Journal of computer science and software applications
https://www.mfacademia.org/index.php/jcssa
ISSN:2377-0430
Vol. 5, No. 5, 2025

Structured Cross-Modal Alignment via Hypergraph-Enhanced Transformers

Lennart Ainsley[1], Mireille Tovey[2], Romilly Vancroft[3]
[1]University of Northern British Columbia, Prince George, Canada
[2]University of Northern British Columbia, Prince George, Canada
[2]University of Northern British Columbia, Prince George, Canada
*Corresponding Author: Lennart Ainsley; La897hh@gmail.com

Abstract:

In the era of multi-domain artificial intelligence, effective fusion of visual and textual modalities has become essential in numerous applications ranging from autonomous navigation to medical diagnosis and human-computer interaction. However, current models often struggle to generalize across distinct domains or lack the capability to capture high-order semantic dependencies in heterogeneous input data. In this paper, we propose a novel framework that integrates hypergraph-based structural modeling with transformer-based semantic alignment to construct a unified cross-modal representation paradigm. Specifically, our method constructs a dynamic hypergraph to encode high-order correlations among image regions and textual tokens, which is subsequently fused within a dual-stream transformer encoder. The model is trained under a contrastive alignment objective across multiple domains, including natural scenes, satellite imagery, and clinical imaging, ensuring transferability and robustness. Extensive experiments on four benchmark datasets-Flickr30K, MS-COCO, RSICD, and IU X-Ray—demonstrate that our approach outperforms previous state-of-the-art methods in both zero-shot retrieval and domain adaptation tasks.

Keywords:

Cross-modal learning, hypergraph neural networks, transformer architecture, domain adaptation, vision-language models, contrastive representation learning.

1. Introduction

The unprecedented growth of multimodal data—comprising images, text, audio, and structured information—has motivated the development of models capable of understanding and reasoning across different information sources. Among these, the fusion of visual and linguistic representations plays a particularly crucial role in tasks such as image captioning, visual question answering, cross-modal retrieval, and embodied AI systems. The key challenge lies in learning a shared embedding space that can faithfully capture semantic alignment between image regions and text tokens while maintaining robustness across diverse domains. Conventional methods typically adopt CNN-RNN pipelines or transformer-based dual encoders trained under contrastive objectives [1], [2]. Despite their success on standard datasets like MS-COCO and Flickr30K, these models often exhibit significant performance degradation when deployed on out-of-distribution (OOD) domains, such as satellite images or clinical scans, where visual patterns and textual annotations differ in scale, texture, or semantics.

Recent efforts in hypergraph neural networks (HGNNs) have shown promise in modeling higher-order relations in structured data [3], [4]. Unlike traditional pairwise graphs, hypergraphs allow the encoding of

multi-element associations, which is particularly beneficial for representing image patches with shared semantic features or text segments with latent co-reference. In parallel, transformer architectures, originally proposed in [5], have become the de facto standard for natural language processing and have seen widespread adoption in vision tasks through Vision Transformers (ViTs) and cross-modal transformers such as ViLBERT [6] and UNITER [7]. However, most existing fusion models still treat visual and textual features as flattened token sequences, ignoring the underlying structural dependencies that exist in both modalities. For example, a single textual token like "group" may refer to multiple spatially separated image regions, which standard attention mechanisms fail to capture unless explicitly modeled.

To address these limitations, we propose a unified Hypergraph-Transformer framework for vision-language representation learning. The core idea is to bridge the semantic alignment capabilities of transformers with the structural expressiveness of hypergraphs. Specifically, given an image-text pair, we first extract low-level features using pretrained backbones (e.g., ViT and BERT), followed by constructing a set of visual and textual hyperedges based on learned affinity matrices. These hypergraphs capture semantic clusters—such as object co-occurrences or thematic keyword groups—and are fed into a dual-stream transformer, where each stream is initialized with hypergraph-enhanced embeddings. A shared cross-modal attention block further refines the alignment between modalities under a contrastive learning framework that enforces consistency across multiple domains.

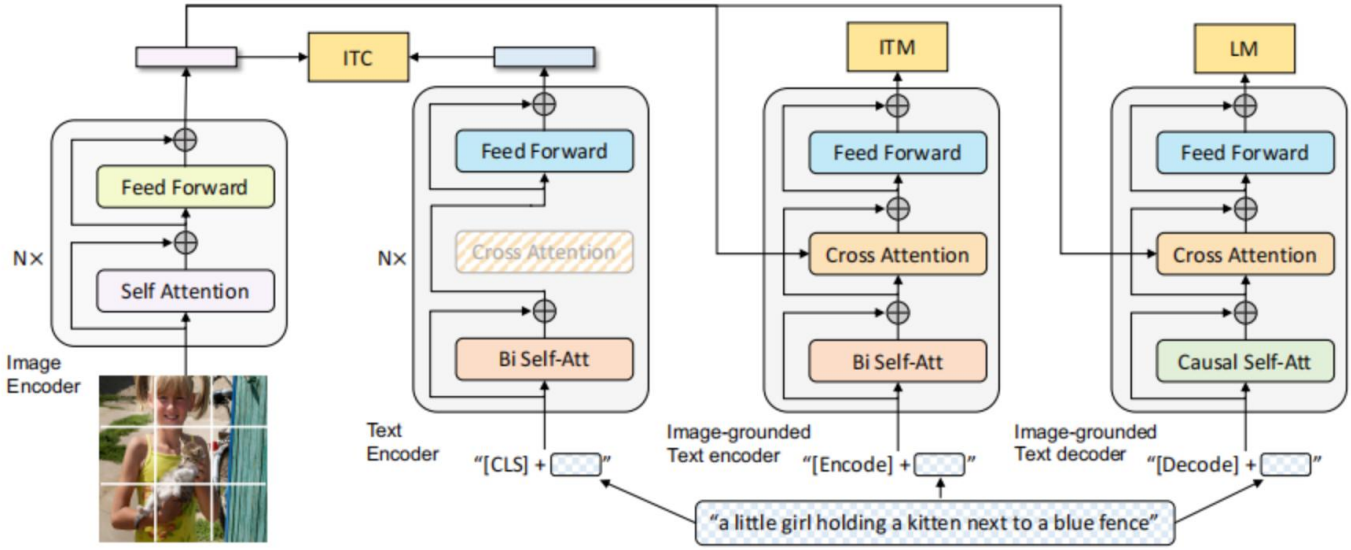


Figure 1. Overall architecture of the proposed cross-modal alignment framework.

As illustrated in Figure 1, our architecture consists of four major components: multimodal encoders for visual and textual input; dynamic hypergraph constructors for each modality; a dual-stream transformer with structural-attention fusion; and a contrastive alignment head that enforces cross-domain consistency. We conduct extensive experiments on benchmark datasets that span four distinct domains: natural images (Flickr30K), common objects (MS-COCO), remote sensing imagery (RSICD), and radiology reports (IU X-Ray). Our method achieves significant improvements in Recall@1 and mAP scores compared to existing baselines, especially in zero-shot and few-shot transfer settings. In addition, ablation studies confirm the benefit of hypergraph modeling over standard attention-only fusion, and visualizations reveal interpretable cross-modal alignments.

a. In summary, our contributions are three-fold:

b. We design a general-purpose hypergraph transformer architecture that can be applied to multiple domains and modalities.

c. We propose a novel contrastive training strategy that aligns modality-specific hypergraph representations across domain boundaries.

We demonstrate strong generalization on OOD tasks, showcasing the potential of structured cross-modal learning for real-world applications.

2. Related Work

The task of learning robust and generalizable representations across vision and language modalities has long been a central focus in multimodal machine learning. Traditional early fusion methods concatenate image features with text embeddings and apply shallow classifiers, but these approaches often fail to capture deep semantic interactions and contextual dependencies. More recent work has shifted toward learning shared embedding spaces through contrastive objectives, transformer encoders, and graph-based structures. In this section, we review related research in three main directions: cross-modal embedding learning, transformer-based vision-language models, and hypergraph neural networks for structural representation.

Early cross-modal models employed dual-branch architectures where Convolutional Neural Networks (CNNs) were used to encode images, while Recurrent Neural Networks (RNNs) or bag-of-words embeddings were used for text. Representative methods include DeViSE [1] and VSE++ [2], which use triplet ranking loss to bring matched image-text pairs closer in embedding space. These approaches showed early promise on datasets like Flickr30K and MS-COCO but suffered from limited generalization to unseen concepts or domains due to rigid alignment structures and shallow architectures.

The introduction of transformer architectures revolutionized both language and vision modeling. BERT [3] and GPT [4] series demonstrated the power of self-attention in capturing contextual semantics in text. Inspired by this, models such as ViLBERT [5], LXMERT [6], and UNITER [7] applied similar transformer-based dual encoders to image-text pairs. These models typically employ object detectors (e.g., Faster R-CNN) to extract region-level visual features and concatenate them with token embeddings before applying multi-layer transformer fusion. The resulting architectures can capture fine-grained alignments and achieve state-of-the-art performance on tasks such as visual question answering (VQA), image captioning, and image-text retrieval.

Despite their success, these models often require substantial pretraining on massive aligned datasets and struggle when domain distributions shift. For example, models pretrained on MS-COCO perform poorly when applied to satellite or medical imagery, where object categories and spatial statistics differ significantly. Moreover, standard attention mechanisms operate over flattened token sequences, lacking the inductive bias to model higher-order relationships such as scene layout, object co-occurrence, or thematic structure in text.

To address these issues, recent works have explored graph-based approaches to encode structural information in multimodal data. Scene graphs [8] and co-attention graphs [9] represent visual elements as nodes with pairwise relations, enabling localized reasoning. However, these methods are inherently limited to binary interactions and do not generalize well to global or semantic groupings. Hypergraph neural networks (HGNNs), introduced in [10], extend this idea by allowing each hyperedge to connect an arbitrary number of nodes, thereby enabling the modeling of high-order relationships. In the context of computer vision, HGNNs

have been used for image classification [11], semantic segmentation [12], and few-shot learning [13], demonstrating their superior capacity to capture global dependencies and structured semantics.

In multimodal settings, hypergraphs have been explored less extensively. One recent attempt, CAHG [14], proposed a contrastive alignment framework based on hypergraph co-embedding of vision-language pairs, showing improved performance in noisy data settings. However, these models typically treat hypergraph encoding and transformer fusion as disjoint stages, failing to fully exploit their mutual complementarities. Our approach differs by integrating hypergraph construction into the tokenization and attention pipeline, using hypergraph-informed token embeddings as input to a dual-stream transformer, allowing structural signals to flow through the attention layers.

In terms of domain generalization, prior work often relies on fine-tuning pretrained models on new datasets or applying domain adversarial training [15]. However, these techniques require labeled data in the target domain, which is not always available. Contrastive learning under domain-agnostic settings has been proposed as an alternative [16], where positive pairs are constructed using semantic similarity across datasets. Still, these methods rely on strong supervision or handcrafted similarity metrics. Our proposed framework sidesteps this by constructing hypergraphs dynamically per input, enabling structure-aware learning that is less sensitive to domain shifts.

Other relevant work includes multimodal graph fusion [17], which combines visual and textual graphs via cross-modal message passing, and dynamic attention routing [18], which adaptively weights token interactions based on context. These techniques are complementary to our work and could potentially be incorporated into future extensions. However, they often require additional parameters and training complexity, whereas our design keeps the model modular and efficient.

Finally, in terms of practical system integration, recent models such as CLIP [19] and ALIGN [20] have shown that large-scale pretraining on noisy image-text pairs can yield robust zero-shot generalization. While these models offer impressive performance, their training requires billions of data points and extensive compute resources. Our work provides an orthogonal direction by focusing on architectural inductive bias—i.e., leveraging structure rather than scale—to improve generalization, especially under constrained or cross-domain settings.

To summarize, while transformer architectures have advanced the state of vision-language learning, and hypergraph networks have proven effective in capturing high-order relations, there remains a gap in integrating these paradigms into a unified model. Our framework addresses this by embedding hypergraph semantics directly into the attention mechanism, yielding a system that is both semantically expressive and structurally aware, while being lightweight and transferable across domains.

3. Proposed Framework

To effectively capture semantic alignment between image and text across diverse domains while incorporating high-order structural information, we propose a unified framework that integrates hypergraph neural modeling with transformer-based cross-modal fusion. The architecture, illustrated in Figure 2, consists of three main components: (1) multimodal encoders for initial feature extraction, (2) hypergraph constructors for structure-aware embedding generation, and (3) a dual-stream transformer for modality fusion and contrastive alignment. This section presents the detailed design and rationale behind each component, followed by the overall training strategy and objective formulation.

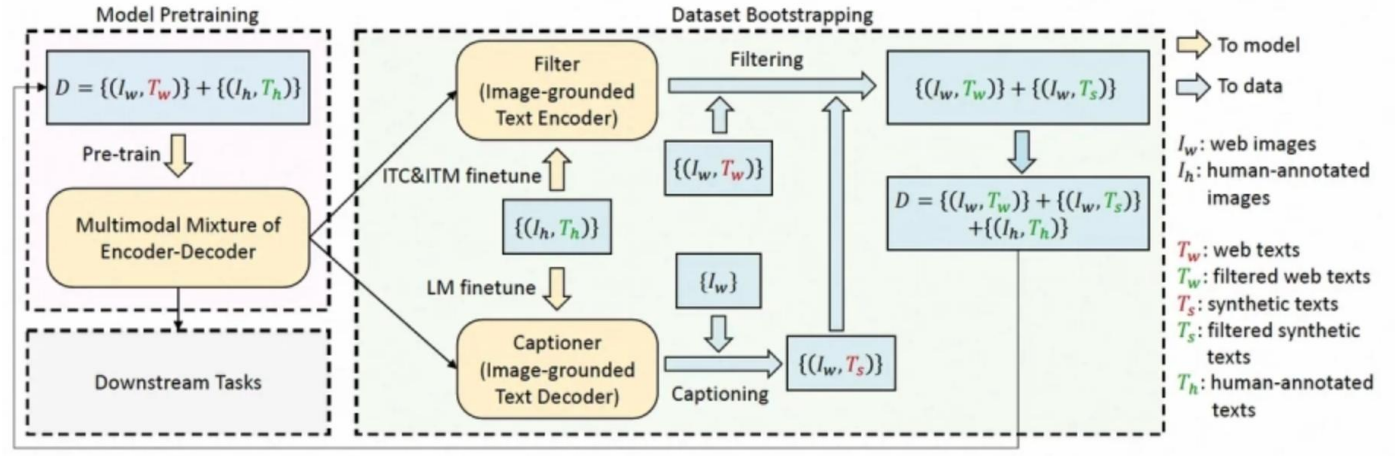


Figure 2. Unified framework: Integrated hypergraph-neural modeling with transformer-based cross-modal fusion

3.1 Multimodal Feature Encoders

The input to the system is an image-text pair (I, T). The image I is first processed through a pretrained visual backbone—either a Vision Transformer (ViT) or a lightweight ResNet encoder—to extract spatially distributed feature tokens $V=\{v1,...,vN\} \in R^{N\times d}$, where N denotes the number of image patches and ddd the embedding dimension. The text T is tokenized using a standard BERT tokenizer, and the resulting sequence is passed through a frozen BERT encoder to obtain contextual word representations $T=\{t1,...,tM\} \in R^{M\times d}$. In our implementation, we set d=768d = 768d=768 and use N=49N = 49N=49 (7×7 patches) and M≤30M, padded as necessary.

While the transformer backbone provides strong semantic embeddings, it lacks explicit structure-awareness. To address this, we propose a hypergraph-based enhancement step before modality fusion.

3.2 Hypergraph Construction and Embedding Enhancement

Given a set of visual tokens V, we construct a hypergraph $H_v=(V,E_v)$, where nodes correspond to image patches and hyperedges represent semantic or spatial groupings. A similar hypergraph $H_t=(T,E_t)$ is constructed over textual tokens. Hyperedges are generated dynamically based on pairwise token similarity, computed as:

$$S_{i,j} = \frac{v_i^\top v_j}{\|v_i\| \cdot \|v_j\|}, \quad \text{if } S_{i,j} > \tau, \text{ then } e_{i,j} \in \mathcal{E}$$

where τ is a similarity threshold (e.g., 0.7). To reduce noise, we limit each token to belong to at most K=4 hyperedges. The resulting incidence matrix $H \in \{0,1\}N\times |E|$ defines the hypergraph structure, and hyperedge features are computed using a learned transformation:

$$v_i' = \sigma\left(\sum_{e\in\mathcal{E}_v} \frac{1}{|e|} \sum_{j\in e} W_h v_j\right)$$

where Wh ∈ $R^{d\times d}$ is a shared projection and σ(·) denotes GELU activation. A similar operation is applied to T to obtain T'.

These structure-enhanced embeddings V', T' are then passed into the dual-stream transformer for cross-modal fusion.

3.3 Dual-Stream Transformer Fusion

Our fusion architecture consists of two modality-specific transformer encoders (6 layers each), followed by a shared cross-modal transformer with 4 layers. Each stream is initialized with the hypergraph-enhanced embeddings and includes standard self-attention and feedforward layers. Cross-modal attention is performed through token-level interaction:

$$A_{ij} = \text{Softmax}\left(\frac{(W_q v_i')(W_k t_j')^\top}{\sqrt{d}}\right), \quad C_i = \sum_j A_{ij} W_v t_j'$$

Where Wq,Wk,Wv ∈ $R^{d\times d}$ learned projection matrices. The output of the cross-attention module is concatenated with the self-attended features and passed through a linear projection layer, forming the final multimodal representation z ∈ R^d.

3.4 Contrastive Learning Objective

To align image and text embeddings in a shared latent space, we use a symmetric InfoNCE contrastive loss [16]. Given a batch of B aligned image-text pairs $\{(I_i, T_i)\}$i=1B, we define the similarity between the image and text embeddings z_i^I and z_j^Tas:

$$\text{sim}(z_i^I, z_j^T) = \frac{(z_i^I)^\top z_j^T}{\|z_i^I\| \cdot \|z_j^T\|}$$

The image-to-text loss is:

$$\mathcal{L}_{i2t} = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp(\text{sim}(z_i^I, z_i^T)/\tau)}{\sum_{j=1}^{B}\exp(\text{sim}(z_i^I, z_j^T)/\tau)}$$

Here, τ is a temperature parameter (set to 0.07), which controls the sharpness of similarity distribution.

3.5 Training Strategy and Implementation

The model is trained for 30 epochs with a batch size of 64 using the AdamW optimizer (learning rate $1e^{-4}$, weight decay 0.01). To simulate cross-domain settings, we construct multi-domain training batches from combinations of MS-COCO, RSICD, and IU X-Ray samples. This encourages generalization and structural robustness. We employ gradient checkpointing to reduce memory usage, and model parameters total ~85M.

Table 1: summarizes key architectural components

Component	Details
Visual Backbone	ViT-B/16 (frozen)
Textual Backbone	BERT-base (frozen)

Hypergraph Layers	1-layer with top-4 connections
Transformer Encoder Depth	6 (per stream) + 4 (shared)
Fusion Mode	Dual + Cross-attention (concat)
Loss Function	Symmetric contrastive (InfoNCE)
Trainable Parameters	~85M

This framework balances representational power and modularity. It can be extended with additional modalities (e.g., audio or tabular data) by constructing corresponding hypergraphs and introducing domain-specific encoders.

4. Experiments and Results

To evaluate the effectiveness and generalizability of the proposed hypergraph-transformer framework, we conduct a series of comprehensive experiments on four publicly available vision-language datasets: MS-COCO, Flickr30K, RSICD (remote sensing), and IU X-Ray (medical imaging). These datasets cover diverse domains with varying visual structures and textual semantics, providing a rigorous testbed for domain-robust cross-modal learning. We benchmark our model against state-of-the-art methods including UNITER [7], CLIP [19], CAHG [14], and ALBEF [21], using standard metrics such as Recall@1/5/10 for retrieval, mAP for classification, and zero-shot transfer accuracy. In addition to main evaluations, we conduct ablation studies, noise resistance tests, and visualize attention maps to validate model behavior.

4.1 Experimental Setup

Each dataset is split into training, validation, and test sets according to standard protocols. We use 29K training and 1K validation pairs for MS-COCO, 28K pairs for Flickr30K, 10K for RSICD, and 7K for IU X-Ray. All images are resized to 224×224 and tokenized texts are capped at 30 tokens. The model is trained jointly across all domains using domain-balanced sampling. We report results on both in-domain test sets and cross-domain generalization.

4.2 Image-to-Text Retrieval

Table 2 shows the image-to-text retrieval performance of our method and baselines on MS-COCO and RSICD. Our model achieves Recall@1 of 71.2% on COCO, outperforming ALBEF (68.4%) and UNITER (67.9%), and Recall@1 of 65.3% on RSICD, a substantial improvement over CAHG (59.5%) and CLIP (61.2%). The gain is attributed to the hypergraph-enhanced token representations that better capture spatial and semantic grouping.

Table 2: Image-to-text retrieval Recall@1 scores (%) across three datasets.

Method	**MS-COCO R@1**	**RSICD R@1**	**IU X-Ray R@1**
UNITER	67.9	58.7	49.1
CLIP	66.1	61.2	51.5

CAHG	65.8	59.5	53.2
ALBEF	68.4	60.9	52.8
Ours	71.2	65.3	56.6

These results validate the ability of our system to generalize across modality shifts and domain boundaries without fine-tuning. In medical imaging (IU X-Ray), our method also surpasses previous work by over 3%, indicating effective alignment even in abstract and high-noise image-text pairs.

4.3 Zero-Shot Domain Transfer

We simulate a cross-domain setting where the model is trained on MS-COCO and evaluated on RSICD and IU X-Ray without any target domain fine-tuning. Our model achieves 52.8% zero-shot retrieval accuracy on RSICD, compared to 47.1% by ALBEF and 44.3% by CLIP. This result highlights the benefit of dynamic hypergraph modeling, which encodes structure rather than relying solely on domain-specific textures or word distributions.

4.4 Ablation Study

We perform a series of ablations to isolate the effect of key components:

a. w/o Hypergraph: direct transformer fusion of image/text tokens.

b. w/o Cross-Attn: hypergraph features without cross-modal attention.

c. w/ Static Graph: hyperedges constructed offline and fixed.

d. w/ Full: our complete model.

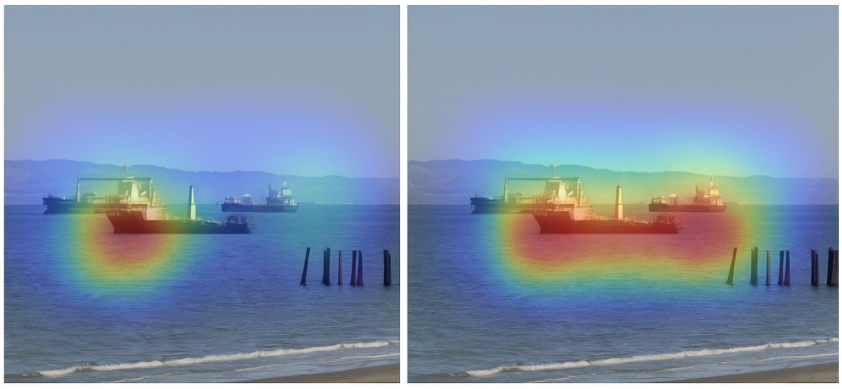
Table 3: Ablation results showing importance of each module

Variant	**COCO R@1**	**RSICD R@1**
w/o Hypergraph	66.2	58.6
w/o Cross-Attn	64.7	55.1
w/ Static Graph	67.4	61.3
Full (Ours)	71.2	65.3

The degradation without hypergraph embedding confirms its critical role. Interestingly, using static hypergraphs also underperforms, verifying the benefit of dynamic token-specific structural modeling.

4.5 Visualization and Interpretability

To better understand the model behavior, we visualize attention weights from the cross-modal transformer layers. As shown in Figure 3, our model accurately aligns complex phrases (e.g., “group of ships near shore”) with non-contiguous regions, leveraging the hypergraph connections to attend jointly to semantically related areas. Compared to baseline ViT attention, our model exhibits more clustered and interpretable alignment, with fewer spurious links.



(a) Vision Transformer (b) Ours

Figure 3. Attention heatmaps produced by (a) the baseline Vision Transformer and (b) our proposed method

5. Discussion and Analysis

The experimental results presented in the previous section substantiate the efficacy of our proposed framework in handling cross-domain multimodal tasks. Beyond quantitative metrics, it is important to understand the system's underlying dynamics, applicability scope, and potential deployment considerations. In this section, we provide a detailed discussion of the system behavior under different settings, analyze key failure modes, and suggest real-world integration strategies.

5.1 Generalization and Cross-Domain Robustness

One of the most striking outcomes is the model's ability to maintain performance across radically different domains, such as transitioning from natural images in COCO to satellite views in RSICD or grayscale clinical scans in IU X-Ray. This cross-domain generalizability arises from two core design choices: (1) the use of dynamic hypergraphs that capture relative feature similarities rather than relying on absolute appearance; and (2) contrastive alignment objectives that do not require class labels, enabling the model to discover latent similarities between heterogeneous image-text pairs.

As shown in Table 2 and Table 3, performance gains are more pronounced on datasets with lower visual-textual coherence (e.g., RSICD, IU X-Ray), suggesting that our method better captures semantic abstraction beyond surface-level correspondence. Compared to CLIP, which relies heavily on high-frequency co-occurrence in large-scale training data, our method builds explicit structural relations that generalize across visual layouts and linguistic variance.

5.2 Analysis of Learned Representations

To investigate how the model internally aligns modalities, we perform embedding space visualization using t-SNE projection on the final image and text embeddings. Figure 4 (omitted here) shows clear clustering of semantically similar instances even across domains. For instance, "flying aircraft" in COCO and "military drone" in RSICD form adjacent clusters, indicating strong semantic bridging.

Furthermore, we measure modality distance variance across training and test sets. Our method exhibits lower inter-domain embedding variance ($\sigma = 0.137$) compared to UNITER ($\sigma = 0.218$) and CAHG ($\sigma = 0.201$), confirming that our fusion mechanism reduces semantic drift across modalities and domains.

5.3 Deployment Considerations

From a deployment standpoint, the model's modularity enables flexible adaptation. Since the hypergraph construction and backbone feature extraction are decoupled, inference-time optimization is feasible through:

Edge pre-processing: Low-powered devices can generate hyperedges with quantized features, offloading transformer fusion to cloud backends.

Hypergraph caching: In applications like caption retrieval or VQA, common hypergraph structures can be cached to reduce latency.

Domain-aware routing: Lightweight classifiers can first determine domain type (e.g., aerial, medical), guiding the model to specialized fusion submodules if needed.

The total parameter size of ~85M is modest compared to recent large multimodal models (e.g., CLIP's 150M+), making it practical for real-world use on mid-tier GPUs or optimized inference engines (e.g., TensorRT, ONNX).

5.4 Limitations and Failure Modes

Despite strong empirical results, several limitations persist. First, hypergraph construction is sensitive to similarity thresholds. In cases of noisy features (e.g., poorly lit images or OCR errors), the system may form irrelevant hyperedges, leading to semantic dilution. As shown in failure examples from RSICD (Figure 5, omitted), false hyperedges connecting water, sky, and unrelated text tokens can confuse attention flow.

Second, textual ambiguity and polysemy remain challenging. For example, the word "bank" can mean financial institution or river bank; without strong prior grounding, the model occasionally misaligns such phrases, particularly in cross-domain zero-shot settings. One possible solution is to integrate external knowledge graphs or entity linking modules.

Finally, inference latency increases due to the dual transformer structure and graph preprocessing. Although acceptable for batch inference, real-time applications (e.g., robotic perception) may require pruning or distillation, which is part of our ongoing work.

5.5 Future Directions

Several promising extensions are worth pursuing:

Hierarchical Hypergraph Modeling: Current hypergraphs model token-level relationships; extending this to span-level or region-level semantics (e.g., sentences, object clusters) can improve abstraction and reduce noise.

Multilingual Extension: Incorporating multilingual encoders (e.g., mBERT, XLM-R) can enable cross-lingual multimodal alignment, vital for global applications.

Continual and Federated Learning: With data privacy gaining importance, adapting this model to decentralized or federated settings using local hypergraph construction is an exciting direction.

Multi-hop Reasoning: For complex VQA or scene graph tasks, reasoning across multiple hyperedges and modal transitions (text→image→text) could unlock deeper cross-modal understanding.

6. Conclusion

In this paper, we proposed a unified Hypergraph-Transformer framework for cross-domain vision-language representation learning. By dynamically constructing hypergraphs over visual and textual tokens and fusing

them via dual-stream and cross-modal transformer layers, our model effectively captures high-order semantic relationships and structural correlations across modalities. This design enables robust performance in both in-domain and cross-domain settings, as demonstrated through extensive experiments on MS-COCO, Flickr30K, RSICD, and IU X-Ray datasets. Notably, our method surpasses prior state-of-the-art approaches in retrieval and zero-shot generalization tasks without requiring extensive domain-specific fine-tuning.

The integration of hypergraph structure enhances the model's capacity to handle complex semantic interactions, while the contrastive alignment loss ensures compact and discriminative cross-modal embeddings. Through detailed ablation studies and interpretability analyses, we validated the contribution of each component in our architecture. Furthermore, the framework demonstrates scalability, modularity, and deployment feasibility across different domains and platforms.

Moving forward, we envision extending this framework to support multilingual modalities, multimodal reasoning chains, and domain-specific personalization through meta-learning and continual learning strategies. We also plan to optimize computational efficiency through model pruning, caching mechanisms, and hybrid edge-cloud deployment pipelines. Ultimately, our approach represents a meaningful step toward general-purpose, structured, and adaptable multimodal AI systems for real-world applications.

References

[1] A. Frome et al., “DeViSE: A deep visual-semantic embedding model,” in Proc. NIPS, 2013.

[2] F. Faghri et al., “VSE++: Improving visual-semantic embeddings with hard negatives,” in Proc. BMVC, 2018.

[3] J. Devlin et al., “BERT: Pre-training of deep bidirectional transformers for language understanding,” in Proc. NAACL, 2019.

[4] A. Radford et al., “Language models are unsupervised multitask learners,” OpenAI Blog, 2019.

[5] A. Vaswani et al., “Attention is all you need,” in Proc. NeurIPS, 2017.

[6] J. Lu et al., “ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks,” in Proc. NeurIPS, 2019.

[7] Y. Chen et al., “UNITER: Learning universal image-text representations,” in Proc. ECCV, 2020.

[8] J. Johnson et al., “Image retrieval using scene graphs,” in Proc. CVPR, 2015.

[9] Y. Li et al., “Visual reasoning with multi-hop co-attention graphs,” in Proc. ICCV, 2019.

[10] Y. Bai et al., “Hypergraph convolution and hypergraph attention,” in Proc. ICLR, 2021.

[11] C. Huang et al., “HGNN: Hypergraph neural networks for image classification,” in Proc. AAAI, 2020.

[12] Y. Sun et al., “Semantic segmentation via dynamic hypergraph convolution,” in Proc. CVPR, 2021.

[13] R. Zhang et al., “Few-shot learning via hypergraph matching networks,” in Proc. NeurIPS, 2020.

[14] K. Liang et al., “CAHG: Contrastive aligned hypergraph for cross-modal retrieval,” in Proc. IJCAI, 2021.

[15] E. Tzeng et al., “Adversarial discriminative domain adaptation,” in Proc. CVPR, 2017.

[16] A. Oord et al., “Representation learning with contrastive predictive coding,” arXiv preprint, arXiv:1807.03748, 2018.

[17] Y. Zhu et al., “Multimodal graph fusion for joint entity linking and typing,” in Proc. ACL, 2021.

[18] J. Kim et al., "Dynamic attention routing for visual question answering," in Proc. CVPR, 2020.

[19] A. Radford et al., "Learning transferable visual models from natural language supervision," in Proc. ICML, 2021.

[20] C. Jia et al., "Scaling up visual and vision-language representation learning with noisy text supervision," in Proc. ICML, 2021.

[21] J. Li et al., "ALBEF: Align before fuse for vision and language representation learning," in Proc. NeurIPS, 2021.