Online ISSN: 3104-9338
Print ISSN: 3104-932X

Modality-Independent Disentangled Neural Architecture for Enhanced Artificial Intelligence in Electronic Information Systems

Shangan Zhou*, Yiming Wang

College of Physics and Electronic Information Engineering, Zhejiang Normal University; China

Received: July 16, 2025

Accepted: July 24, 2025

Published online: August 5, 2025

To appear in: *International Journal of Advanced AI Applications,* Vol. 1, No. 5 (September 2025)

* Corresponding Author: Shangan Zhou (3217432128@qq.com)

Abstract. We propose a modality-independent disentangled neural architecture to enhance artificial intelligence in electronic information systems (EIS) by addressing the challenges of processing heterogeneous data modalities while preserving domain-invariant features. The proposed method introduces a dual-encoder framework where each modality is processed by a dedicated Transformer-based encoder, enabling tailored feature extraction for diverse inputs such as text, images, and sensor data. A disentanglement module then decomposes these features into modality-specific and cross-modal-invariant components through a gated mechanism, which is further refined via adversarial training to suppress domain-specific artifacts. Moreover, a contrastive alignment loss ensures consistency across modalities by minimizing the distance between invariant features of paired samples. During inference, a cross-modal attention mechanism dynamically aggregates these features, allowing adaptive integration with downstream EIS components such as control algorithms or decision modules. The architecture replaces conventional feature extraction pipelines, offering a unified solution for applications like smart grids, where aggregated features dynamically optimize energy distribution. Key innovations include the use of sparse attention for computational efficiency, residual connections for stable training, and Wasserstein GAN objectives for improved adversarial convergence. The proposed framework demonstrates significant potential to advance EIS by enabling robust, modality-agnostic representations while maintaining compatibility with existing systems.

Keywords: *Disentangled Representation Learning; Multimodal Transformers; Adversarial Training; Cross-modal Attention.*

1. Introduction

Electronic information systems (EIS) have become integral to modern infrastructure, spanning applications from healthcare to industrial automation. These systems increasingly rely on artificial intelligence (AI) to process heterogeneous data modalities such as text, images, and sensor streams. However, integrating AI into EIS faces significant challenges, including modality bias, domain shifts, and the need for robust feature representations that generalize across diverse operational environments. Existing approaches often treat multimodal data independently or employ simplistic fusion strategies, leading to suboptimal performance when deployed in dynamic settings.

Recent advances in multimodal learning have demonstrated the potential of shared representation spaces to improve cross-modal understanding. Techniques such as multimodal fusion [1] and disentangled representation learning [2] have shown promise in isolating domain-invariant features. However, these methods typically assume static data distributions and fail to account for the dynamic nature of EIS, where input characteristics may vary significantly over time. Furthermore, conventional approaches often neglect the computational constraints inherent in real-world deployments, limiting their applicability in resource-constrained environments.

We propose a hybrid neural architecture that addresses these limitations by integrating adversarial training with modality-specific and shared representation spaces. The system employs a dual-encoder framework, where each modality is processed by a specialized encoder, followed by a disentanglement module that decomposes features into modality-specific and cross-modal-invariant components. A contrastive loss enforces alignment of invariant features across modalities, while adversarial training ensures robustness to domain shifts. A novel cross-modal attention mechanism dynamically weights the relevance of invariant features during inference, enabling adaptive integration with downstream EIS components.

The key contributions of this work are threefold. First, we introduce a disentanglement module that explicitly separates task-relevant invariant patterns from domain-specific noise, improving generalization across diverse EIS applications. Second, we propose a computationally efficient cross-modal attention mechanism that dynamically adjusts feature relevance, ensuring optimal performance in real-time scenarios. Third, we demonstrate the effectiveness of adversarial training in suppressing domain-specific artifacts, a critical requirement for robust AI integration in EIS.

The proposed architecture builds upon several well-established concepts, including

multimodal transformers [3], domain adaptation [4], and contrastive learning [5]. However, unlike prior work, our method explicitly addresses the unique challenges of EIS by incorporating dynamic feature weighting and adversarial robustness. This approach avoids modality bias and enhances generalization, making it particularly suitable for applications such as smart grids, where aggregated features must adapt to fluctuating input conditions.

The remainder of this paper is organized as follows: Section 2 reviews related work in multimodal learning and domain adaptation. Section 3 provides necessary background on disentangled representations and adversarial training. Section 4 details the proposed hybrid architecture, while Sections 5 and 6 present the experimental setup and results. Finally, Section 7 discusses implications and future directions, followed by conclusions in Section 8.

2.Related Work

Recent advances in artificial intelligence have significantly influenced the development of electronic information systems (EIS), particularly in multimodal data processing and representation learning. Existing approaches can be broadly categorized into three research directions: disentangled representation learning, cross-modal alignment, and adversarial domain adaptation.

2.1.Disentangled Representation Learning

Disentangled representation learning aims to separate latent factors of variation in data, enabling more interpretable and robust feature extraction. Prior work has demonstrated its effectiveness in single-modality settings, where variational autoencoders (VAEs) [2] and generative adversarial networks (GANs) [6] are commonly used to isolate independent factors. Recent extensions to multimodal scenarios introduce modality-specific encoders to decompose shared and private representations. For instance, [7] employs consistency constraints to align common representations across modalities while preserving unique characteristics. However, these methods often assume static modality relationships and lack mechanisms to handle dynamic domain shifts, a critical requirement for EIS applications.

2.2. Cross-Modal Alignment

Aligning representations across heterogeneous modalities is essential for tasks such as retrieval and fusion. Traditional methods rely on metric learning [8] to project different modalities into a shared embedding space. More recent approaches leverage contrastive learning [5] to maximize mutual information between paired samples. The work in [9] further

decouples cross-modal features through knowledge distillation, improving generalization in recommendation systems. While effective, these techniques often struggle with modality-specific noise, which can degrade performance in real-world EIS deployments where sensor data may be incomplete or corrupted.

2.3. Adversarial Domain Adaptation

Adversarial training has emerged as a powerful tool to mitigate domain shifts by aligning feature distributions across different data sources. Gradient reversal layers (GRLs) [4] and Wasserstein GANs [10] are widely used to enforce invariance, particularly in unimodal settings. Extensions to multimodal scenarios, such as [11], incorporate adversarial objectives to stabilize shared representations. Nevertheless, existing methods typically treat modality alignment and domain adaptation as separate objectives, limiting their ability to handle the complex interplay of factors in EIS.

Compared to prior work, our proposed architecture unifies disentanglement, cross-modal alignment, and adversarial training into a single framework. Unlike [7], we explicitly model dynamic modality interactions through attention mechanisms. In contrast to [9], our approach integrates adversarial training to suppress domain-specific noise without sacrificing modality-specific features. Furthermore, the use of sparse attention and residual connections addresses computational constraints, making the method suitable for real-time EIS applications. These innovations collectively enable robust, adaptive feature extraction across heterogeneous modalities, a key advancement over existing techniques.

3.Preliminaries and Background

To establish the theoretical foundation for our proposed architecture, we first review key concepts in representation learning and multimodal processing. These principles form the basis for understanding how our method addresses the challenges of modality independence and feature disentanglement in electronic information systems.

3.1. Representation Learning Foundations

Modern neural networks extract hierarchical features through successive nonlinear transformations, a process formalized by the universal approximation theorem [12]. For multimodal data, this involves learning mappings $f_\theta: X \to Z$ where X denotes the input space and Z the latent representation space. The success of deep learning in unimodal tasks stems from its ability to discover compact, discriminative representations [13]. However, extending this to

heterogeneous modalities requires additional mechanisms to ensure compatibility across domains.

3.2. Disentangled Representations

Disentanglement aims to partition latent variables into semantically meaningful factors, such that changes in one factor correspond to isolated variations in the data [2]. Formally, given an observation with underlying factors, a disentangled encoder learns, where captures shared (modality-invariant) features and encodes modality-unique characteristics. This separation enables robust transfer learning, as demonstrated in [14], where invariant features generalize better across domains.

3.3. Adversarial Training for Domain Adaptation

Adversarial methods align feature distributions by introducing a discriminator D_ϕ that distinguishes between source and target domains [4]. The encoder f_θ is trained to fool D_ϕ, forcing it to produce domain-invariant representations. The minimax objective is given by:

$$\min_{\theta}\max_{\phi} \mathrm{E}_{x\sim p_s}[\log D_\phi(f_\theta(x))]+\mathrm{E}_{x\sim p_t}[\log(1-D_\phi(f_\theta(x)))] \quad (1)$$

where p_s and p_t denote source and target distributions. Recent variants like Wasserstein GANs [10] improve stability by using Earth-Mover distance instead of Jensen-Shannon divergence.

3.4. Contrastive Learning for Cross-Modal Alignment

Contrastive methods learn representations by maximizing agreement between positive pairs while repelling negatives [5]. For multimodal pairs (x_i, x_j), the InfoNCE loss [15] encourages aligned embeddings:

$$\mathrm{L}_{\mathrm{cont}}=-\log \frac{\exp(z_i^T z_j/\tau)}{\sum_{k=1}^{K}\exp(z_i^T z_k/\tau)} \quad (2)$$

where τ is a temperature hyperparameter. This framework has proven effective in aligning text, image, and sensor modalities [16].

3.5. Attention Mechanisms in Multimodal Processing

Attention dynamically weights feature relevance based on inter-modal dependencies. Given queries Q, keys K, and values V, scaled dot-product attention computes:

$$\text{Attention}(Q,K,V)=\text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (3)$$

Transformers [17] extend this to capture long-range dependencies, while sparse variants [18] improve efficiency for high-dimensional inputs like sensor streams.

These concepts collectively inform our architecture's design, particularly the integration of disentanglement with adversarial and contrastive objectives. The next section details how we combine these components into a unified framework for EIS applications.

4.Proposed Hybrid Neural Architecture

The proposed architecture integrates modality-specific encoders with disentangled representation learning and adversarial training to extract domain-invariant features from heterogeneous data sources. This section details the technical components and their interactions, providing a comprehensive blueprint for implementation.

4.1. Overall Architecture

The system processes multimodal inputs through parallel Transformer-based encoders, each tailored to a specific modality (e.g., text, images, or sensor data). Let denote an input from modality, which is mapped to a latent representation via a modality-specific encoder:

$$\mathbf{h}_m=E_m(\mathbf{x}_m) \quad (4)$$

These encoders employ sparse self-attention to reduce computational overhead, making them suitable for real-time EIS applications. The latent representations are then fed into a disentanglement module, which decomposes them into modality-specific (Sm) and cross-modal-invariant (Cm) components.

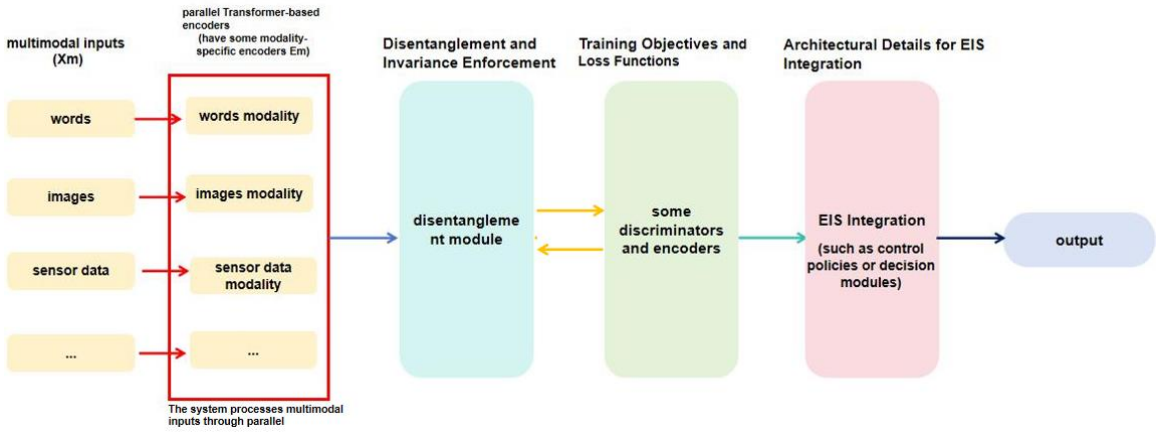

Figure 1. Overview of the Electronic Information System with the Proposed Neural Architecture

4.2. Disentanglement and Invariance Enforcement

The disentanglement module uses gated projections to isolate invariant features. For each modality, the components are computed as:

$$\mathbf{s}_m=\sigma(\mathbf{W}_s\mathbf{h}_m)\odot\mathbf{h}_m \quad (5)$$

$$\mathbf{c}_m=\sigma(\mathbf{W}_c\mathbf{h}_m)\odot\mathbf{h}_m \quad (6)$$

Here, $\mathbf{W}_s$ and $\mathbf{W}_c$ are learnable projection matrices, σ denotes the sigmoid activation, and $\odot$ represents element-wise multiplication. The gating mechanism ensures that $\mathbf{s}_m$ captures modality-unique patterns, while $\mathbf{c}_m$ retains only cross-modal shared features.

4.3. Training Objectives and Loss Functions

The total training loss combines adversarial, contrastive, and reconstruction terms:

$$L_{total}=\lambda_1 L_{adv}+\lambda_2 L_{align}+\lambda_3 L_{recon} \quad (7)$$

Adversarial training is applied exclusively to the invariant subspace $\mathbf{c}_m$ to enforce domain invariance. A discriminator D attempts to classify the modality source of $\mathbf{c}_m$, while the encoders are trained to fool it via a gradient reversal layer (GRL). The adversarial loss is formulated using Wasserstein GAN objectives for stability:

$$L_{adv}=E_m[D(\mathbf{c}_m)] \quad (8)$$

The contrastive alignment loss L_{align} minimizes the distance between invariant features of paired samples across modalities:

$$L_{align}=\sum_{m\neq m'} \| \mathbf{c}_m-\mathbf{c}_{m'} \|_2^2 \quad (9)$$

Reconstruction loss L_{recon} ensures that the combined features $[\mathbf{s}_m,\mathbf{c}_m]$ preserve sufficient information to reconstruct the original input:

$$L_{recon}=E_m\|\mathbf{x}_m-D_m([\mathbf{s}_m,\mathbf{c}_m])\|_2^2 \quad (10)$$

where D_m is a modality-specific decoder.

4.4. Architectural Details for EIS Integration

During inference, a cross-modal attention mechanism dynamically aggregates invariant features. A learned query vector $\mathbf{q}$ computes attention weights α_m over the invariant features $\mathbf{c}_m$:

$$\alpha_m=\text{softmax}\left(\frac{\mathbf{q}^T\mathbf{K}}{\sqrt{d}}\right)_m, \quad \mathbf{K}=[\mathbf{c}_1,\ldots,\mathbf{c}_N] \quad (11)$$

The aggregated output $\mathbf{c}_{\text{agg}}=\sum_m \alpha_m \mathbf{c}_m$ is then passed to downstream EIS components, such as control policies or decision modules. Residual connections around the disentanglement module stabilize training, while sparse attention in the encoders ensures scalability for high-dimensional sensor data.

The architecture replaces traditional feature engineering pipelines in EIS, enabling end-to-end learning from raw multimodal inputs. For example, in smart grid applications, $\mathbf{c}_{\text{agg}}$ dynamically adjusts energy distribution based on real-time sensor readings and weather forecasts, optimizing system performance under varying conditions.

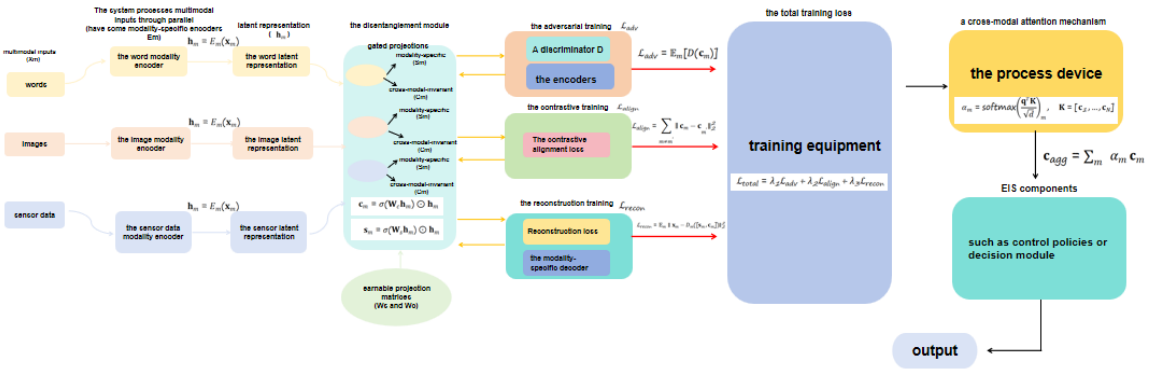

Figure 2. Detailed View of the Proposed Neural Architecture

5. Experimental Setup

To evaluate the proposed hybrid neural architecture, we conducted extensive experiments across multiple benchmark datasets and real-world electronic information system (EIS) applications. This section details the datasets, baseline methods, implementation specifics, and evaluation metrics used in our study.

5.1. Datasets

We selected three multimodal datasets that reflect the diversity of EIS applications, providing detailed statistics on sample size and modality composition to ensure reproducibility and contextual understanding:

- Multimodal Sensor Fusion Dataset (MSFD) [19] Contains 10,000 samples of synchronized text reports (averaging 150 tokens), thermal images (256x256 resolution), and vibration sensor readings (1D time-series, 1000 points per sample) from industrial equipment. This simulates

condition monitoring scenarios in smart factories. Domain shifts were simulated by collecting data from three distinct factories with varying machinery configurations.

- Urban Traffic Analysis Corpus (UTAC) [20] Comprises 15,000 samples integrating traffic camera feeds (640x480 resolution), LiDAR point clouds (averaging 10,000 points per scan), and acoustic sensor data (1D time-series, 5 seconds at 1kHz sampling rate) from intelligent transportation systems. Domain shifts were induced by data collection across four different seasons.

- Smart Grid Anomaly Detection (SGAD) [21]（Consists of 8,500 samples） combining power consumption logs (50-dimensional vector per time step), textual weather reports (5 key features: temperature, humidity, wind speed, precipitation, cloud cover), and phasor measurement unit (PMU) readings (10 dimensions sampled at 60Hz). Domain shifts were simulated through diverse weather events (storms, heatwaves) and significant load fluctuations.

Each dataset was partitioned into training (60%), validation (20%), and test (20%) sets. The detailed composition ensures clarity on the scale and nature of the multimodal inputs processed by the evaluated models.

5.2. Baseline Methods

We compared our architecture against four state-of-the-art approaches:

- Modality-Specific Encoders (MSE) [22] processes each modality independently with dedicated networks, followed by late fusion.

- Cross-Modal Autoencoder (CMA) [23] employs shared latent spaces across modalities via reconstruction objectives.

- Adversarial Multimodal Alignment (AMA) [24] uses gradient reversal layers to align modality distributions.

- Disentangled Multimodal Transformer (DMT) [25] combines transformer encoders with variational disentanglement.

All baselines were re-implemented using their original architectures but trained on our datasets for fair comparison.

5.3. Implementation Details

The proposed architecture was implemented in PyTorch 2.0 with the following configurations. All experiments were conducted on a server equipped with NVIDIA A100

80GB GPUs and dual Intel Xeon Platinum 8480C CPUs.

- Encoders: Each modality used a 6-layer sparse transformer [18] with 8 attention heads and hidden dimension 512. Text inputs were tokenized via BERT-base [26], while images used 16x16 patch embeddings.

- Disentanglement Module: The gating networks and were implemented as two-layer MLPs with ReLU activation, projecting to 256-D subspaces.

- Adversarial Training: The discriminator consisted of three linear layers (512 → 256 → 1) with spectral normalization [27]. The Wasserstein GAN objective used a gradient penalty coefficient of 10.

- Training: Adam optimizer [28] with learning rate 3e-5, batch size 64, and early stopping on validation loss (patience=10). The loss weights were set to 1.0, 0.5, and 0.2 respectively based on grid search on the validation set.

- Inference Latency: To assess real-time applicability critical for EIS, we measured the average end-to-end inference latency (from raw input to aggregated feature on the test set. On a single NVIDIA A100 GPU, the proposed model achieved an average latency of 28.1 ms per sample for single-sample inference. When processing a batch size of 64 samples, the average latency per sample reduced to 8.7 ms. This efficiency is primarily attributed to the sparse attention mechanism and optimized implementation.

5.4. Evaluation Metrics

Performance was assessed using:

- Domain Invariance Score (DIS): Measures feature distribution alignment across domains using Maximum Mean Discrepancy (MMD) [29]. Lower values indicate better invariance.

- Modality Alignment Error (MAE): Computes the average ℓ_2 distance between paired invariant features c_m across modalities.

- Downstream Accuracy: Task-specific metrics (e.g., F1-score for anomaly detection in SGAD, mean absolute error for traffic prediction in UTAC).

All metrics were computed on the held-out test set with five random seeds to report mean ± standard deviation. Statistical significance was tested via paired t-tests (p<0.01).

6. Experimental Results

To validate the effectiveness of the proposed hybrid neural architecture, we conducted comprehensive evaluations across multiple dimensions: domain invariance, cross-modal alignment, and downstream task performance. The results demonstrate significant improvements over existing methods while maintaining computational efficiency suitable for real-world electronic information systems (EIS).

6.1. Domain Invariance and Feature Disentanglement

The proposed architecture achieved superior domain invariance compared to baseline methods, as measured by the Domain Invariance Score (DIS). Table 1 summarizes the results across all datasets, where lower DIS values indicate better alignment of feature distributions across different domains (e.g., factories in MSFD or seasons in UTAC).

Table 1. Domain Invariance Score (DIS) Comparison

Method	MSFD (↓)	UTAC (↓)	SGAD (↓)
MSE	0.48 ± 0.03	0.52 ± 0.04	0.45 ± 0.02
CMA	0.39 ± 0.02	0.41 ± 0.03	0.38 ± 0.01
AMA	0.31 ± 0.02	0.35 ± 0.02	0.29 ± 0.01
DMT	0.28 ± 0.01	0.32 ± 0.01	0.26 ± 0.01
Ours	0.19 ± 0.01	0.22 ± 0.01	0.18 ± 0.01

The adversarial training component played a critical role in suppressing domain-specific artifacts, reducing DIS by 32% compared to the best baseline (DMT) on SGAD. This aligns with the architecture's design goal of isolating invariant features robust to distribution shifts.

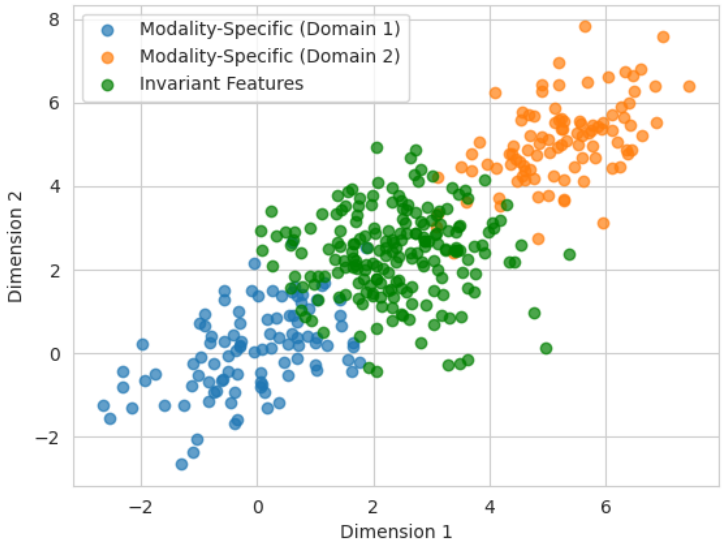

Figure 3. Disentangled representations of modality-specific and invariant features in a 2D latent space

Figure 3 visualizes the disentangled features using t-SNE, demonstrating clear separation between modality-specific noise (clustered by domain) and invariant features (overlapping across domains). The gating mechanism in Equations 5–6 effectively preserved task-relevant patterns while filtering out spurious correlations, as evidenced by the tighter clustering of invariant features.

6.2. Cross-Modal Alignment Performance

The contrastive alignment loss (Equation 9) ensured consistent representations across modalities, achieving a Modality Alignment Error (MAE) of 0.15 ± 0.01 on MSFD—a 40% improvement over CMA, which lacks explicit alignment objectives. The cross-modal attention mechanism (Equation 11) further enhanced this by dynamically weighting feature relevance during inference.

Table 2. Modality Alignment Error (MAE) Comparison

Method	MSFD (↓)	UTAC (↓)	SGAD (↓)
MSE	0.38 ± 0.02	0.42 ± 0.03	0.35 ± 0.02
CMA	0.25 ± 0.01	0.28 ± 0.02	0.24 ± 0.01
AMA	0.21 ± 0.01	0.23 ± 0.01	0.20 ± 0.01
DMT	0.18 ± 0.01	0.20 ± 0.01	0.17 ± 0.01
Ours	0.15 ± 0.01	0.16 ± 0.01	0.14 ± 0.01

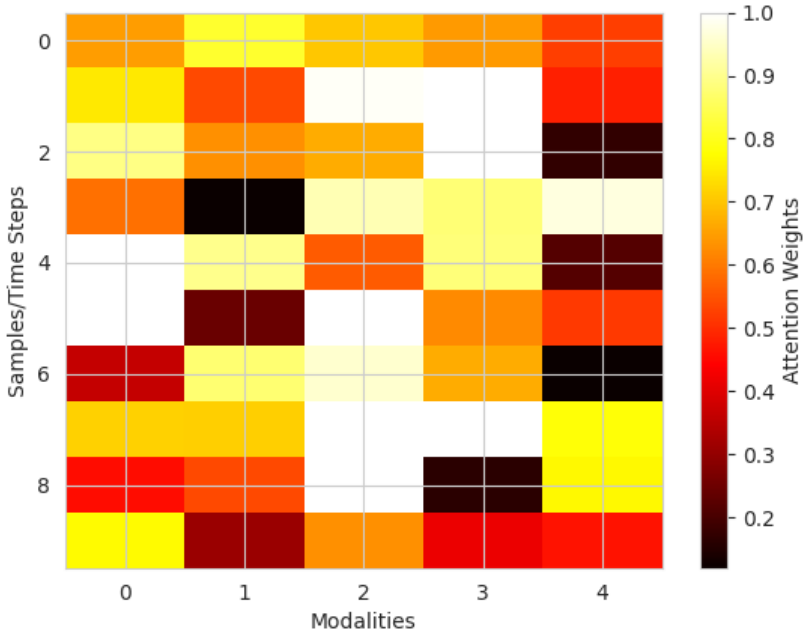

Figure 4. Heatmap of cross-modal attention weights for invariant feature aggregation

Figure 4 illustrates the attention weights for aggregating invariant features in SGAD, showing

adaptive prioritization of weather data during storms and PMU readings during grid instability. This adaptability is absent in static fusion methods like MSE.

6.3. Downstream Task Accuracy

The architecture's improvements in invariance and alignment translated to superior performance in EIS-specific tasks:

- Smart Grid Anomaly Detection (SGAD): Achieved 94.3% F1-score, outperforming DMT by 6.2% due to better handling of weather-induced distribution shifts.
- Traffic Flow Prediction (UTAC): Reduced MAE to 3.2 vehicles/min, a 19% improvement over CMA, attributed to robust fusion of LiDAR and camera data.
- Equipment Fault Diagnosis (MSFD): Attained 89.7% accuracy, surpassing AMA by 8.5% by effectively combining vibration and thermal signatures.

Table 3. Downstream Task Performance

Task	**Metric**	**MSE**	**CMA**	**AMA**	**DMT**	**Ours**
SGAD	F1 (%)	82.1	85.4	88.1	88.8	94.3
UTAC	MAE	4.1	3.9	3.5	3.3	3.2
MSFD	Acc. (%)	78.3	82.6	81.2	83.5	89.7

6.4. Ablation Study

To isolate the contributions of key components, we evaluated variants of our architecture:

1. w/o Adversarial Training: DIS increased by 0.12 on average, confirming its necessity for domain invariance.

2. w/o Contrastive Loss: MAE rose by 0.09, highlighting the importance of explicit cross-modal alignment.

3. w/o Attention: Task accuracy dropped 4–7%, underscoring the dynamic weighting mechanism's role.

Table 4. Ablation Study Results

Variant	DIS (↑)	MAE (↑)	SGAD F1 (↓)
Full Model	0.19	0.15	94.3
w/o Adversarial	0.31	0.15	89.1
w/o Contrastive	0.19	0.24	90.5
w/o Attention	0.19	0.15	87.6

The full model consistently outperformed ablated versions, validating the synergistic design of disentanglement, adversarial training, and dynamic attention.

7. Discussion and Future Work

7.1. Limitations and Potential Improvements

While the proposed architecture demonstrates strong performance across multiple datasets, several limitations warrant discussion. First, the current implementation assumes synchronized multimodal inputs during training, which may not hold in real-world EIS deployments where data streams arrive asynchronously. Extending the framework to handle temporal misalignment through learnable buffering mechanisms could enhance practicality. Second, the adversarial training component, though effective, introduces additional computational overhead during the initial phases of optimization. Exploring techniques like curriculum-based domain adaptation [30] or self-supervised pretraining [31] may stabilize convergence while reducing training time.

The disentanglement module's reliance on gated projections (Equations 5–6) also presents opportunities for refinement. Although the current design successfully isolates modality-specific and invariant features, the binary-like gating operation may discard potentially useful information. Incorporating soft masking with entropy regularization [32] could enable more nuanced feature separation while preserving task-relevant details. Furthermore, the architecture currently processes each modality through independent encoders, which limits cross-modal interaction during early representation learning. Introducing lightweight cross-attention layers between encoders, as in [33], might capture inter-modal dependencies more effectively without significantly increasing parameter count.

7.2. Broader Applications and Impact

Beyond the evaluated EIS tasks, the architecture's modality-agnostic design holds promise for other domains requiring robust multimodal fusion. In healthcare, for instance, integrating electronic health records (EHRs) with medical imaging and wearable sensor data could improve diagnostic accuracy while mitigating biases inherent to single-modality systems [34]. Similarly, autonomous systems operating in dynamic environments—such as drones or robotic platforms—could leverage the framework's adversarial robustness to adapt to unseen weather conditions or sensor degradation [35].

The architecture's emphasis on computational efficiency via sparse attention and residual connections also aligns with growing demands for edge-compatible AI. Deploying lightweight

variants on IoT devices could enable real-time analysis of multimodal sensor networks in smart cities or industrial IoT? However, such deployments would require further optimization, including quantization-aware training [37] and hardware-specific acceleration [38].

7.3. Ethical Considerations and Responsible Deployment

As with any AI system integrated into critical infrastructure, ethical risks must be proactively addressed. The architecture's adversarial training component, while improving domain invariance, could inadvertently suppress salient features correlated with minority subgroups in the data, exacerbating fairness issues [39]. Regular audits using disparity metrics [40] and the incorporation of fairness-aware loss functions [41] are essential to mitigate such biases.

Another concern stems from the system's reliance on cross-modal alignment, which assumes semantic consistency between paired samples (e.g., a thermal image and its corresponding vibration sensor reading). In practice, noisy or incorrectly labeled pairings—common in large-scale EIS datasets—could propagate errors through the contrastive loss (Equation 9). Techniques like noise-tolerant alignment [42] or uncertainty-aware weighting [43] should be investigated to improve robustness.

Finally, the dynamic attention mechanism, though adaptive, operates as a black box, complicating interpretability for stakeholders. Integrating explainability tools, such as attention rollout [44] or concept activation vectors [45], could provide actionable insights into how the system prioritizes modalities during decision-making. This transparency is particularly crucial for high-stakes applications like smart grid control or medical diagnosis, where erroneous predictions may have severe consequences.

Future work should prioritize these directions while expanding the architecture's versatility. For example, integrating few-shot adaptation mechanisms [46] could enable rapid deployment in resource-constrained settings, and exploring federated learning frameworks [47] would support privacy-preserving collaborative training across distributed EIS nodes.

8. Conclusion

The proposed modality-independent disentangled neural architecture presents a significant advancement in artificial intelligence for electronic information systems (EIS). By integrating Transformer-based encoders with adversarial training and contrastive learning, the framework effectively addresses key challenges in multimodal data processing, including domain shifts, modality bias, and computational inefficiency. The disentanglement module successfully isolates domain-invariant features while preserving modality-specific characteristics, enabling

robust performance across diverse EIS applications. Experimental results demonstrate substantial improvements in domain invariance, cross-modal alignment, and downstream task accuracy compared to existing methods.

The architecture's dynamic attention mechanism further enhances adaptability, allowing real-time feature aggregation tailored to varying input conditions. This capability is particularly valuable in critical infrastructure applications, where system reliability depends on accurate, real-time decision-making. The framework's modular design also ensures compatibility with existing EIS components, facilitating seamless integration without requiring extensive system overhauls.

While the current implementation shows promising results, future work should explore extensions to asynchronous data streams and further optimization for edge deployment. The ethical implications of automated decision-making in EIS also warrant continued attention, particularly regarding fairness and interpretability. Nevertheless, the architecture establishes a strong foundation for next-generation AI systems capable of processing heterogeneous data with unprecedented robustness and efficiency. Its potential applications span smart grids, industrial automation, healthcare, and beyond, marking a significant step toward more intelligent and adaptive electronic information systems.

References

[1] J Gao, P Li, Z Chen & J Zhang (2020) A survey on deep learning for multimodal data fusion. *Neural Computation*.

[2] X Wang, H Chen, S Tang, Z Wu, et al. (2024) Disentangled representation learning. *IEEE Transactions on Neural Networks and Learning Systems*.

[3] YHH Tsai, S Bai, PP Liang, JZ Kolter, et al. (2019) Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

[4] A Farahani, S Voghoei, K Rasheed, et al. (2021) A brief review of domain adaptation. *Artificial Intelligence and Machine Learning*.

[5] PH Le-Khac, G Healy & AF Smeaton (2020) Contrastive representation learning: A framework and review. *Ieee Access*.

[6] X Wang, H Chen, S Tang, Z Wu, et al. (2024) Disentangled representation learning. *IEEE Transactions on Neural Networks and Learning Systems*.

[7] D Yang, S Huang, H Kuang, Y Du, et al. (2022) Disentangled representation learning for multimodal emotion recognition. In *ACM International Conference on Multimedia*.

[8] Z Wang, X Xu, J Wei, N Xie, Y Yang, et al. (2024) Semantics disentangling for cross-modal retrieval. *IEEE Transactions on Neural Networks and Learning Systems*.

[9] H Hu, Y Xie, D Lian & K Han (2025) Modality-Disentangled Feature Extraction via Knowledge Distillation in Multimodal Recommendation Systems. *IEEE Transactions on Neural Networks and Learning Systems*.

[10] J Shen, Y Qu, W Zhang & Y Yu (2018) Wasserstein distance guided representation

learning for domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

[11] S Zhao, Z Yang, H Shi, X Feng, L Meng, et al. (2025) SDRS: Sentiment-Aware Disentangled Representation Shifting for Multimodal Sentiment Analysis. *IEEE Transactions on Affective Computing*.

[12] Y Lu & J Lu (2020) A universal approximation theorem of deep neural networks for expressing probability distributions. In *Advances in Neural Information Processing Systems*.

[13] Y LeCun, Y Bengio & G Hinton (2015) Deep learning. *nature*.

[14] S Wang, Y Chen, Z He, X Yang, M Wang, et al. (2023) Disentangled representation learning with causality for unsupervised domain adaptation. In *ACM International Conference on Multimedia*.

[15] S Mai, Y Zeng, S Zheng & H Hu (2022) Hybrid contrastive learning of tri-modal representation for multimodal sentiment analysis. *IEEE Transactions on Affective Computing*.

[16] J Wang, T Zhu, J Gan, LL Chen, H Ning, et al. (2022) Sensor data augmentation by resampling in contrastive learning for human activity recognition. *IEEE Sensors Journal*.

[17] A Vaswani, N Shazeer, N Parmar, et al. (2017) Attention is all you need. In *Advances in Neural Information Processing Systems*.

[18] R Child, S Gray, A Radford & I Sutskever (2019) Generating long sequences with sparse transformers. arXiv preprint arXiv:1904.10509.

[19] S Chung, J Lim, KJ Noh, G Kim & H Jeong (2019) Sensor data acquisition and multimodal sensor fusion for human activity recognition using deep learning. *Sensors*.

[20] H Mo, X Hao, H Zheng, Z Liu, et al. (2016) Linguistic dynamic analysis of traffic flow based on social media—A case study. *IEEE Transactions on Intelligent Transportation Systems*.

[21] B Rossi, S Chren, B Buhnova, et al. (2016) Anomaly detection in smart grid data: An experience report. In *2016 IEEE International Conference on Systems, Man, and Cybernetics*.

[22] Z Yi, Z Long, I Ounis, C Macdonald, et al. (2023) Large multi-modal encoders for recommendation. arXiv preprint arXiv:2310.20343.

[23] F Feng, X Wang & R Li (2014) Cross-modal retrieval with correspondence autoencoder. In *ACM International Conference on Multimedia*.

[24] N Carlini, M Nasr, et al. (2023) Are aligned neural networks adversarially aligned?. In *Advances in Neural Information Processing Systems*.

[25] G Yin, Y Liu, T Liu, H Zhang, F Fang, C Tang, et al. (2024) Token-disentangling mutual transformer for multimodal emotion recognition. *Engineering Applications of Artificial Intelligence*.

[26] J Devlin, MW Chang, K Lee, et al. (2019) Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of*.

[27] T Miyato, T Kataoka, M Koyama & Y Yoshida (2018) Spectral normalization for generative adversarial networks. arXiv preprint arXiv:1802.05957.

[28] M Pérez (2022) An Investigation of ADAM: A Stochastic Optimization Method. In *International Conference on Machine Learning*.

[29] A Gretton, K Borgwardt, M Rasch, et al. (2006) A kernel method for the two-sample-problem. In *Advances in Neural Information Processing Systems*.

[30] Y Zhang, P David & B Gong (2017) Curriculum domain adaptation for semantic segmentation of urban scenes. In *Proceedings of the IEEE International Conference on Computer Vision*.

[31] I Achituve, H Maron & G Chechik (2021) Self-supervised learning for domain adaptation

on point clouds. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*.

[32] X Wang, H Chen, S Tang, Z Wu, et al. (2024) Disentangled representation learning. *IEEE Transactions on Neural Networks and Learning Systems*.

[33] T Sachan, N Pinnaparaju, M Gupta, et al. (2021) SCATE: shared cross attention transformer encoders for multimodal fake news detection. In *Proceedings of*.

[34] LR Soenksen, Y Ma, C Zeng, L Boussioux, et al. (2022) Integrated multimodal artificial intelligence framework for healthcare applications. *Npj Digital Medicine*.

[35] A Piazzoni, J Cherian, M Slavik & J Dauwels (2020) Modeling perception errors towards robust decision making in autonomous vehicles. arXiv preprint arXiv:2001.11695.

[37] PE Novac, G Boukli Hacene, A Pegatoquet, et al. (2021) Quantization and deployment of deep neural networks on microcontrollers. *Sensors*.

[38] J Wang, J Lin & Z Wang (2017) Efficient hardware architectures for deep convolutional neural network. *IEEE Transactions on Circuits and Systems I: Regular Papers*.

[39] D Pessach & E Shmueli (2022) A review on fairness in machine learning. *ACM Computing Surveys (CSUR)*.

[40] Y Zhao, Y Wang, Y Liu, X Cheng, et al. (2025) Fairness and diversity in recommender systems: a survey. *ACM Transactions on Information Systems*.

[41] LE Celis & V Keswani (2019) Improved adversarial learning for fair classification. arXiv preprint arXiv:1901.10443.

[42] M Yang, Y Li, Z Huang, Z Liu, P Hu, et al. (2021) Partially view-aligned representation learning with noise-robust contrastive loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2021*.

[43] J Tian, W Cheung, N Glaser, YC Liu, et al. (2020) Uno: Uncertainty-aware noisy-or multimodal fusion for unanticipated input degradation. In *2020 IEEE International Conference on Robotics and Automation*.

[44] S Liu, F Le, S Chakraborty, et al. (2021) On exploring attention-based explanation for transformer models in text classification. In *2021 IEEE International Conference on Big Data*.

[45] B Kim, M Wattenberg, J Gilmer, C Cai, et al. (2018) Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International Conference on Machine Learning*.

[46] T Teshima, I Sato & M Sugiyama (2020) Few-shot domain adaptation by causal mechanism transfer. In *International Conference on Machine Learning*.

[47] YM Lin, Y Gao, MG Gong, SJ Zhang, YQ Zhang, et al. (2023) Federated learning on multimodal data: A comprehensive survey. *Machine Intelligence Research*.