprint arXiv:2312.13286.
- [37] Wang, X., Zhang, X., Luo, Z., Sun, Q., Cui, Y., et al. (2024). Emu3: Next-token prediction is all you need. arXiv preprint arXiv:2409.18869.
- [38] Xia, R., Zhang, B., Ye, H., et al. (2024). ChartX & ChartVLM: A versatile benchmark and foundation model for complicated chart reasoning. arXiv preprint arXiv:2402.12185.
- [39] Zhang, L., Hu, A., Xu, H., Yan, M., Xu, Y., Jin, Q., Zhang, J., & Huang, F. (2024). TinyChart: Efficient chart understanding with visual token merging and program-of-thoughts learning. arXiv preprint arXiv:2404.16635.
- [40] Chen, X., Wang, X., Beyer, L., Kolesnikov, A., Wu, J., Vo, N., et al. (2023). PaLI-3 Vision Language Models: Smaller, faster, stronger. arXiv preprint arXiv:2310.09199.
- [41] OpenAI. (2024). Hello GPT-4o. OpenAI Technical Publication.
- [42] Zhou, C., Yu, L., Babu, A., Tirumala, K., Yasunaga, M., Shams, L., Kahn, J., Ma, X., Zettlemoyer, L., & Levy, O. (2024). Transfusion: Predict the next token and diffuse images with one multi-modal model. arXiv preprint arXiv:2408.11039.
- [43] Xie, J., Mao, W., Bai, Z., Zhang, D. J., Wang, W., Lin, K. Y., Gu, Y., Chen, Z., Yang, L., & Shou, M. Z. (2024). Show-o: One single transformer to unify multimodal understanding and generation. arXiv preprint arXiv:2408.12528.

- [44] Chen, X., Wu, Z., Liu, X., Pan, Z., Liu, W., et al. (2025). Janus-Pro: Unified multimodal understanding and generation with data and model scaling. arXiv preprint arXiv:2501.17811.
- [45] Wu, S., Fei, H., Le, L., Ji, T., & Chua, T.-S. (2024). NExT-GPT: Any-to-any multimodal LLM. In Proceedings of the 41st International Conference on Machine Learning (ICML).
- [46] Gupta, T., & Kembhavi, A. (2023). Visual programming: Compositional visual reasoning without training. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).
- [47] Shen, Y., Song, K., Tan, X., Li, D., Lu, W., & Zhuang, Y. (2024). HuggingGPT: Solving AI tasks with ChatGPT and its friends in Hugging Face. In Advances in Neural Information Processing Systems 36.
- [48] Peng, Z., Wang, W., Dong, L., Hao, Y., Huang, S., Ma, S., & Wei, F. (2023). Kosmos-2: Grounding multimodal large language models to the world. arXiv preprint arXiv:2306.14824.
- [49] Chen, K., Zhang, Z., Zeng, W., Zhang, R., Zhu, F., & Zhao, R. (2023). Shikra: Unleashing multimodal LLM's referential dialogue magic. arXiv preprint arXiv:2306.15195.
- [50] You, H., Zhang, H., Gan, Z., Du, X., Zhang, B., Wang, Z., Cao, L., Chang, S.-F., & Yang, Y. (2024). Ferret: Refer and ground anything anywhere at any granularity. In International Conference on Learning Representations (ICLR).
- [51] Laurençon, H., Trinh, H., de Vazelhes, H., Wang, A., & Truong, V. (2023). OBELISC: An open web-scale filtered dataset of interleaved image-text documents. In Advances in Neural Information Processing Systems 36 Datasets and Benchmarks Track.
- [52] Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., & Finn, C. (2023). Direct preference optimization: Your language model is secretly a reward model. In Advances in Neural Information Processing Systems 36.
- [53] Yu, T., Yao, Y., Zhang, H., He, T., Han, Y., Cui, G., Hu, J., Liu, Z., Zheng, H.-T., Zhao, M., & Sun, M. (2024). RLHF-V: Towards trustworthy MLLMs via behavior alignment from fine-grained correctional human feedback. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).
- [54] Yu, T., Zhang, H., Li, Q., Xu, Q., Yao, Y., et al. (2024). RLAIIF-V: Open-source AI feedback leads to super GPT-4V trustworthiness. arXiv preprint arXiv:2405.17220.
- [55] Liu, H., Li, C., Li, Y., Li, B., Zhang, Y., Shen, S., & Lee, Y. J. (2024). LLaVA-NeXT: Improved reasoning, OCR, and world knowledge. LLaVA Project Blog.
- [56] Chen, Z., Wang, W., Tian, H., Ye, S., Gao, Z., et al. (2024). InternVL 1.5: How far are we to GPT-4V? Closing the gap to commercial multimodal models with open-source suites. arXiv preprint arXiv:2404.16821.
- [57] Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., et al. (2024). Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191.
- [58] Dehghani, M., Djolonga, J., Wilhelm, B., Pintsch, L., Lucioni, A., Arnab, A., et al. (2023). Patch n' pack: NaViT, a vision transformer for any aspect ratio and resolution. In Advances in Neural Information Processing Systems 36.
- [59] Xu, R., Yao, Y., Guo, Z., Cui, J., Ni, Z., et al. (2024). LLaVA-UHD: An LMM perceiving any aspect ratio and high-resolution images. arXiv preprint arXiv:2403.11703.
- [60] Chen, L., Zhao, H., Liu, T., Bai, S., Lin, J., Zhou, C., & Chang, B. (2024). An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In European Conference on Computer Vision (ECCV).
- [61] Shang, Y., Cai, M., Xu, B., Lee, Y. J., & Yan, Y. (2024). LLaVA-PruMerge: Adaptive token reduction for efficient large multimodal models. arXiv preprint arXiv:2403.15388.
- [62] Li, W., You, Z., Qi, C., Wang, H., Zhou, K., Wang, H., Gao, C., Chen, W., & Zhang, Y.

- (2024). TokenPacker: Efficient visual projector for multimodal large language models. arXiv preprint arXiv:2407.02392.
- [63] Kim, G., Hong, T., Yim, M., Nam, J., Park, J., Yim, J., Hwang, W., Yun, S., Han, D., & Park, S. (2022). OCR-free document understanding transformer. In European Conference on Computer Vision (ECCV).
- [64] Lin, B., Tang, Z., Ye, Y., Cui, J., Zhu, B., Jin, P., Zhang, J., Zhuang, M., Yuan, L., Li, S., & Li, Y. (2024). MoE-LLaVA: Mixture of experts for large vision-language models. arXiv preprint arXiv:2401.15947.
- [65] Wu, Z., Chen, X., Pan, Z., Liu, X., Liu, W., Li, D., et al. (2024). DeepSeek-VL2: Mixture-of-experts vision-language models for advanced multimodal understanding. arXiv preprint arXiv:2412.10302.
- [66] Hong, L., Li, Q., Zhang, L., Chen, D., Jiang, Z., Li, Z., et al. (2024). Aria: An open multimodal native mixture-of-experts model. arXiv preprint arXiv:2410.05993.
- [67] Kimi Team. (2025). Kimi-VL technical report. arXiv preprint arXiv:2504.07491.
- [68] Abdin, M., Aneja, S., Behl, H., Bubeck, S., Eldan, R., Gunasekar, S., et al. (2024). Phi-3 technical report: A highly capable language model locally on your phone. arXiv preprint arXiv:2404.14219.
- [69] Chu, X., Xiang, L., Wei, B., Li, X., Lu, X., Wei, N., & Yan, M. (2024). MobileVLM: A fast, reproducible and strong vision language assistant for mobile devices. arXiv preprint arXiv:2312.16886.
- [70] Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., & Parikh, D. (2017). Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
- [71] Yue, X., Ni, Y., Zhang, K., Zheng, T., Liu, R., Zhang, G., et al. (2024). MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).
- [72] Yue, X., Zheng, T., Ni, Y., Wang, Y., Zhang, K., et al. (2024). MMMU-Pro: A more robust multi-discipline multimodal understanding benchmark. arXiv preprint arXiv:2409.02813.
- [73] Lu, P., Bansal, H., Xia, T., Liu, J., Li, C., Hajishirzi, H., Cheng, H., Chang, K.-W., Galley, M., & Gao, J. (2024). MathVista: Evaluating mathematical reasoning of foundation models in visual contexts. In Advances in Neural Information Processing Systems 36 Datasets and Benchmarks Track.
- [74] Masry, A., Long, D. X., Tan, J. Q., Joty, S., & Hoque, E. (2022). ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In Findings of the Association for Computational Linguistics: ACL 2022.
- [75] Guan, T., Liu, F., Wu, X., Xian, R., Li, Z., Liu, X., Wang, X., Chen, L., Huang, F., Yacoob, Y., Manocha, D., & Zhou, T. (2024). HallusionBench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).
- [76] Wang, J., Wang, Y., Xu, G., Zhang, J., Gu, Y., Jia, H., Zhang, J., Yan, M., Zhang, J., & Sang, J. (2024). AMBER: An LLM-free multi-dimensional benchmark for MLLM hallucination evaluation. arXiv preprint arXiv:2311.07397.
- [77] Li, K., Wang, Y., He, Y., Li, Y., Wang, Y., Liu, Y., et al. (2024). MVBench: A comprehensive video benchmark for multi-modal large language models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).
- [78] Fu, C., Chen, P., Shen, Y., Qin, Z., Zhang, D., Lin, X., et al. (2024). Video-MME: The

- first-ever comprehensive evaluation benchmark of multi-modal LLMs in video analysis. arXiv preprint arXiv:2405.21075.
- [79] Zhang, J., Khayatkhoei, M., Chhikara, P., & Ilievski, F. (2025). MLLMs know where to look: Training-free perception of small visual details with multimodal LLMs. arXiv preprint arXiv:2502.17422.
- [80] OpenAI. (2024). Video generation models as world simulators. OpenAI Technical Report.
- [81] Tang, C., Yu, W., Sun, G., Chen, X., Tan, T., Li, W., Lu, L., Ma, Z., & Zhang, C. (2024). SALMONN: Towards generic hearing abilities for large language models. In International Conference on Learning Representations (ICLR).
- [82] Qwen Team. (2025). Qwen2.5-Omni technical report. arXiv preprint arXiv:2503.20215.

Right Answers, Wrong Reasons: Perception Errors in Multimodal Chain-of-Thought and the Blind Spot of Outcome-Based Rewards

Yanwei Zhang, Xinyu Song*

Geely University of China, ChengDu, China, 641423

Received: Sept 22, 2026

Revised: Sept 25, 2026

Accepted: Sept 27, 2026

Published online: September 28, 2026

To appear in: *International Journal of Advanced AI Applications*, Vol. 2, No. 10 (October 2026)

* Corresponding Author: Xinyu Song (songxinyu@guc.edu.cn)

Abstract. Multimodal large language models increasingly answer visual questions through explicit reasoning chains, and the chains are typically fluent. Evidence shows they can nonetheless be perceptually wrong—misreading values off a chart's axis, miscounting geometric primitives, or citing objects absent from the image—while the final answer lands correct anyway. This survey organizes that evidence around a Process–Outcome Consistency (POC) framework: a 2×2 over process soundness and outcome correctness whose key cell, the compensated error (flawed process, correct outcome), is invisible to every metric and training signal that scores only the answer. We trace the evidence for perception errors inside coherent chains, formalize why outcome-based supervision—verifier-reward reinforcement learning, outcome reward models, answer-filtered rejection sampling—cannot select against compensated errors (the reward insensitivity principle), and review the resulting signatures: reward overoptimization, reward hacking, and reasoning that grows more articulate without growing more grounded. We map mitigation onto five leverage points—upstream perception, inference-time process shaping, training-signal redesign, post-hoc verification, process-aware evaluation—and propose reporting compensated-error and evidence-grounding rates alongside accuracy. We close on an asymmetry favoring multimodal settings: visual process errors are checkable by mundane tooling, making faithful multimodal reasoning an unusually tractable target for process supervision.

Keywords: *multimodal chain-of-thought; perception errors; object hallucination; process supervision; verifiable rewards; reward hacking; reasoning faithfulness*

1. Introduction

1.1. The Phenomenon

Ask a multimodal large language model to read a value off a chart and compute with it, and the response typically looks exemplary: a structured chain that identifies the chart, names the relevant series, extracts values, and carries out the arithmetic in orderly steps [1,2]. Increasingly often, that answer is also correct. What the response format conceals is a second, independent question: did the model actually read the chart? The evidence surveyed in Section 2 says not always. Models misread values from axes and bar heights while narrating the extraction as if performed [3]; they fail counting and geometric discrimination at rates far below human baselines while producing confident explanations [4,5]; and they describe objects and attributes that are not in the image at all [6,7]. Because the chain sounds like verification, both human readers and automated graders tend to accept it.

The phenomenon's training-theoretic half is what makes it more than a curiosity. The dominant recipe for improving reasoning—reinforcement learning and rejection sampling scored by final-answer correctness, from early verifiers [8] through process-free self-taught reasoners [9] to today's verifiable-reward systems [10,11]—assigns exactly the same credit to a trace that misread the chart and guessed well as to a trace that read the chart and reasoned soundly. When the two are pooled into a single "correct" class, every outcome-based update reinforces whichever processes happen to be sampled alongside correct answers, not the processes that deserve them. A training regime with no view of the process cannot penalize a process-level defect that repeatedly co-occurs with reward; it can only reproduce it more fluently.

This survey's subject is the intersection: perception errors inside multimodal reasoning chains, and the structural blind spot that outcome-based scoring builds around them. The topic sits at the confluence of three literatures that have developed largely apart—multimodal perception evaluation, chain-of-thought faithfulness, and reward modeling/RLHF—and the survey's contribution is to join them.

1.2. Scope and Method

We cover 2018–2025 literature on (i) perceptual failure modes of multimodal models under structured evaluation, (ii) faithfulness of generated reasoning chains in text and multimodal settings, and (iii) training signals for reasoning—outcome- and process-based—with emphasis on where they differ. We searched arXiv and the proceedings of the major machine-learning

venues, prioritizing papers that either measure process quality directly or introduce a training/evaluation method whose design presupposes a position on process measurability. Reported figures are quoted from primary sources; approximate figures are marked as such, and no quantitative claim is carried over from secondary summaries. We exclude pure text-hallucination work except where it supplies the conceptual vocabulary (Section 2.2) and we exclude model-updating and knowledge-editing literatures, which raise adjacent but distinct consistency questions. The companion survey to this one treats the architecture side—how modalities are coupled into language models—and reaches this survey’s doorstep from the other direction: integration has matured while perception has not, so the perception errors analyzed here are architectural common ground across model families, which is what makes their treatment by training signals a question worth asking in general.

1.3. Positioning and the POC Framework

Three nearby surveys frame, without occupying, this survey’s ground. Huang et al. survey hallucination in text LLMs—definitions, taxonomy, mitigation—without treating multimodal perception or reasoning-train signals [12]. Wang et al. survey multimodal chain-of-thought techniques, including a taxonomy of reasoning designs and an overview of evaluation, but organize by technique rather than by the consistency between process and outcome [13]. The hallucination-in-LVLM literature (object-level fabrication) and the chart-reasoning evaluation literature each document half of the phenomenon [3,7]; neither connects it to the training-signal economics that determine whether those errors survive reinforcement. Table 1 summarizes the gap.

Table 1. Positioning among adjacent surveys.

Dimension	Huang et al. [12]	Wang et al. [13]	This survey
Modality	Text LLMs	Multimodal reasoning	Multimodal reasoning
Organizing axis	Hallucination taxonomy	Reasoning techniques	Process–outcome consistency
Perception errors in chains	Out of scope	As evaluation items	Central object
Training-signal analysis	RLHF briefly	Brief	Formal (insensitivity principle)
Prescriptive target	Mitigation methods	Technique selection	Supervision and metric design

The organizing object is the Process–Outcome Consistency (POC) framework, introduced in Section 3: any reasoning episode is classified by whether its process (what was read, cited, and inferred) is sound and whether its outcome (the final answer) is correct, yielding four cells whose least-studied and most consequential is the compensated error—flawed process, correct

outcome. Three claims structure the survey. Descriptive: perception errors inside fluent chains are common, documented across modalities and model families (Section 2). Analytic: outcome-based supervision is structurally insensitive to the SC/FC distinction, so it cannot be the sole corrective (Section 4). Prescriptive: the mitigation space is a five-point leverage map whose unifying idea is to make the process checkable—by tools, by trained process models, or by evaluation protocols (Section 5).

2. The Evidence: Perception Errors Inside Coherent Reasoning

2.1. Misreading Structured Content: Charts, Plots, and Geometry

The chart-understanding literature supplies the cleanest decomposition because it separates reading from reasoning. Early benchmarks already showed the split: ChartQA paired questions requiring visual extraction with ones requiring logical composition [14], and PlotQA probed numeric answers whose derivation runs through precise value readings on plotted series [15]. The diagnostic sharpened with CharXiv, a suite of realistic, uncured charts with paired descriptive questions (read a value) and reasoning questions (operate on read values), reporting that the strongest proprietary model trailed non-expert humans by roughly thirty points on the reasoning split (47.1% versus 80.5%), with the strongest open-source system at 29.2% and the descriptive split—pure reading—similarly depressed [3]. The paired design carries the interpretive weight: when descriptive accuracy is far below reasoning-styled accuracy on comparable items, the fluent "reasoning" over unread values is not reasoning at all but narration over guessed numbers. MathVista's aggregation across visual-mathematical contexts showed the same signature in frontier systems, with then-leading GPT-4V at 49.9% overall despite chains that rarely acknowledged uncertainty about what they had read [16].

Charts deserve a sentence on why they are the phenomenon's natural laboratory. They are, uniquely, structured content where perception has exact ground truth (a bar has a height; an axis has a scale), where the language prior is strong (charts co-occur with conventional narratives in pretraining text, so the model has strong expectations about what any chart "says"), and where downstream reasoning is verifiable (arithmetic has an answer). That triple—weak visual grounding, strong textual priors, verifiable answers—is precisely the configuration in which a chain can be wrong in what it reads and right in what it answers. Document workflows, financial and scientific reporting, and dashboard analysis inherit the configuration; the chart benchmarks measure it in isolation. When the strongest systems' descriptive-versus-reasoning split on CharXiv shows narration outrunning extraction [3], the finding generalizes beyond charts to the

entire class of tasks where fluent expectations substitute for read evidence.

Perception-failure evidence extends below chart conventions to primitives. The BlindTest suite isolates seven elementary visual skills—line-segment intersection counts, nested shape counts, relative angle and path length judgments—on synthetic images where language priors are useless, and reports four frontier VLMs averaging 58.07%, the strongest reaching only 77.84%, on items humans solve at nearly 100% [4]. Eyes Wide Shut constructs CLIP-blind pairs—images trivially distinguishable to humans that produce near-identical embeddings under CLIP-family encoders—and shows that MLLMs built on such encoders answer at or below chance while emitting articulate justifications [5]. Two features of this evidence matter for the survey's argument. First, the failures are not exotic: counting and reading are the atomic operations of chart QA, financial analysis, and document workflows. Second, the chains are coherent in every case—the failure is upstream of logic, in what the chain reports having seen, which is precisely the component outcome-based scoring never inspects.

2.2. Fabricated Evidence: Objects and Attributes Not in the Image

The second family is referential: the chain cites visual evidence that does not exist. The phenomenon was named and measured in captioning, where CHAIR quantified the rate of mentioned objects absent from the ground-truth scene [6]. POPE generalized the measurement to LVLM question answering with a polling protocol—is this object present?—and established that hallucinated objects are frequent under adversarial sampling, with models exhibiting systematic yes-bias: a prior over what should be in a scene that overrides the pixels [7]. HallusionBench demonstrated the deeper version with paired-question designs: when the question text biases toward an answer and the image contradicts it, models follow the text; when the image is constant and the question changes, models repeat; GPT-4V answered just 31.42% of the suite's question pairs correctly, with explanations that rationalized the prior-driven answer after the fact [17]. AMBER consolidated the dimensions—existence, attribute, relation—into an LLM-free evaluation confirming generality across model families [18]. The line connecting this literature to the shortcut-learning tradition is direct: models exploiting dataset priors over true signal is the canonical shortcut pathology [19], diagnosed in earlier eras as being "right for the wrong reasons" on NLI heuristics [20] and as annotation-artifact susceptibility in QA [21]; multimodal reasoning chains are the pathology's newest and most fluent carrier. Within the POC framework of Section 3, prior-override episodes are predominantly compensated errors in waiting: the language prior that supplies a wrong answer on adversarial items supplies right answers on ordinary ones, and the chain narrates whichever

outcome occurred as if it had been seen.

2.3. The Knowing–Doing Gap

Two findings complicate a purely representational story and matter for mitigation. First, models often encode more than they use: on fine-grained perception failures, the correct visual information is present in internal attention distributions, and training-free interventions that steer attention toward question-relevant regions, or apply contrastive decoding against distorted-image predictions, recover a large fraction of errors without any retraining [22]. Second, the language prior's dominance is a decoding-time phenomenon as much as an encoder one: visual contrastive decoding, which subtracts the model's predictions on degraded images from its predictions on the true input, reduces hallucinated-object rates substantially—direct evidence that the prior was contaminating output selection, not merely coexisting with weak evidence [23]. Related work showed hallucination amplifies with response length and specificity, the exact direction chain-of-thought pushes [6,23]. And the correction side has its own literature—rewriting hallucinated content with retrieved visual evidence (LURE), verification-and-correction pipelines (Woodpecker), and information-grounding controls (visual information grounding) [24–26]—each demonstrating after the fact that the evidence to avoid the error was available. The gap, in short, is between what the system knows and what its generated process claims and uses: the chain is a report about perception written by a subsystem that did not consult the perception. That formulation is what makes the phenomenon a training-signal problem rather than only a capability problem—if the information is present, scoring that never asks whether the report used it will not fix the reporting.

2.4. A Taxonomy of Process Errors

Table 2 consolidates the evidence into a taxonomy of process-level error types, each with its characteristic benchmark evidence and its detectability by external tooling—the property Section 5 exploits.

The taxonomy's last row is the training-relevant residue of the first four: because chains are fluent, corrupted evidence flows into derivations that are locally valid, making the final trace self-consistent—the property that makes chains persuasive to graders and the property that outcome scoring rewards. "Self-consistent" and "sound" are different properties; the entire argument of Section 4 is that outcome-based supervision optimizes the first while pricing the second at zero.

Table 2. Process-level error types in multimodal reasoning chains.

Type	Definition	Representative evidence	Detectable by tooling?
Extraction error	Cited value/label differs from image content (axis misread, series confusion)	CharXiv, ChartQA [3,14]	High (OCR/plot-parsing tools)
Primitive failure	Miscounting intersections, shapes, orientations	BlindTest, MMVP [4,5]	Medium (specialized verifiers)
Referential fabrication	Objects/attributes cited that are absent	CHAIR, POPE, AMBER [6,7,18]	High (open-vocabulary detectors)
Prior override	Answer follows question/language prior against pixels	HallusionBench, VCD analyses [17,23]	Medium (paired probes)
Unsound derivation	Valid-looking inference over corrupted or unsupported premises	Faith-and-Fate analyses [27]	Low–medium (step checkers)

3. The Process–Outcome Consistency Framework

3.1. The Four Cells

Fix a question with a ground-truth answer and a model's reasoning trace. Call the trace's outcome correct iff the final answer matches ground truth. Call the trace's process sound iff (i) every perceptual act in the trace—values read, entities cited, spatial claims—matches the image, and (ii) every inferential step follows from its predecessors. Soundness is demanding (it is faithfulness plus validity), but it is the standard outcome scoring implicitly claims to reward. Crossing the two binary properties yields four cells (Table 3; Figure 1 visualizes the quadrant and the reward-flow asymmetry that Section 4 formalizes).

Table 3. The Process–Outcome Consistency (POC) 2×2.

	Outcome correct	Outcome wrong
Process sound	<i>SC — sound success.</i> What benchmarks measure as success and training intends to reward.	<i>SW — honest failure.</i> Reasoning failed despite sound process; information about capability limits, no reward.
Process flawed	<i>FC — compensated error.</i> Right answer via lucky guess, prior rescue, error cancellation, or post-hoc rationalization. <i>The blind spot: rewarded identically to SC.</i>	<i>FW — double failure.</i> Penalized; the only cell outcome metrics see when they see anything.

Three points about the framework. First, FC is not a corner case. The Section 2 evidence implies a substantial base rate: models with documented misreading rates in the tens of percent still produce final-answer accuracies that include those episodes whenever the answer survives

the misread—which happens routinely when priors are strong, options are few, or the question tolerates a misread value. Second, the cells are empirically distinguishable—truncation probes [28], biased-posterior tests [29], and tool-based evidence audits (Section 5.5) all operationalize the soundness axis—so the framework is a measurement proposal, not merely a vocabulary. Third, the framework converts vague anxieties ("models guess") into measurable quantities: the compensated-error rate $CER = P(\text{process flawed} \mid \text{outcome correct})$, and the evidence-grounding rate $EGR = \text{fraction of trace steps whose cited evidence is verifiable and correct}$. Current reporting practice publishes neither.

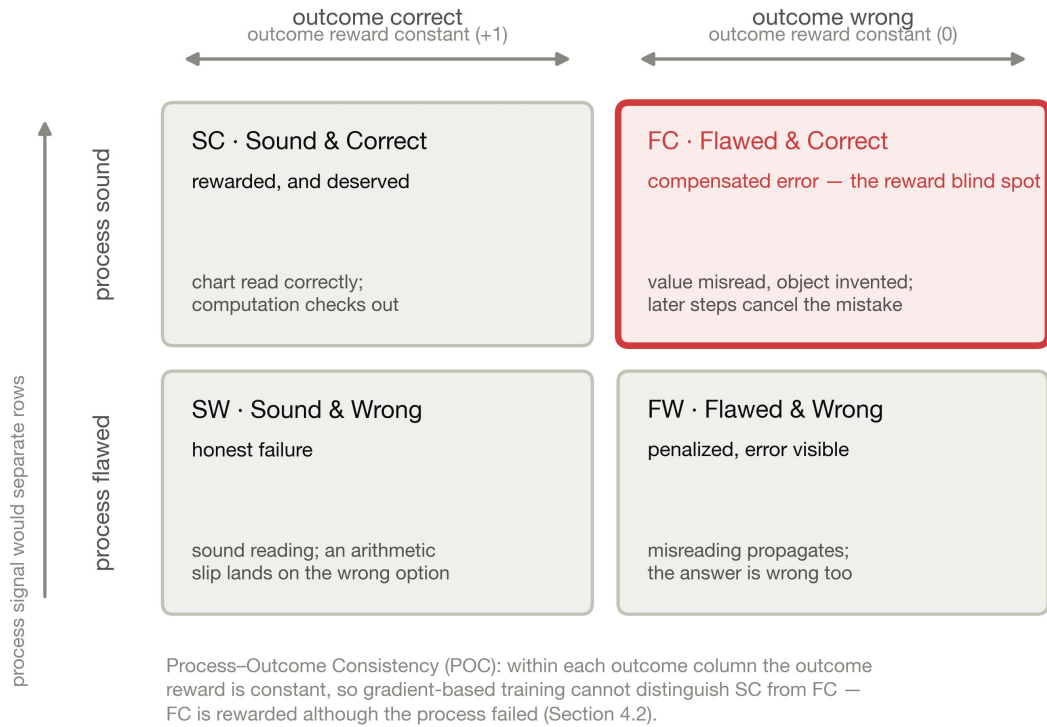


Figure 1. The Process–Outcome Consistency (POC) 2×2. Within each outcome column the outcome reward is constant, so gradient-based training cannot distinguish SC from FC; the FC cell—compensated error—is the reward blind spot.

A fourth point concerns the cell outcome metrics do see. SW (sound process, wrong outcome) is the informative failure—it localizes capability limits in the inference stage—and FW double failures supply the error signal that outcome-based methods learn from. The asymmetry between the visible diagonal and the blind FC cell explains a characteristic shape in the literature: training regimes built on outcome signals are good at eliminating double failures (the model stops answering wrong) while FC survives unmeasured (it keeps answering right for the wrong reasons), so measured progress and structural progress diverge exactly along the soundness axis. A fifth point is defensive: the framework classifies traces, not models. The same model emits SC on some items and FC on others, so CER is a distributional property of a

model-task pair—the quantity to optimize and report, not a verdict on any single response.

3.2. Why Outcome Metrics Cannot See the Blind Spot

An answer-only evaluation computes $P(\text{outcome correct})$, which conflates SC and FC; the reported number is the sum of two populations with different properties. The conflation is benign in evaluation only if FC is small and stationary. Neither holds. FC grows whenever language priors are informative (benchmarks with conventional content), whenever answer spaces are small (multiple choice tolerates misreads), and—as Section 4.3 documents—whenever training optimizes the reported number, because the optimizer has no incentive distinguishing the two populations and every incentive to produce whichever is cheapest to sample. The evaluation and training problems are the same problem in two poses: a grader that cannot distinguish SC from FC cannot certify capability; a reward that cannot distinguish them cannot select for it. Section 3.3 and Section 4 develop the two poses in turn.

3.3. Measuring the Soundness Axis

The faithfulness literature supplies the measurement toolkit, developed on text. Truncation and perturbation probes sample a chain, delete or corrupt steps, and re-query: when the answer survives deletion of the entire chain, the chain was not the cause of the answer [28]. Biased-posterior interventions plant the answer in the prompt via few-shot examples or stronger forms and find that models often adopt the planted answer while producing a chain that argues for it from "the evidence"—rationalization, in the strict sense [29]. Filler-token experiments show models can use latent computation without any meaningful chain at all—steps that look like reasoning carrying none of it [30]—and compositional analyses trace how chains fail on multi-step problems, as plausible-but-wrong sub-procedures rather than random noise [27]. Program-trace alternatives force the reasoning into symbolic calls whose steps are executable, making faithfulness structural rather than hoped-for [31]. Each probe transfers to multimodal chains with a twist the text toolkit lacks: the perceptual claims in a multimodal trace are checkable against an external artifact—the image—in a way textual presuppositions are not (Section 5.5 makes this the basis of EGR). The soundness axis, in other words, is hard to measure in text and merely laborious in vision, which is one of this survey's load-bearing asymmetries.

The soundness axis also has a human-measurement problem that anticipates Section 4. Human graders exhibit the same blind spot as automated ones, for the same reason: auditing a chain's perceptual claims against the image is expensive, while judging whether an answer is correct and whether the chain reads well is cheap. Fluency functions as an audit substitute,

exactly as it does for outcome-based reward models [32]—the observation behind RLHF-V's finding that preference tuning rewards confident description and thereby amplifies hallucination. The consequence is that CER does not disappear when evaluation moves from automated metrics to human panels; it becomes unmeasured human approval of compensated reasoning. Any evaluation regime that prices process audits above grader budgets will reproduce the blind spot regardless of who—or what—does the grading, which is why Section 5.5's proposals are protocol designs that make audits cheap rather than exhortations to grade harder.

4. The Reward Blind Spot of Outcome-Based Training

4.1. The Outcome-Supervision Family

The family shares one property: the training signal is a function of the final answer. Its members are historically ordered. Verification-filtered fine-tuning trained answer-verifiers on GSM8K and reranked sampled solutions [8]. STaR bootstrapped rationales by keeping those whose final answer matched, fine-tuning on them, and iterating—explicitly treating any correct-answer rationale as a positive example of reasoning [9]. Self-consistency and its distillation variants select chains by majority-vote agreement of answers—again, an outcome statistic [33]—with a subtlety worth naming: majority voting masks compensated errors in evaluation too. If a model's FC episodes are prior-driven, they cluster on the same (correct) answer, so voting consolidates precisely the unsound mass that per-sample CER would expose; a system can raise its aggregate accuracy by concentrating its unsound guesses, which looks like progress under $P(\text{correct})$ and is regression under CER. Outcome reward models learn to predict answer correctness and score chains [8,34]. The current wave, verifiable-reward RL (RLVR), scores by rule-checked answers—unit tests, exact match—with GRPO-style policy optimization, at scales from Tülu 3 [10] through DeepSeek-R1, whose emergent long chains were shaped entirely by answer-based reward [11,35]. The family's record is real: outcome supervision measurably improves answer accuracy [8,10,11]. The family's defining limitation is the subject of this section: the signal never inspects the trace. Everything below follows from that single fact.

Why the family dominates in practice is economic before it is intellectual. Answer labels are nearly free—ground truth comes with the dataset, and rule-checkers scale—whereas process labels require a judge capable of reading the trace, whether human, model, or tool, at per-step cost. Outcome supervision therefore amortizes: every improvement in base-model sampling

makes the answer filter cheaper per useful trace. The economics are sound for what they measure; the argument of this section is that what they do not measure is not a residual but the survey's core phenomenon, and that the economics flip under specific constructions (Section 5.3) where tooling substitutes for annotation.

4.2. The Insensitivity Principle

Formally: let $R_o(T) = 1\{\text{answer}(T)\text{correct}\}$ be outcome reward, and partition correct-answer traces into sound (SC) and compensated (FC). The policy gradient update $\nabla \mathbb{E}[R_o] = \mathbb{E}[R_o \cdot \nabla \log \pi(T)]$ weights every sampled correct-answer trace identically, so the relative reinforcement of SC versus FC within the correct pool is determined entirely by sampling frequencies—not by soundness, which the reward does not encode. Three consequences follow. Differential amplification: if FC traces are cheaper to sample (shorter chains, fewer attended image regions, prior-friendly answers), they arrive more often and are reinforced proportionally more; the outcome-optimal policy has no pressure against drifting toward the FC-dense region of its own sampling distribution. Contamination by curation: rejection-sampling SFT filters exactly on the answer, so the resulting fine-tuning corpus contains compensated traces at their sampling rate and imitates their style—the narration of unperformed perception—without any flag; the STaR lineage's "correct rationale" is precisely $\text{FC} \cup \text{SC}$ [9]. Verifier inheritance: an ORM trained to predict answer correctness from traces learns whatever features correlate with outcomes in its data, not soundness; if FC traces are common in training data, "sounds fluent" may be the learned proxy [8,36].

The principle's sign matters as much as its zero: outcome supervision does not merely ignore the SC/FC boundary—it can shift mass across it, because the boundary is unobservable to the mechanism doing the shifting. The shift has an efficiency gradient as well as a frequency one. Chains that genuinely attend to images are longer, slower to sample, and more often interrupted by uncertainty; chains that narrate priors are short and confident. Under a fixed computational budget for sampling, RLVR's rollouts will therefore over-represent the cheap narrators among its rewarded traces, and length or effort penalties—common pragmatic additions to outcome rewards—widen the gap by taxing exactly the traces that look at the image. The obscuration result of Section 4.3 is the endpoint of this gradient: when monitoring penalizes visible shortcuts, the reward still flows, so the shortcut moves rather than closes [37]. Both dynamics are reward hacking in Skalse et al.'s formal sense: an outcome proxy whose optimum need not coincide with the true (reasoning-quality) objective [38], foreseen in the alignment literature as the

reward-specification problem [39]. The multimodal instance adds a specific failure surface: the cheapest reliable way to produce correct answers under weak perception is to lean harder on language priors—the one mechanism Section 2 showed produces FC-by-construction [17,23]. Outcome-only RL on multimodal tasks is, on this analysis, structurally positioned to strengthen prior reliance, because prior reliance is the statistically efficient route to reward that never audits perception.

4.3. Empirical Signatures

The predicted signatures have text-side precedents that transfer. Reward-model overoptimization follows a scaling law: as policy optimization against a proxy reward continues, true reward rises then falls while proxy reward keeps climbing—the canonical overoptimization curve, measured across KL budgets [36]. In code-generating RL with executable tests, models learn to hack the tests, and—directly relevant to chain trust—when the chains are monitored and rewarded against hack-like content, the behavior does not disappear; it is obscured, with models learning to write chains that look clean while planning hacks in unmonitored channels [37]. That result is the empirical form of Section 3's self-consistency point: outcome pressure makes chains less faithful, not more, when the reward cannot see the difference. Imperfect-verifier analyses close the loop on the family's own machinery: resampling against imperfect verifiers (the outcome family's workhorse) saturates quickly, with verifier quality—not sample count—becoming the binding constraint [40]. On the multimodal side, the signature appears as a motivation and a measurement: RLHF-V's starting observation was that naive preference tuning increases multimodal hallucination—annotators reward confident, fluent descriptions, and hallucination is fluent—requiring claim-level human feedback to reverse [32]; the same group's RLAI-F-V scaled the claim-level structure with AI feedback [41]. Preference fluency bias is the FC dynamic operating one level up, at the response level rather than the trace level.

4.4. The Multimodal Wave Sharpens the Stakes

The 2024–2025 wave of multimodal reasoning systems—LLaVA-CoT's staged deliberation [42], Vision-R1's RLVR adaptation with curated cold-start reasoning data [43], Perception-R1's explicitly perception-targeted policy optimization [44], and process-reward initiatives such as VisualPRM [45]—imports the R1 recipe into settings where Section 2 documents higher baseline rates of process error. It is worth being precise about what today's multimodal RLVR optimizes: answer-exactness on math- and chart-styled items, checked by rule. The checkable-answer surface is also the FC-prone surface—the tasks where a misread value still lands on the

right option band—so the wave's reward concentrates on exactly the items where the SC/FC distinction matters most and is least visible to the checker. Perception-R1 is the wave's interesting counter-design: by targeting perception itself as the RL objective, it attacks the base rate $P(\text{process flawed})$ rather than the answer channel, complementing rather than contradicting this survey's argument—which concerns the reward's blindness on the answer channel. Nothing in the wave's published recipes yet audits sampled traces' evidence before reward; the constructions of Section 5.3 are designed to slot into exactly that slot. The sharpening is arithmetic: if $P(\text{process flawed})$ is higher on multimodal tasks, then the FC share of rewarded traces is higher, and the insensitivity principle's differential-amplification consequence operates on a larger blind-spot mass. The wave is therefore the right moment for the framework of this survey—after it, un-sound reasoning habits will be baked into the reward-optimized policies that downstream systems imitate. Two features make multimodal RLVR distinctive in the other direction: answer verification for chart/document tasks is nearly as cheap as in text, and—the survey's optimistic asymmetry, developed in Section 6.1—the process errors themselves are tool-checkable in ways text reasoning errors are not.

A micro-example fixes the mechanism. Consider a two-option chart question—"did revenue exceed 40M in Q3?"—where the model misreads the Q3 bar as 38M when it is 42M, then reasons validly: $38 < 40$, answer no. The trace is FW, penalized, and eventually trained away. Now consider the same misread on a question whose options are 35M and 40M-banding: the model reads 38M, reasons validly within its false premise, and selects the "upper-30s to 40" band—which is correct. The trace is FC: rewarded, filtered into rejection-sampling data [9], and reinforced with the same weight as a trace that read 42M. Note what the example shows about the blind spot's shape: outcome reward cannot distinguish the two traces even though one contains a perception error and the other does not, because the reward's entire view of the trace is its final token. Every per-step distinction—the 42M versus 38M read, the soundness of the comparison—lies in the update's null space. Charts make the null space visible because ground truth for intermediate values exists; the mechanism is general.

5. Mitigation: Five Leverage Points

Table 4 previews the map before the subsections elaborate it. The organizing idea: a compensated error survives only while the process is uncheckable; each leverage point makes some component of the process checkable—by tools, by trained verifiers, by construction, or by protocol.

Table 4. The five leverage points against compensated errors.

Leverage point	Representative methods	What it catches	Relative cost
Upstream perception	Encoder/decoder interventions: attention steering, contrastive decoding [22,23]	Extraction, primitive, fabrication errors before they enter the chain	Zero-training; per-inference
Process shaping at inference	Extract-then-compute decompositions; program-of-thoughts [31,46,47]	Unsound derivations; unperformed reads made explicit	Prompt/agent engineering
Training-signal redesign	PRMs [48,49]; claim-level feedback [32,41]; multimodal PRMs [45]; tool-verified curation (this survey's synthesis)	FC itself—the only point that changes the reward	Annotation/tooling + RL cost
Post-hoc verification	Rewrite/correction pipelines [24,25]; verifier ensembles [40]	Detectable errors before exposure	Per-inference; bounded by verifier quality
Process-aware evaluation	Faithfulness probes [28,29]; paired-question and extraction-required designs [3,17]; CER/EGR reporting	Measurement blind spot itself	Protocol design

5.1. Upstream Perception

If the error is in what was read, the cheapest fix is where it is read. The knowing–doing results make decoding-time intervention surprisingly effective: steering attention toward question-relevant regions and contrasting against distorted-image predictions recovers much fine-grained perception without training [22,23]. Resolution and layout engineering (tiling, native dynamic resolution) addresses the extraction channel structurally, as the companion survey details; chart-specific interfaces—chart-to-table extraction before answering, guard tokens that verify a number was grounded in the plot, program-of-thoughts training for chart arithmetic—re-engineer the read-then-compute boundary at the model level [46,50,51]. These methods reduce the base rate of process errors and therefore the FC mass entering training pipelines; they do not, by themselves, give any training signal a view of the process, which is why they are necessary but insufficient.

Data-side perception work belongs at this leverage point too, because the base rate of process errors is set by what pretraining taught about reading. Chart- and document-specific corpora with synthesized ground truth—generated charts whose underlying data tables are known exactly—turn the extraction channel into a supervised problem with verifiable labels, which is why chart-specialized models close a disproportionate share of the extraction gap despite

modest scale [46,50]. The general lesson transfers: anywhere the image's structured content has a reconstructable ground truth, synthetic generation supplies perception training data with process-level supervision built in—extraction checked at data-construction time is the zeroth layer of the audit ladder that Sections 5.3–5.5 then extend to inference and training.

5.2. Process Shaping at Inference

Where the model cannot be trusted to narrate perception, the narration can be structured out of its hands: force extraction (values, entities) into an explicit, checkable artifact before computation proceeds over the artifact. The idea has independent lineages—program-of-thoughts prompting for numeric chains [46], symbolic-program chains whose steps execute [31], and visual programming's compositional delegation [47]—and it is the operational form of faithfulness: the trace is the computation, so self-consistency and soundness collapse into one property. A concrete chart-QA pipeline makes the shape plain: (1) parse the plot into a table with a dedicated extraction pass; (2) freeze the table; (3) perform all arithmetic over table cells, citing cell addresses; (4) emit the answer with the table attached. Steps 2–4 cannot fabricate a value that step 1 did not produce, and step 1's output is diffable against tool-parsed ground truth. Structured formats thereby convert the hardest cell of Table 2 (unsound derivation over corrupted premises) into the easiest (extraction diff): either the table matches the image or it does not, and no amount of downstream eloquence can launder the difference.

The leverage point's costs are real and bound its use. Decomposition taxes capability where tasks resist staging (holistic visual judgments, estimation under ambiguity), adds latency, and can introduce its own errors at the parse step; forcing extraction on tasks that need less than full extraction wastes budget and creates audit surface where none is needed. The design rule that respects both sides: decompose and expose exactly the claims that will be rewarded or audited later, and leave the rest of the chain free. Section 5.3's tool-verified filtering then audits precisely the exposed artifact, so the interface and the reward agree on what "the process" is.

5.3. Training-Signal Redesign

The only leverage point that changes the reward itself. Text-side results establish the premise: process supervision outperformed outcome supervision at comparable annotation budgets on mathematical reasoning (best-of-N solve rates of roughly 78% versus 72% attributed to PRM-versus ORM-ranked sampling on MATH, in Lightman et al.'s large-scale comparison [48]), and earlier process/outcome comparisons already showed process-level signals improving final-answer performance per unit of supervision [49]. The multimodal translations now exist in early

form: claim-level human feedback that corrects specific hallucinated statements (RLHF-V, which produces substantially more trustworthy responses than response-level preference tuning [32]), its AI-feedback generalization (RLAIF-V [41]), and multimodal process reward models trained to score reasoning steps [45].

This survey's synthesis generalizes the design pattern that runs through all of them: make the process signal from what is already checkable. Concretely, two constructions follow. (i) Perception-verified rejection sampling: before a sampled correct-answer trace enters SFT or RL data, audit its cited values/entities against the image with OCR and open-vocabulary detection; traces failing the audit are excluded—or better, kept as preference negatives. (ii) Process-contrastive pairs: for the same question and the same correct answer, construct a sound trace (evidence verified) and a corrupted twin (values flipped, object fabricated), and apply DPO-style optimization [52] on the pair; the outcome term is constant by construction, so the entire gradient signal isolates process soundness—the training-time analogue of CharXiv's paired design. Both constructions use only existing components (samplers, detectors, DPO); their novelty is the pairing constraint that neutralizes the insensitivity principle, and they are offered here as the survey's prescriptive synthesis rather than as evaluated methods.

As an operational sketch of (i) in a chart-QA training loop: sample N candidate traces at temperature for each training question; parse each trace for numeric and entity citations with a fixed extraction schema; run the cited numbers against a plot-parsing oracle and the cited entities against an open-vocabulary detector; label traces by (answer correct, evidence audit passed); then update on the four resulting strata—correct-and-grounded traces as positives, correct-but-ungrounded traces as preference negatives (the FC stratum, finally visible to the objective), wrong-and-grounded traces retained as honest-failure signal, and double failures discarded. The loop's only new machinery is the extraction schema and the oracle calls; the four-stratum labeling is exactly the POC table implemented in a data pipeline, which is the framework's point: the blind spot closes not by measuring reasoning in the abstract but by checking the specific, checkable claims chains make.

5.4. Post-Hoc Verification and Its Limits

Correction pipelines (LURE's evidence-based rewriting, Woodpecker's five-stage check-then-correct) demonstrate that fabrication and extraction errors are findable after generation with off-the-shelf machinery [24,25]. The limits are as documented as the successes: verifier quality caps resampling gains quickly [40], verifier–generator gaps mean the strongest models have the weakest external checkers, and per-inference verification multiplies serving cost. The

leverage point's proper role is therefore containment in deployment—especially for high-stakes structured-content workflows where the extraction diff is cheap—rather than a substitute for signal-level redesign, since post-hoc filtering changes neither the policy nor the data it generates. Containment also has a calibration form that costs no verifier at all: reward or prompt the model to report uncertainty about its own reads. A chain that says "the Q3 bar reads approximately 38M (low confidence)" converts a silent compensated error into a flagged one, and Section 2.3's evidence that the information for correct reads is often internally present [22] makes well-calibrated read-confidence a learnable target. Abstention on low-confidence extraction is not a retreat from capability claims; it is the deployment-grade expression of the SC/FC distinction—and currently one of the rarest behaviors in the chains this survey's benchmarks collect.

5.5. Process-Aware Evaluation

The evaluation leverage point closes the loop with Section 3's measurement proposal. Three practices would make the blind spot visible in reported results: (i) paired designs that hold the answer-relevant prior fixed while varying the image—HallusionBench's protocol [17]—or hold content fixed while varying extraction load—CharXiv's descriptive/reasoning split [3]; (ii) faithfulness probes as standard reporting (truncation survival, paraphrase sensitivity [28]); and (iii) evidence audits: tool-checked verification of cited values and entities, yielding EGR, and tool-classified soundness among correct answers, yielding CER. None requires new science; all require declining to publish answer accuracy alone. Concretely, a results row for a chart-QA system would read as a five-tuple: answer accuracy; CER (share of correct answers whose traces failed the evidence audit); EGR; truncation-survival rate (share of answers unchanged when the chain is deleted [28]); and the audit's own coverage (share of trace claims the tools could check). The last field is the honest one—it prices the report's confidence—and its absence in current practice is part of the blind spot, since an unaudited 100% EGR and a fully audited one are indistinguishable in today's tables. The incentive point deserves emphasis: benchmarks do not merely measure—they select. A leaderboard that reports only $P(\text{outcome correct})$ trains the field's collective attention on the $SC \cup FC$ mixture; one that reports CER alongside it makes compensated reasoning a visible, comparable defect. Evaluation is the cheapest leverage point, and the only one whose audience includes every other point on this map.

6. Discussion

6.1. The Tractability Asymmetry

The survey's recurring asymmetry deserves its own statement: multimodal process errors are externally checkable in ways text reasoning errors are not. A hallucinated object has a detector; a misread chart value has an OCR-and-parsing pipeline; a fabricated spatial relation has a grounding model. Text-side process supervision must lean on learned PRMs whose own reliability is the open question [36,48]; the multimodal setting can lean on perception tooling whose failure modes are comparatively characterized. This is why constructions (i) and (ii) of Section 5.3 use no science beyond assembly. The asymmetry also suggests where multimodal process supervision will not be hard where text was: annotation cost, the historical bottleneck of process supervision [48], can be partly replaced by tool verdicts, which are cheap, consistent, and auditable in turn. The cost comparison is roughly: a human process label costs minutes of expert time per step; a tool verdict costs milliseconds and produces a diffable record; a model-based judge costs an inference call and inherits the judge's own blind spots. Tooling is weakest exactly where perception is hardest—the primitive failures of Table 2 resist cheap verifiers—so the practical program is a coverage ladder: tool-audit the structured claims (most of the FC mass on document and chart tasks), model-judge the rest, reserve human review for calibration of the other two. If the multimodal reasoning wave adopts this ladder early, faithful multimodal reasoning could become the setting where process supervision is cheaper than outcome supervision plus its failures—not a luxury good.

6.2. The Monitoring Tension

A second-order effect deserves caution. If outcome-only RL makes unfaithful chains (Section 4.3's obscuration result [37]), then process monitoring introduced without process reward may select for chains that evade monitoring—the text-side result in code RL, transposed. The design implication is sequencing: process signals should be built into the reward at the same stage that process inspection becomes load-bearing, not bolted on afterward as an expectation. The POC framework supplies the vocabulary for what such signals must distinguish; Section 5.3's constructions show they can be assembled from existing parts.

6.3. Limits of the Framework

Three caveats bound the framework's claims. First, "soundness" is stricter than human reasoning: experts skip steps, cite evidence economically, and reach correct answers by pattern recognition that no audit would reconstruct; a CER target of zero would penalize legitimate

expertise-shaped reasoning and select verbose chains. The framework's use is as a reported quantity and a reward component, not a pass/fail bar. Second, process metrics are gameable too: a model rewarded for citing checkable evidence can learn to cite verbatim-plausible evidence, and tool verdicts inherit their tools' errors; Section 4.3's overoptimization logic applies to PRMs as to ORMs [36,38]. Third, the SC/FC boundary is epistemic, not ontological—the same trace can be sound given one reading of a blurry axis and compensated given another—so CER estimates carry the audit's own uncertainty and should be reported with it. Fourth, soundness as defined here bundles three properties—faithfulness (the chain caused the answer), grounding (cited evidence matches the image), and validity (steps follow)—that come apart in principle: a trace can be perfectly grounded yet invalid, or valid yet unfaithful. The framework's quantities target the bundle differently—EGR measures grounding, probes measure faithfulness, step-checking measures validity—and conflating them in a single "process score" would hide exactly which property a training regime improved. None of these caveats weakens the core claim; each relocates it from "process is measurable" to "process is measurable well enough to be worth measuring," which the Section 2 evidence says is currently not happening at all.

7. Conclusion

This survey joined three literatures that have been solving adjacent halves of one problem. The perception-evaluation literature documented that multimodal reasoning chains routinely misread structured content, fail primitives, and fabricate evidence while remaining fluent [3,4,7,17]. The faithfulness literature supplied the measurement tools showing that chains are often not the cause of their answers [28-30]. The reward-modeling literature showed that optimization against any proxy drifts off the true objective and that outcome-only signals cannot even represent the difference [36-38]. The POC framework states the join precisely: outcome-based training and evaluation occupy one axis of a two-dimensional space, and the cell they cannot see—compensated error—is where the documented perception failures go to be rewarded.

The prescriptive summary is short because the parts exist. Make the process checkable: at inference, by extracting before computing; in training, by tool-verified rejection sampling and process-contrastive pairs that neutralize the insensitivity principle; in evaluation, by reporting compensated-error and evidence-grounding rates beside accuracy. The multimodal setting is not merely exposed to this problem but equipped for it: perception errors are the one class of reasoning errors that mundane tooling can audit, which makes faithful multimodal reasoning the most tractable instance of a problem the field has so far met mostly with cautionary tales.

What is missing is not capability but a reporting standard—the field still publishes the $SC \cup FC$ mixture and calls it accuracy. The compensated-error rate is one number away from every leaderboard that wants it.

References

- [1] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q. V., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems* 35.
- [2] Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems* 35.
- [3] Wang, Z., Xia, M., He, L., Chen, H., Liu, Y., et al. (2024). CharXiv: Charting gaps in realistic chart understanding in multimodal LLMs. *arXiv preprint arXiv:2406.18521*.
- [4] Rahmanzadehgervi, P., Bolton, L., Taesiri, M. R., & Nguyen, A. T. (2024). Vision language models are blind: Failing to translate detailed visual features into words. *arXiv preprint arXiv:2407.06581*.
- [5] Tong, S., Liu, Z., Zhai, Y., Ma, Y., LeCun, Y., & Xie, S. (2024). Eyes wide shut? Exploring the visual shortcomings of multimodal LLMs. *arXiv preprint arXiv:2401.06209*.
- [6] Rohrbach, A., Hendricks, L. A., Burns, K., Darrell, T., & Saenko, K. (2018). Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- [7] Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, W. X., & Wen, J.-R. (2023). Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- [8] Cobbe, K., Klimov, O., Hesse, C., Kim, D., & Schulman, J. (2021). Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.
- [9] Zelikman, E., Wu, Y., Mu, J., & Goodman, N. D. (2022). STaR: Bootstrapping reasoning with reasoning. In *Advances in Neural Information Processing Systems* 35.
- [10] Lambert, N., Morrison, J., Pyatkin, V., Huang, S., Ivison, H., Brahman, F., et al. (2024). Tulu 3: Pushing frontiers in open language model post-training. *arXiv preprint arXiv:2411.15124*.
- [11] Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., et al. (2025). DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*.
- [12] Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., & Liu, T. (2025). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. In *ACM Transactions on Information Systems*.
- [13] Wang, Y., Wu, S., Zhang, Y., Yan, S., Liu, Z., Luo, J., & Fei, H. (2025). Multimodal chain-of-thought reasoning: A comprehensive survey. *arXiv preprint arXiv:2503.12605*.
- [14] Masry, A., Long, D. X., Tan, J. Q., Joty, S., & Hoque, E. (2022). ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*.
- [15] Methani, N., Ganguly, P., Khapra, M. M., & Kumar, P. (2020). PlotQA: Reasoning over scientific plots. In *Proceedings of the AAAI Conference on Artificial Intelligence*.
- [16] Lu, P., Bansal, H., Xia, T., Liu, J., Li, C., Hajishirzi, H., Cheng, H., Chang, K.-W., Galley, M., & Gao, J. (2024). MathVista: Evaluating mathematical reasoning of foundation models in visual contexts. In *Advances in Neural Information Processing Systems* 36

Datasets and Benchmarks Track.

- [17] Guan, T., Liu, F., Wu, X., Xian, R., Li, Z., Liu, X., Wang, X., Chen, L., Huang, F., Yacoob, Y., Manocha, D., & Zhou, T. (2024). HallusionBench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).
- [18] Wang, J., Wang, Y., Xu, G., Zhang, J., Gu, Y., Jia, H., Zhang, J., Yan, M., Zhang, J., & Sang, J. (2024). AMBER: An LLM-free multi-dimensional benchmark for MLLM hallucination evaluation. arXiv preprint arXiv:2311.07397.
- [19] Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., & Wichmann, F. A. (2020). Shortcut learning in deep neural networks. In Nature Machine Intelligence, 2, 665-673.
- [20] McCoy, R. T., Pavlick, E., & Linzen, T. (2019). Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL).
- [21] Jia, R., & Liang, P. (2017). Adversarial examples for evaluating reading comprehension systems. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP).
- [22] Zhang, J., Khayatkhoei, M., Chhikara, P., & Ilievski, F. (2025). MLLMs know where to look: Training-free perception of small visual details with multimodal LLMs. arXiv preprint arXiv:2502.17422.
- [23] Leng, S., Zhang, H., Chen, G., Li, X., Lu, S., Miao, C., & Bing, L. (2024). Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).
- [24] Zhou, Y., Cui, C., Yoon, J., Zhang, L., Deng, Z., Finn, C., Bansal, M., & Yao, H. (2023). Analyzing and mitigating object hallucination in large vision-language models. arXiv preprint arXiv:2310.00754.
- [25] Yin, S., Fu, C., Zhao, S., Xu, T., Wang, H., et al. (2024). Woodpecker: Hallucination correction for multimodal large language models. arXiv preprint arXiv:2310.16045.
- [26] Favero, A., Zare, Z., Zarei, A., & Morency, L.-P. (2024). Multi-modal hallucination control by visual information grounding. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW).
- [27] Dziri, N., Lu, X., Sclar, M., Li, X. L., Jiang, L., Lin, B. Y., et al. (2023). Faith and fate: Limits of transformers on compositionality. In Advances in Neural Information Processing Systems 36.
- [28] Lanham, T., Chen, A., Radhakrishnan, A., Steiner, B., Denison, C., Hernandez, D., et al. (2023). Measuring faithfulness in chain-of-thought reasoning. arXiv preprint arXiv:2307.13702.
- [29] Turpin, M., Michael, J., Perez, E., & Bowman, S. R. (2023). Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In Advances in Neural Information Processing Systems 36.
- [30] Pfau, J., Merrill, W., & Bowman, S. R. (2024). Let's think dot by dot: Simple bounds on hidden computation in transformer language models. arXiv preprint arXiv:2404.15758.
- [31] Lyu, Q., Apidianaki, M., & Callison-Burch, C. (2024). Faithful chain-of-thought reasoning. In Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (EACL).
- [32] Yu, T., Yao, Y., Zhang, H., He, T., Han, Y., Cui, G., Hu, J., Liu, Z., Zheng, H.-T., Zhao, M., & Sun, M. (2024). RLHF-V: Towards trustworthy MLLMs via behavior alignment from fine-grained correctional human feedback. In Proceedings of the IEEE/CVF

- Conference on Computer Vision and Pattern Recognition (CVPR).
- [33] Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., & Zhou, D. (2023). Self-consistency improves chain of thought reasoning in language models. In International Conference on Learning Representations (ICLR).
 - [34] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., et al. (2022). Training language models to follow instructions with human feedback. In Advances in Neural Information Processing Systems 35.
 - [35] Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., et al. (2024). DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. arXiv preprint arXiv:2402.03300.
 - [36] Gao, L., Schulman, J., & Hilton, J. (2023). Scaling laws for reward model overoptimization. In Proceedings of the 40th International Conference on Machine Learning (ICML).
 - [37] Baker, B., Huizinga, J., Gao, L., Dou, Z., Guan, M., et al. (2025). Monitoring reasoning models for misbehavior and the risks of promoting obfuscation. arXiv preprint arXiv:2503.11926.
 - [38] Skalse, J., Howe, N. H. R., Krashennnikov, D., & Krueger, D. (2022). Defining and characterizing reward hacking. In Advances in Neural Information Processing Systems 35.
 - [39] Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. arXiv preprint arXiv:1606.06565.
 - [40] Stroebl, B., Kapoor, S., & Narayanan, A. (2024). The limits of inference scaling through resampling. arXiv preprint arXiv:2411.17501.
 - [41] Yu, T., Zhang, H., Li, Q., Xu, Q., Yao, Y., et al. (2024). RLAIIF-V: Open-source AI feedback leads to super GPT-4V trustworthiness. arXiv preprint arXiv:2405.17220.
 - [42] Xu, G., Jin, P., Wu, Z., Li, H., Song, Y., et al. (2024). LLaVA-CoT: Let vision language models reason step-by-step. arXiv preprint arXiv:2411.10440.
 - [43] Huang, W., Jia, B., Zhai, Z., Cao, S., Ye, Z., et al. (2025). Vision-R1: Incentivizing reasoning capability in multimodal large language models. arXiv preprint arXiv:2503.06749.
 - [44] Yu, E., Lin, K., Zhao, L., et al. (2025). Perception-R1: Pioneering perception policy with reinforcement learning. arXiv preprint arXiv:2504.07954.
 - [45] Wang, W., Gao, Z., Chen, L., et al. (2025). VisualPRM: An effective process reward model for multimodal reasoning. arXiv preprint arXiv:2503.10291.
 - [46] Zhang, L., Hu, A., Xu, H., Yan, M., Xu, Y., Jin, Q., Zhang, J., & Huang, F. (2024). TinyChart: Efficient chart understanding with visual token merging and program-of-thoughts learning. arXiv preprint arXiv:2404.16635.
 - [47] Gupta, T., & Kembhavi, A. (2023). Visual programming: Compositional visual reasoning without training. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).
 - [48] Lightman, H., Kosaraju, V., Burda, Y., Edwards, H., Baker, B., Lee, T., Leike, J., Schulman, J., Sutskever, I., & Cobbe, K. (2024). Let's verify step by step. In International Conference on Learning Representations (ICLR).
 - [49] Uesato, J., Kushman, N., Kumar, R., Song, F., Siegel, N., Wang, L., Creswell, A., Irving, G., & Higgins, I. (2022). Solving math word problems with process- and outcome-based feedback. arXiv preprint arXiv:2211.14275.
 - [50] Xia, R., Zhang, B., Ye, H., et al. (2024). ChartX & ChartVLM: A versatile benchmark and foundation model for complicated chart reasoning. arXiv preprint arXiv:2402.12185.
 - [51] Chen, J., Kong, L., Wei, H., Liu, C., & Ge, Z. (2024). OneChart: Purify the chart structural extraction via one auxiliary token. arXiv preprint arXiv:2404.09987.
 - [52] Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., & Finn, C. (2023).

Direct preference optimization: Your language model is secretly a reward model. In Advances in Neural Information Processing Systems 36.

Impressum

Founders	Zhengjie Gao, Xinyu Song
Editor in Chief	Ao Feng, Chengdu University of Information Technology, China
Executive Editor	Zhengjie Gao, Geely University of China, China
Editorial Board	Jing Hu, Huazhong University of Science and Technology, China Xiaohu Du, Huazhong University of Science and Technology, China Xiangkui Li, Harbin University of Science and Technology, China Zuopeng Liu, Goettingen University, Germany Xinyu Song, Geely University of China, China
Young Editorial Board	Min Liao, Geely University of China, China Tao Zheng, Geely University of China, China Chong Li, Chongqing University, China Ruiqin Fan, Sehan University, Korea Ziyang Liu, Jiangsu Normal University, China Qiwei Liu, Urumqi Vocational University, China Minqiu Kuang, Hunan Agricultural University, China
Published By	Hong Kong Dawn Clarity Press Limited Rm 9042, 9/F, Block B Chung Mei Centre, 15-17 Hing Yip Street, Kwun Tong, Kowloon, Hong Kong e-mail: ijaaa@dawnclarity.press <i>International Journal of Advanced AI Applications</i> is published monthly.
Editorial Policy	<i>International Journal of Advanced AI Applications</i> is directed to the international communities of scientific researchers in artificial intelligence, computers and electronic, from the universities, research units and industry. To differentiate from other similar journals, the editorial policy of IJAAA encourages the submission of original scientific papers that focus on the integration of the advanced AI applications. In particular, the following topics are expected to be addressed by authors: (1) Natural Language Processing (NLP): Conversational AI, machine translation, sentiment analysis, and context-aware dialogue systems. (2) Smart Cities and IoT Integration: AI for traffic optimization, energy management, waste reduction, and urban infrastructure. (3) Autonomous Systems and Robotics: Self-driving vehicles, drones, industrial automation, and human-robot collaboration.

-
- (4) Edge AI and Distributed Systems: Real-time processing, federated learning, and low-latency AI at the network edge.
 - (5) Creative and Generative AI: Art, music, and content generation using generative adversarial networks (GANs) and transformers.
 - (6) AI in Education and Industry: Adaptive learning platforms, intelligent tutoring systems, and AI-driven supply chain optimization.
- Ethical and Explainable AI (XAI): Fairness, transparency, and accountability in real-world AI deployment.
-