Multi-View Differential Mixing and Graph-Guided Structural Region Selection for Cross-Modal Alignment

Linlin Ji[1], **Li Liu**[1,2*]

[1]School of Information Science and Engineering, Shandong Normal University, Jinan, China
[2]Shandong Province Key Laboratory of Independent and Reliable Computing Technology and Equipment, Jinan
17616243627@163.com, liuli_790209@163.com

Abstract

Cross-modal alignment is a promising yet challenging task in multimodal learning. Existing methods typically assess it by measuring the cross-modal semantic similarity from both global and local perspectives. However, these methods often neglect their potential interdependence. Specifically, global matching methods suffer from the over-compression of local features, while local matching methods rarely consider the inherent spatial topology of image patches. To address these limitations, we propose MG-Net, a unified framework with two collaborative modules: Multi-View Differential Mixer (MDM) and Graph-Guided Structural Region Selector (GSRS). The MDM is designed to capture discriminative global representations. It generates a series of views by decomposing feature vectors through multi-order differential operations, and adaptively fuses them via a lightweight Mixture-of-Experts (MoE) network. Meanwhile, the GSRS organizes image patches as a spatial graph and employs text-guided contextual reasoning to select spatially coherent and semantically complete structural regions. Extensive experiments on the Flickr30K and MS-COCO benchmarks demonstrate that the proposed MG-Net outperforms state-of-the-art methods in most cases.

Code — https://github.com/tiantian176/MG-Net.

Introduction

Cross-modal alignment is dedicated to bridging the semantic gap between different modalities, such as visual and textual modalities. (Zhang et al. 2022; Li et al. 2019; Zhang and Lu 2018). As a key technology in the field of cross-modal understanding and generation, cross-modal alignment plays a pivotal role in various downstream tasks, such as cross-modal retrieval (Jing et al. 2020), visual grounding (Xu et al. 2009), and visual question answering (Guo et al. 2024).

Existing cross-modal alignment methods can be roughly grouped into two categories: global alignment methods and local alignment methods. The former aims to map the entire content from different modalities into a shared embedding space with independent encoders, scoring cross-modal semantic similarity by comparing the global embedding vectors (Pan, Wu, and Zhang 2023; Huang et al. 2024; Diao et al. 2021; Ji and Liu 2025). Local alignment methods focus on learning fine-grained correspondences between local units of different modalities (e.g., visual regions in an image and word fragments in a text), thereby enabling the capture of more complex and subtle semantic associations (Fu et al. 2024; Liu et al. 2025, 2022; Zhu et al. 2022; Xu et al. 2018).

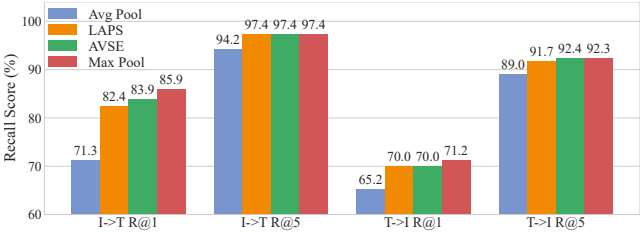


Figure 1: The Impact of Global Aggregation Strategy on Model Performance.

To improve the performance of cross-modal alignment, current methods tend to score the global- and local-level alignments, and then combine them to measure the semantic similarity between different modalities. However, existing methods mostly learn the global representation by aggregating the local features extracted by pre-trained encoders, like BERT (Devlin et al. 2019) and CLIP (Radford et al. 2021). But, we argue that these methods exaggerate the designs of aggregation. As shown in Fig. 1, we evaluate average pooling, max pooling and two alignment methods (LAPS (Fu et al. 2024) and AVSE (Liu et al. 2025)), on the image-text and text-image retrieval tasks, and list their performance w.r.t Recall@1 and Recall@5. From the observation, different simple pooling strategies exhibit vast performance disparities: the widely used average pooling leads to suboptimal performance when mapping image patches and text words to a shared space. In contrast, max pooling achieves surprisingly strong performance, with results even surpassing some more complex state-of-the-art methods in two metrics (Fu et al. 2024; Liu et al. 2025). The reason might be that the most discriminative signal in each modality conveys the representative features. Nevertheless, such a simple strategy is hard to apply in most scenarios, since the rich contextual information from all other local regions could be discarded. Therefore, how to measure the representativeness of each

*Corresponding author.

local feature in global representation learning and aggregate them is the first challenge we are facing.

For local representation learning, existing methods typically segment images into fixed-size grids of non-overlapping patches (Fu et al. 2024; Liu et al. 2025). This inherently disregards the image's spatial structure and semantic integrity (Li et al. 2017), treating each patch as an independent entity that is merely a disconnected fragment of a visual object. When performing alignment by selecting these patches in isolation without considering their spatial topological relationships, this often leads to selected visual regions that are spatially scattered and semantically incoherent, as shown in Fig. 2, thereby harming the representation of local features. Hence, how to utilize the global spatial structure information to guide the local representation learning is the other challenge.

To overcome the aforementioned challenges, we propose a novel cross-modal alignment framework equipped with two collaborative key modules: a Multi-View Differential Mixer (MDM) and a Graph-Guided Structural Region Selector (GSRS). To address the limitations of global representations, MDM generates a series of views that reveal the internal dynamics and structure of features by computing multi-order differences on image patches or text features. Subsequently, a lightweight Mixture-of-Experts (MoE) network adaptively fuses these diverse views, thereby enriching the hierarchy and discriminative power of the global representation while preserving key information.

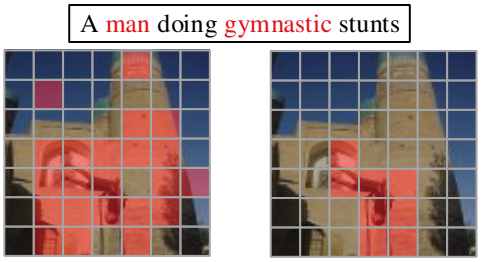
For local representation learning, the GSRS module constructs a spatial topological graph from all image patches to explicitly model their adjacency relationships and spatial dependencies (Veličković et al. 2017). Based on this graph, the model performs contextual reasoning to adaptively identify and select structural visual regions that is most relevant to the text description and spatially coherent. It is able to enhance the semantic integrity of the selected visual evidence. The contributions of this paper are summarized as follows:

- After probing the prior studies, we explore the influence of global and local representation learning on each other.
- We propose MG-Net, a unified framework with two collaborative modules, including MDM for generating diverse global views via multi-order differences to enrich global representation, and GSRS for graph-based spatial reasoning to ensure coherent and semantically relevant local feature selection.
- We perform extensive experiments on two public datasets to demonstrate that our proposed model outperforms several state-of-the-art methods in most cases.

Related Work

Global Alignment Methods

Global alignment methods primarily map image and text features into a common embedding space, where similarity is computed via cosine similarity of their embeddings. VSE++ (Faghri et al. 2017) improves visual-semantic embeddings by employing online hard negative mining to pull matched positive pairs closer while pushing negative pairs apart. GPO (Chen et al. 2021) introduces a Generalized Pooling Operator that adaptively learns the optimal pooling function for different modalities. Similarly, HREM (Fu et al. 2023) aligns representations by utilizing both global average and max pooling. However, these methods typically rely on a single aggregation strategy, whereas our work fuses global feature vectors from multiple perspectives.



(a) Exist method (b) Our method

Figure 2: The effect of global structure information on local spatial relationships.

Local Alignment Methods

To enhance cross-modal alignment precision, many works perform fine-grained interaction between local features before deriving a final similarity score (Zhou et al. 2020; Zhang et al. 2024; Jiang and Ye 2023). The seminal SCAN (Lee et al. 2018) uses stacked cross-attention for mutual image-text region alignment. Subsequent works like IMRAM (Chen et al. 2020) and CAAN (Zhang et al. 2020) introduce more sophisticated mechanisms to further improve this fine-grained alignment. However, these fine-grained methods suffer from high computational complexity and numerous invalid alignments. To mitigate this, LAPS (Fu et al. 2024) introduces a pruning framework that selects important patches based on saliency. Nevertheless, these methods still typically neglect the spatial coherence of visual patches, often selecting a scattered collection of patches that are semantically incoherent as a whole.

Methods

In this section, we elaborate on the design of our proposed MG-Net, as illustrated in Fig. 3. In particular, we detail the Token Feature Extraction, followed by the MDM and GSRS.

Token Feature Extraction

We begin by employing a pre-trained Transformer architecture (Vaswani et al. 2017) as the feature encoder for both image and text inputs, extracting sequences of visual and textual tokens, respectively.

Visual Patches. Given an image I, we employ a Vision Transformer (ViT) (Dosovitskiy et al. 2020) or Swin Transformer (Liu et al. 2021) as the visual encoder. The image is partitioned into N non-overlapping patches, based on its spatial grid. Then, these patches are treated as a sequence of visual tokens and fed into the visual encoder. After that, we obtain a set of visual patch features $V = \{v_1, \ldots, v_N\} \in \mathbb{R}^{N \times d}$, where d is the feature dimension.

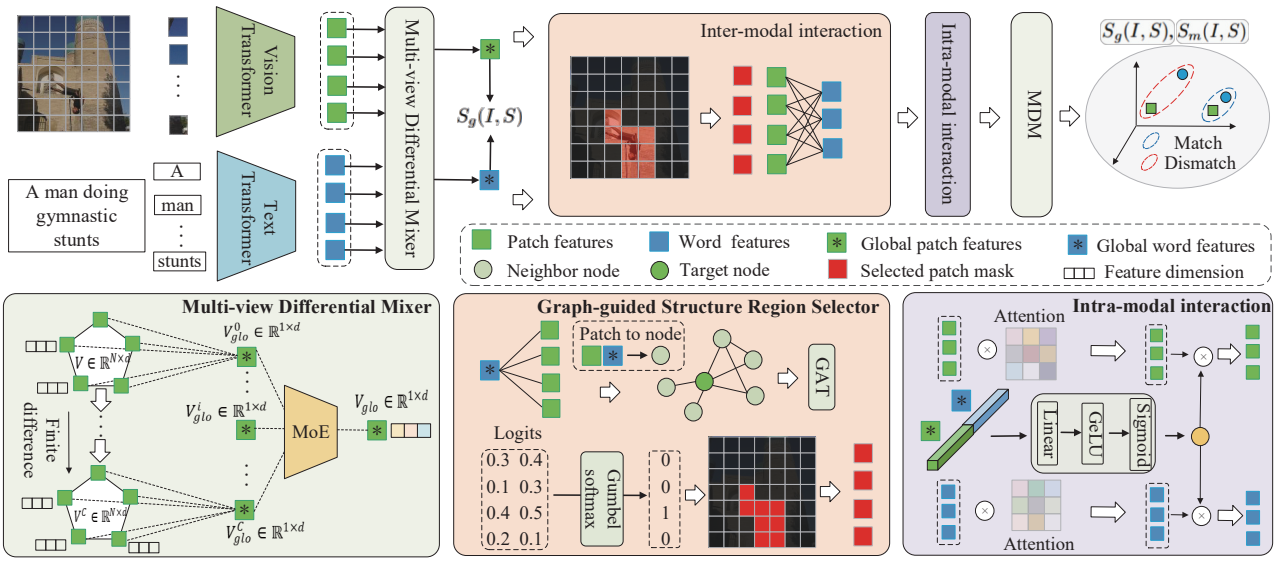


Figure 3: The architecture of our MG-Net framework.

Text Tokens. For a given sentence S, we utilize BERT (Devlin et al. 2019) as the text encoder. Analogously, the sentence is tokenized into a sequence of linguistic words and then fed into the encoder. This allows us to obtain a set of textual word features $T = \{t_1, \ldots, t_M\} \in \mathbb{R}^{M \times d}$, where M is the number of words in the sentence.

Multi-View Differential Mixer

With the obtained visual and textual features, existing global matching methods typically adopt the simple global pooling to combine them into visual and textual global representations, respectively. However, such an operation may cause the dimensionality reduction and information compression, leading to the loss of critical information.

To address this problem, we propose a Multi-View Differential Mixer that consists of a base global representation and multiple differential global representations. It is able to characterize image and text features with richer information, as illustrated in Fig. 3. Toward this end, we construct the base global representation by aggregating the information from the most discriminative patches. In particular, we select K patch feature vectors with the largest L1-norms, and then aggregate them via a norm-weighted average to highlight the most informative regions in the image. The base global representation tensor is defined as follows:

$$V_{glo}^0 = \frac{\sum_{j=1}^{K} \|v_j\|_1 \cdot v_j}{\sum_{j=1}^{K} \|v_j\|_1}, \tag{1}$$

where v_j represents the features of j-th patch and $V_{glo}^0 \in \mathbb{R}^{1\times d}$ is the base global representation. In addition, $\|\cdot\|_1$ denotes the L1-norm. After that, we compute the differential global features, in order to capture the representative features. We follow the fact that the representativeness of a patch feature is directly proportional to the magnitude of variation between adjacent data within that patch. Since differencing is an effective tool for measuring the degree of such changes, we apply it to calculate the differential global features. Formally, the first-order difference (Δv_i) is defined as the difference between adjacent feature elements:

$$(\Delta^1 v_i)_j = v_{i,j+1} - v_{i,j}, \tag{2}$$

where $j \in \{1, \ldots, d-1\}$ is the feature dimension index. The c-th order differential decomposition is denoted as:

$$(\Delta^c v_i)_j = (\Delta^{c-1} v_i)_{j+1} - (\Delta^{c-1} v_i)_j. \tag{3}$$

Since the magnitude of the c-th order difference vector serves as an effective indicator of a patch's feature variation intensity under c-th order abstraction, we define its L2-norm as a Differential Saliency Score:

$$s_{i,c} = \|\Delta^c v_i\|_2, \tag{4}$$

where $s_{i,c}$ denotes the score of the i-th patch with c-th order differential decomposition. Based on this saliency score, we then select the indices of the top-K most salient patches for each differential order c, denoted as $\mathcal{I}_c$:

$$\mathcal{I}_c = \underset{i\in\{1,\ldots,N\}}{\text{arg top-}K}(s_{i,c}). \tag{5}$$

Using the feature indices $\{\mathcal{I}_1, \ldots, \mathcal{I}_C\}$ obtained from the differential energy, we retrieve the corresponding patch subsets $\{V^1, \ldots, V^C\}$ from the original set V, where each $V^c = \{v_1^c, \ldots, v_K^c\}$. These subsets are then aggregated to form higher-order views:

$$V_{glo}^c = \frac{\sum_{j=1}^{K} \|v_j^c\|_1 \cdot v_j^c}{\sum_{j=1}^{K} \|v_j^c\|_1}, \tag{6}$$

where $V_{glo}^c \in \mathbb{R}^{1\times d}$ represents the global visual feature for the c-th order. This process yields a set of views $\{V_{glo}^0, V_{glo}^1, \ldots, V_{glo}^C\}$.

Mixture-of-Experts. To effectively integrate the outputs of our multi-order differential analysis, we formulate the final fusion step as a lightweight Mixture-of-Experts (MoE) layer. This layer dynamically predicts the contribution of each view by generating a probability distribution over them. Let $\{V_{glo}^c\}_{c=0}^C$ be the set of all view representations. The final fused visual feature $V_{glo} \in \mathbb{R}^{1\times d}$ can be formulated as:

$$V_{glo} = \sum \text{Softmax}(H_1(\{v_{glo}^i\}_{i=0}^C)) \cdot (\{v_{glo}^i\}_{i=0}^C), \quad (7)$$

where H_1 denotes a multilayer perceptron (MLP) network.

For the text modality, we conduct the same operations on the word features T. To handle variable sentence lengths, we select the top-$\min(K, L)$ feature vectors, where L is the actual number of words in the sentence. Then, we obtain the global textual representation tensor $T_{glo} \in \mathbb{R}^{1\times d}$.

With the final global visual feature V_{glo} and global textual feature T_{glo}, we compute the cosine similarity between the image I and sentence S as:

$$S_g(I, S) = \frac{V_{glo} \cdot T_{glo}}{||V_{glo}|| \cdot ||T_{glo}||}. \quad (8)$$

Graph-Guided Structural Region Selector

To select structured regions that align with textual semantics, it is crucial to learn the text-awareness visual features for the selection. To this end, we first enhance node features under textual guidance. Specifically, we use the global text representation $T_{glo} \in \mathbb{R}^{1\times d}$ to generate a scaling factor $\gamma_i \in \mathbb{R}^{1\times d}$ and a bias term $\beta_i \in \mathbb{R}^{1\times d}$ for each visual patch in the image, where $i \in \{1, \ldots, N\}$ indexes the i-th patch. It can be formulated as:

$$[\gamma_1, \ldots, \gamma_N, \beta_1, \ldots, \beta_N] = H_2(T_{glo}), \quad (9)$$

where H_2 represents an MLP. Subsequently, for the original visual patch feature $v_i \in \mathbb{R}^{1\times d}$, we utilize the corresponding text-generated parameters γ_i and β_i to perform an element-wise affine transformation. The result is then concatenated with the text feature to obtain the text-guided visual patch feature $v_i' \in \mathbb{R}^{1\times 2d}$:

$$v_i' = \text{Concat}(\gamma_i \odot v_i + \beta_i, T_{glo}), \quad (10)$$

where $\odot$ and Concat$(\cdot)$ denote the element-wise multiplication and concatenation operations, respectively. This yields the final set of visual features $V' = \{v_1', \ldots, v_N'\}$.

We treat each patch v_i' as a node in a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. To capture the local 2D topology of the image, we build a spatial adjacency graph. Its edges $\mathcal{E}$ connect nodes that are spatially adjacent in the original image grid. This is defined by an adjacency matrix $\mathcal{A} \in \{0, 1\}^{N\times N}$:

$$\mathcal{A}_{ij} = \begin{cases} 1 & \text{if } \max(|r_i - r_j|, |c_i - c_j|) = 1, \\ 0 & \text{otherwise.} \end{cases} \quad (11)$$

We conduct a Graph Attention Network (GAT) to propagate and aggregate contextual information. For a node v_i', a similarity coefficient e_{ij} is computed for each neighbor $v_j \in \mathcal{N}(v_i) \cup \{v_i\}$ (including itself):

$$e_{ij} = H_3\left(Concat(Wv_i', Wv_j')\right), \quad (12)$$

where W is a learnable linear transformation matrix shared across all nodes for extracting higher-level features, and H_3 represents an MLP. The final attention weights α_{ij} are then obtained by normalizing these coefficients across all neighbors of node v_i:

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{v_k \in \mathcal{N}(v_i)\cup\{v_i\}} \exp(e_{ik})}, \quad (13)$$

where α_{ij} represent attention weights quantifying the contributions from the neighboring nodes v_j' to the central node v_i'. Finally, we use these weights to perform a weighted aggregation of the features of all neighboring nodes to obtain the updated feature representation h_i':

$$h_i^{'} = \sigma \sum_{v_j \in \mathcal{N}(v_i)\cup\{v_i\}} \alpha_{ij} W v_j'. \quad (14)$$

After that, each patch feature effectively aggregates the most text-relevant contextual information from its spatial neighborhood. It is able to transform the semantic information from an isolated feature vector into a context-rich structural representation h_i'.

To maintain gradient flow during the discrete selection process, we employ the Gumbel-Softmax trick (Maddison, Mnih, and Teh 2016), which is a differentiable relaxation of the discrete categorical sampling process. For each node's logit l_i (derived from h_i' via an MLP), its corresponding selection probability vector m_i is calculated as:

$$m_i = \text{Softmax}\left(\frac{l_i + g_i}{\tau}\right), \quad (15)$$

where g_i is a noise term independently sampled from a standard Gumbel(0, 1) distribution, and τ is a temperature coefficient that controls the smoothness of the distribution. The first element of the vector m_i represents the probability of the i-th patch being selected. The selection probabilities of all patches constitute the mask $M = \{m_i\}_{i=1}^N$.

Inter- and Intra-modal Interaction

After obtaining the mask matrix, we can filter the original visual features: $V' = V \cdot M$, where $V' \in \mathbb{R}^{N\times d}$ denoteds the masked visual features. Subsequently, we compute the visual-text cross-attention:

$$V_{cross} = \text{Cross-Attention}(T, V', V'), \quad (16)$$

where $V_{cross} \in \mathbb{R}^{M\times d}$ represents the cross-modal visual features enhanced by the semantic information from the text.

Next, to evaluate the importance of each local feature within its own modality, we introduce an intra-modal contextual alignment mechanism. It computes the semantic consistency between each local feature and the global context of that modality, as shown in Fig. 3.

For the visual modality, we first construct a vector $\bar{v}_{glo}$ representing the global visual context by averaging all M patch features from the cross-attention output, $V_{cross} = \{v_i\}_{i=1}^M$, and then applying L2 normalization:

$$\bar{v}_{glo} = \frac{\frac{1}{M}\sum_{i=1}^M v_i}{\left\|\frac{1}{M}\sum_{i=1}^M v_i\right\|_2}. \quad (17)$$

Method	MS-COCO 1K							MS-COCO 5K						
	Image-to-Text			Text-to-Image			rSum	Image-to-Text			Text-to-Image			rSum
	R@1	R@5	R@10	R@1	R@5	R@10		R@1	R@5	R@10	R@1	R@5	R@10	
ViT-Base-224 + BERT–base, 14×14 patches														
VSE++ (Faghri et al. 2017)	75.0	94.6	98.0	62.7	89.4	94.9	514.6	52.4	80.3	88.8	40.6	70.4	81.1	413.4
SCAN (Lee et al. 2018)	76.0	95.4	98.1	64.5	90.8	95.8	520.6	53.9	81.8	90.0	42.9	72.3	82.5	423.5
SGR (Diao et al. 2021)	77.2	95.0	98.0	65.1	90.7	95.8	521.8	54.9	82.8	90.5	42.8	72.2	82.5	425.8
CHAN (Pan, Wu, and Zhang 2023)	77.1	95.1	98.1	65.0	91.0	96.0	522.2	56.3	83.2	90.1	43.0	72.6	82.8	428.0
LAPS (Fu et al. 2024)	78.7	95.5	98.3	66.2	91.3	96.2	526.3	57.5	84.0	90.8	44.5	74.0	83.6	434.4
PICO (Ma et al. 2025)	78.8	95.9	**98.8**	66.3	**91.6**	**96.5**	527.9	57.5	84.1	91.2	44.9	**74.3**	**83.8**	435.8
Ours	**80.5**	**96.1**	98.4	**66.7**	90.8	95.6	**528.1**	**60.1**	**85.5**	**91.5**	**45.5**	73.9	83.1	**439.7**
Swin-Base-224 + BERT–base, 7×7 patches														
VSE++ (Faghri et al. 2017)	83.3	97.5	99.3	71.0	93.0	96.7	540.9	64.0	88.2	94.2	49.9	78.0	86.6	460.9
SCAN (Lee et al. 2018)	80.9	97.0	99.1	69.7	93.1	97.1	536.9	60.7	86.6	93.2	48.1	77.1	86.1	451.8
SGR (Diao et al. 2021)	81.2	97.1	99.1	69.9	93.2	97.2	537.7	61.0	86.7	93.2	48.6	77.2	86.3	453.1
CHAN (Pan, Wu, and Zhang 2023)	81.6	97.2	99.3	70.6	93.7	**97.6**	539.8	64.1	87.9	93.5	49.1	77.3	86.1	458.0
LAPS (Fu et al. 2024)	84.0	97.6	99.3	72.1	93.7	97.3	544.1	64.5	89.2	94.4	51.6	78.9	87.2	465.8
PICO (Ma et al. 2025)	84.2	97.9	99.5	72.1	**93.8**	97.4	544.9	64.6	89.7	94.8	51.7	**79.2**	**87.5**	467.5
Ours	**84.4**	**97.9**	**99.5**	**73.1**	93.3	97.2	**545.4**	**67.9**	**89.8**	**94.9**	**51.8**	78.7	87.0	**470.1**
Swin-Base-384 + BERT–base, 12×12 patches														
VSE++ (Faghri et al. 2017)	82.9	97.7	99.4	71.3	93.5	97.3	542.1	63.0	88.5	94.3	50.1	78.9	87.4	462.2
SCAN (Lee et al. 2018)	81.6	96.8	99.1	69.1	92.7	96.7	536.1	61.1	87.3	93.3	47.8	76.9	85.9	452.4
SGR (Diao et al. 2021)	81.9	96.7	99.1	69.3	92.8	96.7	536.6	62.8	87.0	92.9	48.1	77.0	86.0	453.8
CHAN (Pan, Wu, and Zhang 2023)	83.1	97.3	99.2	70.4	93.1	97.1	540.2	63.4	88.4	94.1	49.2	77.9	86.6	459.5
LAPS (Fu et al. 2024)	84.1	97.4	99.2	72.1	93.9	97.4	544.1	67.1	88.6	94.3	53.0	79.5	87.6	470.1
PICO (Ma et al. 2025)	84.4	97.8	99.5	72.5	**94.3**	**97.9**	546.4	**67.4**	89.0	94.5	53.1	79.8	88.0	471.8
Ours	**84.6**	**98.6**	**99.5**	**73.3**	94.0	97.5	**546.8**	66.3	**89.5**	**95.1**	**53.1**	**80.2**	**88.1**	**472.3**

Table 1: Comparison with state-of-the-art methods on MS-COCO datasets. The best results are marked in bold.

We assume that greater alignment between a local feature and the context vector indicates higher semantic importance within the modality. Therefore, we define the saliency score map S_V of local features as,

$$S_V = \bar{v}_{glo} \cdot V_{cross}, \tag{18}$$

where $S_V = (S_{v_1}, S_{v_2}, \ldots, S_{v_M}) \in \mathbb{R}^{M \times d}$, and S_{v_i} represents the saliency score of each local feature. We perform the same operations on the text modality to compute its global context representation $\bar{t}_{glo}$ and its textual salience map $S_T \in \mathbb{R}^{M \times d}$.

Next, to produce a gating signal for modulating these salience scores, we concatenate the input global image feature $\bar{v}_{glo} \in \mathbb{R}^{1 \times d}$ and the global text feature $\bar{t}_{glo} \in \mathbb{R}^{1 \times d}$ along the feature dimension, forming a joint global feature $\bar{f}_{glo} \in \mathbb{R}^{1 \times 2d}$. This joint feature is passed through an MLP, $H_4(\cdot)$, and a Sigmoid activation function σ to compute a gating vector g:

$$g = \sigma(H_4(\bar{f}_{glo})). \tag{19}$$

The modulated salience scores are then computed as:

$$S'_V = g \cdot S_V, \quad S'_T = (1 - g) \cdot S_T. \tag{20}$$

Finally, the modulated scores S'_V and S'_T are processed by the MDM to obtain the final global representations V_m and T_m. The final cosine similarity $S_m(I, S)$ is computed as:

$$S_m(I, S) = \frac{V_m \cdot T_m}{||V_m|| \cdot ||T_m||}. \tag{21}$$

Objective Function

To train the proposed model, we use a bidirectional triplet loss with hard negative mining (Faghri et al. 2017).

$$\begin{aligned} \mathcal{L}_{\text{triplet}} = \sum_{(I,S)} \Big(& [\alpha - S(I,S) + S(I,\hat{S})]_+ \\ & + [\alpha - S(I,S) + S(\hat{I},S)]_+ \Big), \end{aligned} \tag{22}$$

where α is the margin parameter, and $\hat{I}$ and $\hat{S}$ represent the hardest negative samples within the current training mini-batch that have the highest similarity score. We apply this triplet loss to both the similarity score from our MDM module, $S_g(I, S)$, and the final similarity score, $S_m(I, S)$, to obtain $\mathcal{L}_g$ and $\mathcal{L}_m$, respectively. The final loss is defined as:

$$\mathcal{L} = \mathcal{L}_g + \lambda \mathcal{L}_m, \tag{23}$$

where λ is a hyperparameter that balances the two components of the loss function.

Experiments

Datasets and Evaluation Metrics

Following prior works (Faghri et al. 2017; Lee et al. 2018), we train and evaluate our model on two standard benchmarks: Flickr30K (Young et al. 2014) and MS-COCO (Lin et al. 2014). In these datasets, each image is associated with

Method	Image-to-Text			Text-to-Image			rSum
	R@1	R@5	R@10	R@1	R@5	R@10	
ViT-Base-224 + BERT–base, 14×14 patches							
VSE++	71.8	92.8	96.5	59.4	84.7	90.9	496.1
SCAN	69.5	90.9	95.6	56.4	83.1	90.0	485.6
SGR	69.7	90.8	95.2	59.1	84.1	89.9	488.7
CHAN	69.2	91.8	95.0	58.4	84.9	90.6	489.9
LAPS	74.0	93.4	97.4	62.5	87.3	92.7	507.3
PICO	74.5	94.0	**98.2**	63.0	**88.5**	**93.1**	511.3
Ours	**78.9**	**94.8**	97.7	**63.4**	87.5	92.7	**515.0**
Swin-Base-224 + BERT–base, 7×7 patches							
VSE++	82.5	96.5	98.9	70.0	91.4	95.1	534.4
SCAN	79.0	95.9	98.2	67.7	90.6	94.9	526.3
SGR	80.4	97.0	98.7	66.9	90.2	94.5	527.6
CHAN	81.4	97.0	98.6	68.5	90.6	94.5	530.6
LAPS	82.4	97.4	99.5	70.0	91.7	95.4	536.3
PICO	82.9	97.9	99.6	70.3	92.2	95.6	538.5
Ours	**87.8**	**98.6**	**99.7**	**72.8**	**92.6**	**95.7**	**547.2**
Swin-Base-384 + BERT–base, 12×12 patches							
VSE++	83.3	97.5	99.2	71.1	93.2	96.2	540.6
SCAN	81.9	96.9	98.9	70.0	92.7	95.8	536.1
SGR	80.7	96.8	99.0	69.9	91.7	95.3	533.4
CHAN	81.2	96.7	98.8	70.3	92.2	95.9	535.0
LAPS	85.1	97.7	99.2	74.0	93.0	96.3	545.3
PICO	85.8	98.1	99.4	74.5	93.5	96.9	548.2
Ours	**89.1**	**99.5**	**100.0**	**76.4**	**94.4**	**97.0**	**556.4**

Table 2: Comparison with state-of-the-art methods on the Flickr30K datasets. The best results are marked in bold.

five corresponding text captions. The evaluation metrics are Recall at K (R@K, for K=1, 5, 10), which measures the percentage of ground-truth items retrieved in the top-K list, and rSum, which is the sum of all six R@K scores for both image-to-text and text-to-image retrieval.

Implementation Details

To ensure a fair comparison, we keep our vision-text backbones consistent with existing cross-modal alignment methods (Ma et al. 2025; Fu et al. 2024). For the visual encoder, we use either a ViT with a patch size of 16×16 pixels, or a Swin with a patch size of 32×32 pixels. For the text encoder, we use BERT (Devlin et al. 2019). All encoders used are the base versions and finetuned end-to-end with our heads.

Images are resized to a resolution of either 224×224 or 384×384, resulting in a sequence of 14×14 patches for ViT, and 7×7 or 12×12 patches for Swin, respectively. And, we use a linear layer on top of the encoders to unify the feature dimension to $d = 512$. The model is trained for 30 epochs using the AdamW optimizer (Loshchilov and Hutter 2017).

Comparison with State-of-the-Art

To demonstrate the superiority of our proposed method, we compare its performance against recent state-of-the-art methods on the Flickr30K and MS-COCO datasets.

The experimental results are summarized in Table 1 and Table 2. Compared to the recent state-of-the-art methods (Ma et al. 2025; Fu et al. 2024), our approach demonstrates superior performance across nearly all metrics. On Flickr30K with Swin-384, MG-Net attains an rSum of 556.4 (+8.2 over PICO's 548.2) and a saturated 100% R@10 (image-to-text), indicating highly reliable matching quality.

Method		Image-to-Text			Text-to-Image			rSum
		R@1	R@5	R@10	R@1	R@5	R@10	
MDM	Avg	71.3	94.2	97.3	65.2	89.0	93.3	510.2
	Max	85.9	97.4	99.4	71.2	92.3	95.4	541.7
	C=0	84.3	97.7	99.4	71.2	92.5	95.6	540.8
	C=1	85.7	97.9	99.1	71.4	91.8	95.6	541.5
	C=2	85.0	98.4	99.5	72.7	92.4	96.0	544.0
	C=3	86.1	97.9	99.1	72.3	92.4	95.6	543.5
GSRS	w/o G	85.0	98.4	99.5	72.7	92.4	**96.0**	544.0
	w/o T	86.2	98.6	99.5	72.7	92.4	95.7	545.1
	Full	**87.8**	**98.6**	**99.7**	**72.8**	**92.6**	95.7	**547.2**

Table 3: Ablation study comparing different modules of our method on the Flickr30K dataset. "Avg" denotes average pooling, "Max" max pooling; "G" the GSRS Module, and "T" the Text-Guided component.

Ablation Study

To verify the effectiveness of our proposed modules, we conduct a series of detailed ablation studies on the Flickr30K dataset. We evaluate the individual contributions of the MDM and the GSRS module, and their synergistic effect when combined. The results are listed in Table 3.

We evaluate the effectiveness of the MDM module by comparing it against two widely used baseline aggregation strategies: average pooling and max pooling. We analyze the effect of different differential orders (C) on the MDM module. Using only the 0-th order view (C=0), i.e., our base global representation, the performance (rSum 540.8) is already comparable to that of max pooling. However, as the number of differential orders increases, thereby introducing higher-order views extracted from the internal dynamics of the features, the model's performance continuously improves. When C=2, the model achieves an rSum of 544.0, demonstrating that our proposed multi-view decomposition and fusion strategy can effectively capture rich information beyond a single maximum signal, thus enhancing the discriminative power of the global representation.

Qualitative Analysis

To intuitively verify the effectiveness of our proposed GSRS module in selecting visual regions, we visualize the structured regions it generates, as shown in Fig.4. From the visualization results, it is evident that, unlike prior methods that may select scattered and isolated image patches, our GSRS module successfully selects visual regions that are semantically complete and spatially coherent. For instance, when the text describes a specific object (e.g., "young boy mow a of flowers a toy lawn mower."), the attention of the GSRS module is precisely focused on the complete contour of that

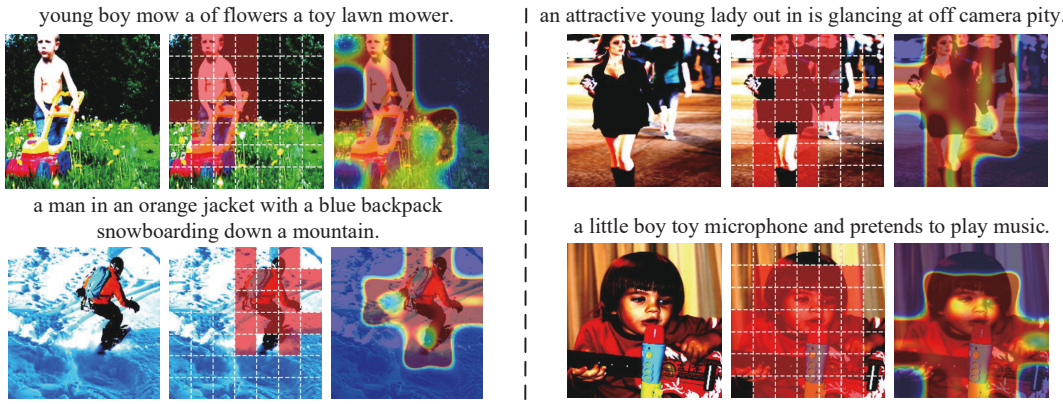


Figure 4: Validation of the GSRS module's effectiveness in selecting spatially coherent visual regions.

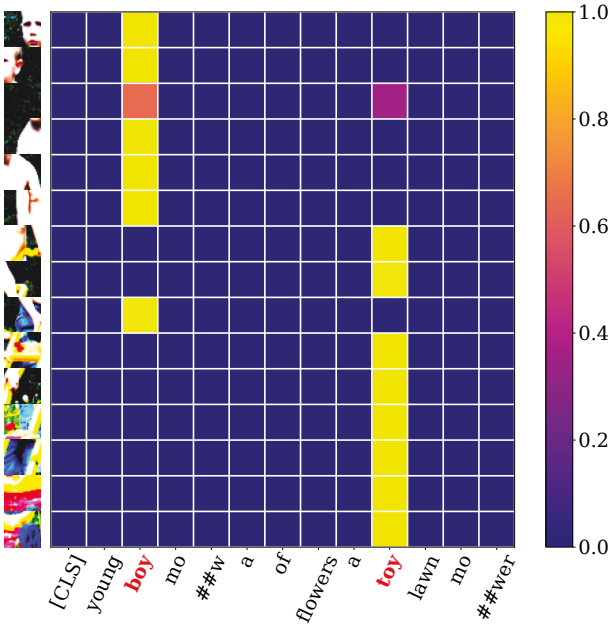


Figure 5: Fine-grained Sparse Alignment Matrix. The horizontal axis represents the individual words in the input text, while the vertical axis represents the visual image patches selected by our module.

object in the image, rather than being scattered across multiple irrelevant, fragmented patches. This provides strong evidence for the effectiveness of our graph-guided mechanism: by explicitly modeling the spatial topological relationships between image patches, the model is able to perform contextual reasoning to identify and integrate adjacent regions that collectively form a complete semantic concept.

Model Interpretability

To further investigate the fine-grained correspondence between our selected visual regions and the words in the text description, we visualize the "selected patch-word" alignment matrix. As shown in Fig.5, it is clear from the visualization that the alignment scores are highly sparse, concentrating significantly on a few key semantic columns. The vast majority of the selected image patches exhibit extremely high alignment scores with the columns for 'boy' and 'toy', while their alignment scores with background or functional words (such as 'a', 'with') are generally close to zero. This not only validates that our selected visual regions are highly relevant but also reveals the model's ability to deconstruct a complex visual scene into different semantic parts and establish clear connections with specific words in the text. This, in turn, provides strong support and high interpretability for the model's final matching decision.

Computational Complexity Analysis

Computing up to the C-th order differences over N patches of dimension d in MDM yields $O(CNd)$ complexity, comparable to global pooling $O(Nd)$. Since C is small ($C \leq 3$), this introduces negligible overhead while enriching feature hierarchy through lightweight differencing and expert fusion. GSRS builds a sparse 8-neighborhood grid with $|E|{=}O(N)$ and applies a single-head GAT, limiting message passing to local neighbors instead of dense $O(N^2)$ connections. Thus, the model keeps linear complexity in N and scales well to higher resolutions. In practice, MG-Net runs inference in 7.01 s on Flickr30K (Swin-224), over $2\times$ faster than LAPS (17.37 s) and faster than VSE++ (9.78 s).

Conclusion

In this paper, we tackle two key challenges in image-text matching—the information bottleneck in global representations and the lack of spatial structure in local matching—by proposing a novel framework, MG-Net. The framework features an innovative Multi-View Differential Mixer, which produces more discriminative global representations via differential decomposition and dynamic expert fusion. Meanwhile, a Graph-Guided Structural Region Selector elevates visual grounding for matching from isolated patches to semantically coherent structural regions by modeling the spatial topology of image patches. Extensive experiments on two widely-used datasets demonstrate that MG-Net surpasses current state-of-the-art methods in most cases.

Acknowledgments

This work is supported by the fundamental research project of Shandong, China (Nos. ZR2024ZD08, ZR2024MF043)

References

Chen, H.; Ding, G.; Liu, X.; Lin, Z.; Liu, J.; and Han, J. 2020. Imram: Iterative matching with recurrent attention memory for cross-modal image-text retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12655–12663.

Chen, J.; Hu, H.; Wu, H.; Jiang, Y.; and Wang, C. 2021. Learning the best pooling strategy for visual semantic embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15789–15798.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, volume 1 (long and short papers)*, 4171–4186.

Diao, H.; Zhang, Y.; Ma, L.; and Lu, H. 2021. Similarity reasoning and filtration for image-text matching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 1218–1226.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Faghri, F.; Fleet, D. J.; Kiros, J. R.; and Fidler, S. 2017. Vse++: Improving visual-semantic embeddings with hard negatives. *arXiv preprint arXiv:1707.05612*.

Fu, Z.; Mao, Z.; Song, Y.; and Zhang, Y. 2023. Learning semantic relationship among instances for image-text matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15159–15168.

Fu, Z.; Zhang, L.; Xia, H.; and Mao, Z. 2024. Linguistic-aware patch slimming framework for fine-grained cross-modal alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 26307–26316.

Guo, K.; Tian, D.; Hu, Y.; Lin, C.; Qian, Z.; Sun, Y.; Zhou, J.; Duan, X.; Gao, J.; and Yin, B. 2024. Cfmmc-align: Coarse-fine multi-modal contrastive alignment network for traffic event video question answering. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(11): 10538–10550.

Huang, S.; Fu, W.; Zhang, Z.; and Liu, S. 2024. Global-local fusion based on adversarial sample generation for image-text matching. *Information Fusion*, 103: 102084.

Ji, L.; and Liu, L. 2025. Multi-Scale Feature Fusion Based on Piecewise Polynomial Activation Function for Image-Text matching. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(11): 11627–11640.

Jiang, D.; and Ye, M. 2023. Cross-modal implicit relation reasoning and aligning for text-to-image person retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2787–2797.

Jing, Y.; Wang, W.; Wang, L.; and Tan, T. 2020. Cross-modal cross-domain moment alignment network for person search. In *Proceedings of the IEEE/CVF Conference on Computer vision and Pattern Recognition*, 10678–10686.

Lee, K.-H.; Chen, X.; Hua, G.; Hu, H.; and He, X. 2018. Stacked cross attention for image-text matching. In *Proceedings of the European Conference on Computer Vision*, 201–216.

Li, C.; Lin, L.; Zuo, W.; and Tang, J. 2017. Learning patch-based dynamic graph for visual tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 4126–4132.

Li, K.; Zhang, Y.; Li, K.; Li, Y.; and Fu, Y. 2019. Visual semantic reasoning for image-text matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4654–4662.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Proceedings of the European Conference on Computer Vision*, 740–755.

Liu, X.; He, Y.; Cheung, Y.-M.; Xu, X.; and Wang, N. 2022. Learning relationship-enhanced semantic graph for fine-grained image–text matching. *IEEE Transactions on Cybernetics*, 54(2): 948–961.

Liu, Y.; Liu, M.; Huang, S.; and Lv, J. 2025. Asymmetric Visual Semantic Embedding Framework for Efficient Vision-Language Alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 5676–5684.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10012–10022.

Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Ma, X.; Xu, L.; Fang, L.; Zhang, C.; and Cui, L. 2025. Reliable Cross-modal Alignment via Prototype Iterative Construction. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 3847–3855.

Maddison, C. J.; Mnih, A.; and Teh, Y. W. 2016. The concrete distribution: A continuous relaxation of discrete random variables. *arXiv preprint arXiv:1611.00712*.

Pan, Z.; Wu, F.; and Zhang, B. 2023. Fine-grained image-text matching by cross-modal hard aligning network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19275–19284.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*, 8748–8763.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30*, 5998–6008.

Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Lio, P.; and Bengio, Y. 2017. Graph attention networks. *arXiv preprint arXiv:1710.10903*.

Xu, D.; Han, L.; Tan, M.; and Li, Y. F. 2009. Ceiling-based visual positioning for an indoor mobile robot with monocular vision. *IEEE Transactions on Industrial Electronics*, 56(5): 1617–1628.

Xu, T.; Zhang, P.; Huang, Q.; Zhang, H.; Gan, Z.; Huang, X.; and He, X. 2018. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1316–1324.

Young, P.; Lai, A.; Hodosh, M.; and Hockenmaier, J. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2: 67–78.

Zhang, G.; Sohn, K.; Hahn, M.; Shi, H.; and Essa, I. 2024. Finestyle: Fine-grained controllable style personalization for text-to-image models. In *Advances in Neural Information Processing Systems 37*, 52937–52961.

Zhang, K.; Mao, Z.; Wang, Q.; and Zhang, Y. 2022. Negative-aware attention framework for image-text matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15661–15670.

Zhang, Q.; Lei, Z.; Zhang, Z.; and Li, S. Z. 2020. Context-aware attention network for image-text retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3536–3545.

Zhang, Y.; and Lu, H. 2018. Deep cross-modal projection learning for image-text matching. In *Proceedings of the European Conference on Computer Vision*, 686–701.

Zhou, Y.; Wang, M.; Liu, D.; Hu, Z.; and Zhang, H. 2020. More grounded image captioning by distilling image-text matching model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4777–4786.

Zhu, J.; Li, Z.; Zeng, Y.; Wei, J.; and Ma, H. 2022. Image-text matching with fine-grained relational dependency and bidirectional attention-based generative networks. In *Proceedings of the 30th ACM International Conference on Multimedia*, 395–403.