

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.

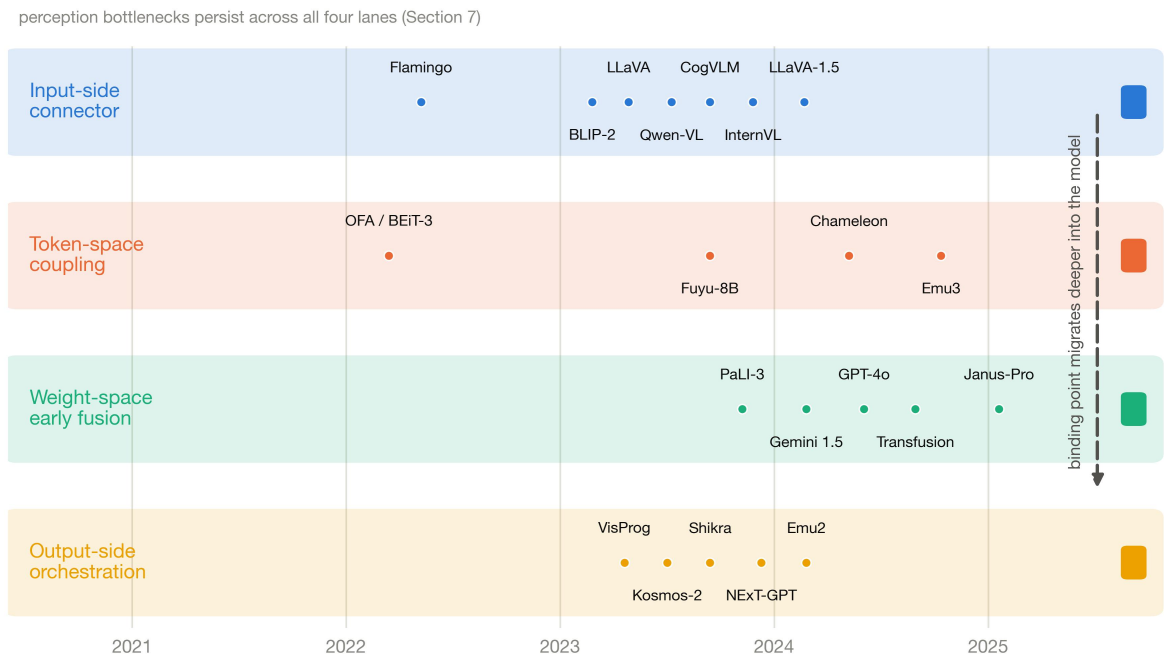


Figure 1. The coupling-locus taxonomy as a timeline (2020–2025). Each lane collects representative systems for one locus; over the period the binding point migrates from input-side connectors toward weight-space early fusion, while the perception benchmarks of Section 7 fail systems in all four lanes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

- 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.