Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Multi-Scale Semantic Alignment for Enhanced Image-Text Retrieval

Haifang Mo
South-Central Minzu University
Zili Su
South-Central Minzu University
Huili Zhang
zhl@mail.scuec.edu.cn

South-Central Minzu University
Meng Xia
South-Central Minzu University
Linhui Cheng
South-Central Minzu University
Qian Luo
South-Central Minzu University

Research Article

Keywords: Cross-Modal Image-Text Retrieval, Multi-Scale Feature Extraction, Adaptive Similarity Fusion, Semantic Complexity, Dynamic Sparse Aggregation

Posted Date: April 21st, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8805622/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: No competing interests reported.

Multi-Scale Semantic Alignment for Enhanced Image-Text Retrieval

Haifang Mo[1,2], Zili Su[1,2], Huili Zhang[1,3*], Meng Xia[1,2], Linhui Cheng[1,3], Qian Luo[1,2]

[1*]School of Computer Science, South-Central Minzu University, Wuhan, 430074, Hubei, China.
[2]Hubei Provincial Engineering Research Center for Intelligent Management of Manufacturing Enterprise, Wuhan 430074, Hubei, China.
[3]Hubei Provincial Engineering Research Center of Agricultural Blockchain and Intelligent Management, Wuhan 430074, Hubei, China.

*Corresponding author(s). E-mail(s): zhl@mail.scuec.edu.cn;
Contributing authors: mohaifang@mail.scuec.edu.cn;
2023120512@mail.scuec.edu.cn; xiameng@mail.scuec.edu.cn;
3007407@mail.scuec.edu.cn; 2023120504@mail.scuec.edu.cn;

Abstract

Cross-modal image-text retrieval aims to precisely match visual content with natural language descriptions, a task pivotal in multimodal understanding. Despite advancements in feature extraction and alignment, mainstream methods are constrained by isolated global or local matching strategies. This paper introduces a Cross-Modal Retrieval Network with Multi-Scale Feature Enhancement and Semantic-Aware Adaptive Fusion. The proposed method integrates visual and textual representations across different hierarchies, dynamically filters redundant visual information, and intensifies focus on regionally relevant image-text pairs. An adaptive fusion mechanism, guided by semantic complexity, intelligently balances global and local similarities. Experiments on Flickr30K and MS-COCO datasets demonstrate superior performance, validating the effectiveness and robustness of our approach.Our code is available at https://github.com/SUZILI7/MSSA

Keywords: Cross-Modal Image-Text Retrieval; Multi-Scale Feature Extraction; Adaptive Similarity Fusion; Semantic Complexity; Dynamic Sparse Aggregation

1 Introduction

Driven by advancements in AI and social media, multimodal data—notably images and text—has grown exponentially [1].Consequently, uncovering correlations between these heterogeneous modalities is a critical imperative. As a core cross-modal task, image-text retrieval (ITR) aims to match semantically relevant content across modalities (e.g., text-to-image search). Fundamentally, this task bridges the vision-language "semantic gap" offering significant theoretical value and supporting practical applications like e-commerce search [2] and visual question answering (VQA) [3].

Early research primarily mapped global image and text features into a common latent subspace [4, 5], but often failed to achieve fine-grained alignment. To mitigate this, attention-based mechanisms [6], and Transformer-based unified encoders [7–9]have been introduced to explore deeper semantic interactions.

Despite these advancements, two critical limitations persist. First, Vision Transformer (ViT)-based encoders typically partition images into numerous fixed-size patches [10], many of which contain redundant background noise that degrades retrieval precision. Second, existing approaches often rely on a single alignment strategy—either global or local—rendering them inflexible when handling descriptions of varying complexity. As illustrated in Fig. 1, simple descriptions often favor global matching, whereas complex scenarios necessitate fine-grained local alignment.

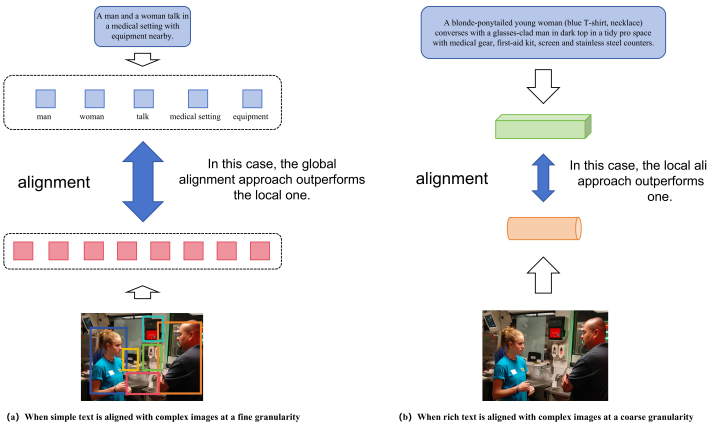


Fig. 1 Limitations of existing alignment paradigms. Fine-grained methods often suffer from insufficient semantic guidance when dealing with short, concise texts, whereas coarse-grained methods are highly susceptible to semantic misalignment for richly descriptive texts.

To address these limitations, we propose the Multi-Scale Semantic Alignment (MSSA) framework. Our approach integrates visual and textual representations across different hierarchies to ensure both global consistency and local discriminative power. Specifically, a dynamic sparse aggregation mechanism is utilized to filter redundant

visual information and intensify focus on relevant image-text pairs. Crucially, we introduce a Semantic-Aware Adaptive Fusion (SAAF) mechanism that quantifies textual complexity (e.g., length and entity density) to dynamically balance global and local similarities. This enables an intelligent strategy that adaptively modulates alignment granularity based on the semantic richness of the text.

The main contributions of this study are summarized as follows:

- We propose a Multi-Scale Feature Integration (MSFI) mechanism to enhance the representational capacity of global features. By leveraging the final-layer global context as a supervisory signal to guide the adaptive weighting of intermediate hierarchical features, MSFI effectively integrates multi-scale visual information while mitigating noise and information bottlenecks.
- We introduce a Hard-Mask Visual-Text Matching (HMVTM) module to enhance fine-grained alignment. Inspired by the hard assignment encoding paradigm [11], HMVTM targets core regions with the highest cross-modal relevance and filters irrelevant local pairs, significantly improving retrieval precision.
- We design the SAAF mechanism guided by textual complexity. By quantifying attributes such as length, semantic density, and lexical diversity, SAAF dynamically modulates the fusion weights of global and local similarities to achieve context-aware retrieval.
- We conduct extensive experiments on the Flickr30k and MS-COCO benchmarks. The results demonstrate that our method significantly outperforms state-of-the-art approaches, validating its effectiveness and robustness in diverse scenarios.

The remainder of this paper is organized as follows: Section 2 reviews related work; Section 3 details the proposed method; Section 4 presents experimental results; and Section 5 concludes the study.

2 Related Work

2.1 Global and Local Feature Matching

Global and Local Feature Matching Early research established joint embedding frameworks using CNNs and GRUs [12] , with VSE++ [13] introducing hard negative mining to enhance global alignment. To capture fine-grained details, local alignment methods like SCAN [6] pioneered cross-attention for region-word matching. Subsequent refinements have modeled intrinsic region relationships [14], incorporated negative constraints [15], or filtered representative pairs to suppress redundancy [11]. To handle noise in Vision Transformer patches, frameworks such as HREM [16], IRRA [17], and LAPS [18] have explored multi-level semantic capture, implicit reasoning, and language-aware filtering. Additionally, the Visual Anchor Alignment module [19] was proposed to mitigate imprecise local selection by mining representative image regions.

2.2 Multi-Level Alignment and Pre-training

Large-scale pre-training models, including CLIP [9], ALIGN [20], ALBEF [8], and the BLIP series [21, 22] leverage massive datasets to learn generalizable features. Despite

their success, integrating multi-level features remains crucial to prevent information loss. Researchers have sought to jointly optimize global and fine-grained alignments [23, 24] or utilize saliency detection [25] to boost performance. Notably, Zhu et al. [19] introduced multi-level semantic consistency (MSCA) to partition textual information into hierarchical layers. From a feature integration perspective, Ma et al. [26] proposed the MLSS framework, combining Transformers with GCN-based multi-label classifiers to supplement local details.

2.3 Research Motivation

While existing methods [6, 11, 19, 26] have advanced the field, most rely on static alignment strategies that do not adapt to varying textual complexity. Our work addresses this gap by integrating alignment strategies with textual complexity analysis, providing a dynamic mechanism that balances global consistency with local fine-grained details.

3 Proposed Method

The architecture of the proposed Multi-Scale Semantic Alignment (MSSA) framework is illustrated in Fig. 2. It builds upon the foundational structure of LAPS [18] and is primarily composed of four key components: the Image-Text Feature Extraction Module, the Sparse Aggregation Module, the Global/Local Similarity Calculation Module, and the Similarity Fusion Module. Detailed internal structures and the underlying mechanisms of these core components are further elaborated in the subsequent subsections.

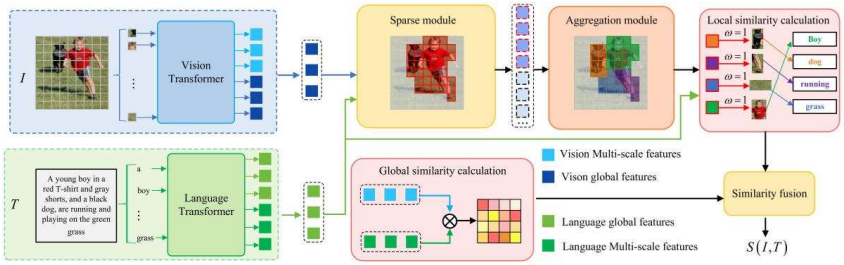


Fig. 2 The overall architecture of the proposed MSSA framework. For a given image-text pair (I, T), a dual-stream Transformer encoder is first employed to extract visual patch features and textual word features, incorporating multi-scale and global information. Subsequently, the image sparsification module identifies visually relevant patches guided by a language supervision mechanism. The aggregation module then consolidates these patches to generate distinctive semantic representations. On this basis, the local similarity calculation module matches each textual token with the most compelling evidence in the visual space. Concurrently, the global similarity is computed via dot-product operations on the multi-scale features. Finally, the similarity fusion module (guided by semantic complexity analysis) adaptively outputs the final similarity score of the image-text pair.

3.1 Image-Text Feature Extraction Module

The core objective of this module is to obtain cross-modal features that possess both global semantic consistency and local fine-grained discriminative power. To achieve this, we adopt a symmetrical design incorporating "Global Feature Extraction" and "Multi-Scale Feature Integration." This architecture effectively guarantees the consistency of cross-modal feature alignment and the completeness of the representation.

3.1.1 Image Feature Extraction via MSFI

The image feature extraction module is constructed based on the Vision Transformer (ViT-base) [27]. The input image is first resized to a resolution of 224×224 and fed into the ViT model to extract multi-level visual features. To preserve visual information across different semantic hierarchies, the model simultaneously extracts the hidden states from the backbone network's intermediate layers (specifically, layers 1, 3, and 5) to serve as multi-scale local features. These features are then projected into a unified dimension D. The global image feature V_{global} is obtained from the feature vector corresponding to the [CLS] token in the final layer, which is similarly projected into the D-dimensional space.

To further augment the representational capacity of the global features, we propose an attention-guided Multi-Scale Feature Integration (MSFI) mechanism. Specifically, the global feature is first expanded to match the spatial dimensions of the local features at each layer. It then interacts with these local features individually to generate fused embeddings. Subsequently, these embeddings are processed through a fully connected layer followed by Softmax normalization to generate attention weights for each local feature. Finally, the weighted multi-scale features are aggregated along the spatial dimension and fused via a linear layer to yield the enhanced multi-scale global image feature. The framework is illustrated in Fig. 3.

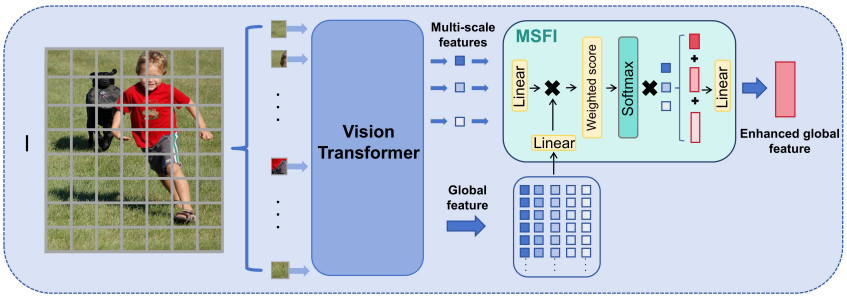


Fig. 3 Image Feature Extraction Module

For the features at the k-th scale (where $k \in \{1, 3, 5\}$), the weighted aggregated result is computed via a global feature-guided attention mechanism:

$$\mathbf{u}^{(k)} = \sum_{i=1}^{N} \mathbf{v}_i^{(k)} \cdot \text{Softmax}\left(\text{MLP}(\mathbf{v}_i^{(k)}) \odot \text{MLP}(V_{glo})\right) \tag{1}$$

where $\mathbf{v}_i^{(k)}$ denotes the feature of the i-th image patch at the k-th scale. The aggregated features from the three scales (corresponding to layers 1, 3, and 5) are concatenated and then projected through a linear layer to generate the enhanced global image feature:

$$\widetilde{V}_{glo} = \text{Linear}\left(\left[\mathbf{u}^{(1)}, \mathbf{u}^{(3)}, \mathbf{u}^{(5)}\right]\right). \tag{2}$$

3.1.2 Text Feature Extraction

Text feature extraction utilizes the BERT [28] model as its backbone. The input text is first tokenized and then fed into BERT to obtain the hidden state representations for each token. The global text feature T_{glo} is derived from the [CLS] token of the last layer and projected into a d-dimensional space. To capture the fine-grained semantic structure within the text, the model simultaneously extracts the hidden states from the first layer of BERT to serve as local features, which interact with the global feature via a bottom-up attention mechanism.

Specifically, the global and local features are mapped through separate fully connected layers and combined via element-wise multiplication to generate a fused representation. This representation is further processed by a fully connected layer and normalized using Softmax to yield attention weights for each token. Finally, the local features are weighted and summed according to these weights to obtain the enhanced global text feature $\widetilde{T}_{glo}$.

The attention weights for the local text features are computed under the guidance of the global feature, formulated as follows:

$$\beta_j = \text{Softmax}\left(\text{MLP}(\mathbf{t}_j) \odot \text{MLP}(T_{glo})\right) \tag{3}$$

where $\mathbf{t}_j \in \mathbb{R}^d$ denotes the local feature of the j-th token. The enhanced global text feature is obtained by computing the weighted sum of the local features across all tokens:

$$\widetilde{T}_{glo} = \sum_{j=1}^{M} \mathbf{t_j} \cdot \beta_j. \tag{4}$$

3.2 Sparse Aggregation Module

The large number of image patches generated by the visual encoder often contain redundant background information irrelevant to text semantics. Furthermore, individual image patches may lead to matching ambiguity due to semantic incompleteness. These issues hinder the effective utilization of the features extracted in Section 3. To

address this, our model adopts the "Sparse-Aggregation" two-stage framework proposed in LAPS and improves its sparsification strategy, aiming to achieve adaptive selection of key image patches.

3.2.1 Dynamic Sparsification

Dynamic sparsification implies that the number of retained visual patches, K, is determined dynamically rather than using a fixed selection ratio, thereby enabling the model to adapt to diverse matching scenarios.

Specifically, a base sparsity ratio r_{base} is first initialized. A fine-tuning offset is then added to this base ratio to obtain the dynamic sparsity ratio r_{dyn} calculated as follows:

$$r_{dyn} = \text{Clamp}\left(r_{base} + \mu \cdot \left(\sigma\left(W_p V_{glo}\right) - 0.5\right), 0.2, 0.8\right) \tag{5}$$

where W_p denotes the linear projection weights, and $\sigma(\cdot)$ represents the Sigmoid function. We center the output of the Sigmoid function by subtracting 0.5 and introduce a scaling factor μ (set to 0.1 in our experiments). This constrains the learned offset within a small range of $[-\mu/2, \mu/2]$. Finally, the number of retained visual patches K is determined by:

$$K = \lceil r_{dyn} \times N \rceil \tag{6}$$

where N denotes the total number of original visual patches.

Once the number of retained patches K is determined, we implement a saliency-based selection strategy. A learnable parameter α_v is introduced to dynamically balance the contributions of image self-attention and cross-modal attention. This enables the model to adaptively decide whether to prioritize visually salient objects or text-relevant regions during the pruning process. Accordingly, the comprehensive importance score for each visual patch is calculated as:

$$v_{score} = \alpha_v \cdot a_{self} + (1 - \alpha_v) \cdot a_{cross}. \tag{7}$$

Subsequently, all visual patches are sorted in descending order based on their scores v_{score}, and the top-K patches are selected to form the retained visual features, denoted as V_{select}.

To prevent the loss of non-key yet beneficial semantic information caused by over-sparsification, we perform a weighted average on the discarded patches based on their image self-attention scores. This aggregates them into a fused visual feature V_{extra}. This representation aims to preserve information from secondary regions in the image, serving as a supplement to the set of key aggregated visual patches.

3.2.2 Key Feature Aggregation

Feature aggregation facilitates the integration of correlated information, effectively mitigating the issue of semantic fragmentation [29] that may arise among key visual patches following dynamic sparsification. The aggregation module employs a network composed of a normalization layer, two linear layers, and a GELU activation function to learn a set of weights for each visual patch in V_{select}. Based on these weights,

multiple relevant visual patches are aggregated via weighted summation into several aggregated visual patches, where each represents a more complete semantic unit. Finally, these aggregated visual patches are concatenated with the previously generated fused visual feature V_{extra} to form the final optimized image feature V_{agg}. This feature possesses a more comprehensive semantic structure, providing a higher-quality representation for the subsequent similarity calculation.

3.3 Similarity Measurement

3.3.1 Global Similarity calculation

To enhance the model's capability in capturing the holistic semantic information of image-text pairs, we design a global similarity calculation module. Specifically, the global features of both images and texts are first subjected to L_2 normalization. This step aims to eliminate discrepancies in feature magnitudes, thereby stabilizing the subsequent similarity calculations. Subsequently, the global similarity is computed using the dot product.

Formally, for a batch of images, let $\widetilde{v}_{glo} \in \mathbb{R}^{B_v \times d}$ denote the normalized image features (where B_v represents the batch size and d is the feature dimension). Let $\widetilde{t}_{glo}^{(i)} \in \mathbb{R}^{1 \times d}$ denote the normalized feature of the i-th text. The global similarity between the i-th text T_i and all images I within the batch is calculated as follows:

$$S_{glo}(I, T_i) = \widetilde{v}_{glo} \cdot (\widetilde{t}_{glo}^{(i)})^T. \tag{8}$$

3.3.2 Local Similarity Calculation via HMVTM

Traditional local similarity calculations are prone to interference from irrelevant matches, leading to the dilution of key alignment signals. To address this, we design the HMVTM module, which employs Visual Hard Attention [11] to achieve precise alignment by focusing each text token on its best-matching visual evidence.

First, we apply L_2 normalization to V_{agg} and the text tokens to eliminate magnitude discrepancies. Subsequently, we compute the cosine similarity between them to obtain the fine-grained similarity matrix S, where $S_{i,j}$ denotes the matching degree between the i-th visual patch and the j-th text token.

To mask invalid regions (padding), we construct mask matrices for both visual and textual features. For visual features V_{agg} let l_v denote the actual number of patches for each sample in the batch, and $K_{\max}$ be the maximum number of patches. We construct an index matrix $\mathcal{I}_v \in \mathbb{R}^{B_v \times K_{\max}}$, and define the visual mask as $M^v = \mathbb{I}(\mathcal{I}_v < l_v)$, where $\mathbb{I}$ is the indicator function (taking the value 1 if the condition is met, and 0 otherwise). Similarly, based on the effective text token length l_t and the maximum token count $M_{\max}$, we construct the text mask M^t. The joint mask M is obtained via matrix multiplication:

$$M = M^v (M^t)^\top. \tag{9}$$

Subsequently, guided by M, we fill the invalid regions in S with $-\infty$:

$$\widetilde{S}_{i,j} = \begin{cases} S_{i,j}, & \text{if } M_{i,j} = 1 \\ -\infty, & \text{otherwise.} \end{cases} \tag{10}$$

Finally, we perform max pooling on $\widetilde{S}$ along the visual dimension. This implies that for each text token t_j, we select only the similarity of the best-matching visual patch as its alignment score s_j. This "Hard Alignment" mechanism effectively filters out noise introduced by irrelevant patches, thereby focusing on the most critical image-text correspondences.

Finally, the final local similarity is obtained via a Log-Sum-Exp (LSE) pooling operation with a configurable temperature coefficient τ (default set to 0.1):

$$S_{loc}(I,T) = \log\left(\sum_{j=1}^{l_t} \exp(s_j/\tau)\right). \tag{11}$$

3.4 Similarity Fusion via SAAF

In image-text retrieval, the importance of global and local cues varies significantly with the textual context. Simple descriptions focusing on isolated objects rely predominantly on global semantic consistency. Conversely, elongated queries depicting complex scenes with intricate spatial relations demand precise alignment of local entities. To address this discrepancy, we propose the Semantic-Aware Adaptive Fusion (SAAF) strategy, as illustrated in Fig. 4. This mechanism dynamically modulates the contributions of global similarity Sglobal and local similarity Slocal based on the intrinsic semantic complexity of the text.

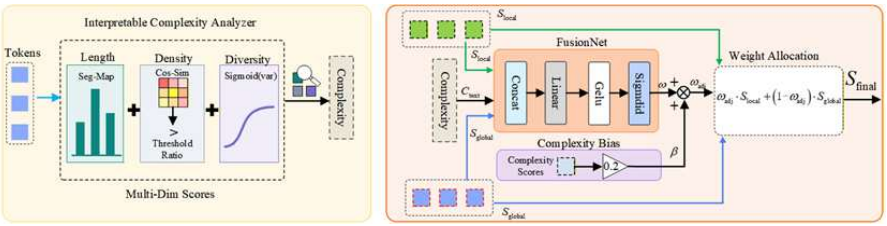


Fig. 4 Semantic Complexity-Guided Similarity Fusion Module

First, we design an Interpretable Complexity Analyzer. This module comprehensively evaluates the semantic complexity of the text from three complementary dimensions, ultimately generating a normalized complexity score C_{text}:

1. **Text Length** (C_{len}): According to the actual number of tokens of the text (n_{word}) it is mapped to four preset intervals: if $n_{word} < 5$, the score is 0.2; if $5 \leq n_{word} < 10$, the score is 0.4; if $10 \leq n_{word} < 20$, the score is 0.6; if $n_{word} \geq 20$, the score is 0.8.

2. **Semantic Density** ($C_{density}$): This metric quantifies the concentration of non-redundant semantic concepts within the text. To avoid overestimating density due to synonyms (e.g., "puppy" and "dog"), we employ embedding-based cosine similarity for robust redundancy detection. Specifically, for a token sequence $T = t_1, t_2, \ldots, t_{|T|}$, a token t_j is marked as redundant if its maximum similarity with any other token t_k ($k \neq j$) exceeds a threshold θ (set to 0.9). Accordingly, $C_{density}$ is formulated as:

$$C_{\text{density}} = \frac{1}{|T|} \sum_{j=1}^{|T|} \begin{cases} 1, & \text{if } \max\limits_{1 \leq k < j} \cos(t_j, t_k) \leq 0.9 \\ 0, & \text{otherwise} \end{cases} \tag{12}$$

3. **Lexical Diversity** (C_{div}): This metric evaluates the breadth of the semantic space covered by the text [30]. We quantify this by computing the average dispersion of token embeddings across feature dimensions [31]. Formally, let t_{ji} denote the value of the j-th token at the i-th dimension. C_{div} is defined as:

$$C_{\text{div}} = \sigma \left(\frac{1}{D} \sum_{i=1}^{D} \text{Var}\left(t_{i1}, t_{i2}, \ldots, t_{i|T_i|}\right) - b \right) \tag{13}$$

where D is the embedding dimension, $\text{Var}(\cdot)$ denotes the variance, $\sigma(\cdot)$ is the Sigmoid function, and b is a bias term.

To synthesize a comprehensive complexity score C_{text}, we aggregate the three metrics using a scheme that prioritizes semantic density for short texts, while emphasizing length and diversity for longer ones. A lightweight fusion network then predicts a base local weight ω, taking S_{loc}, S_{glo}, and C_{text} as inputs.

To amplify the regulatory impact of text complexity, we introduce a bias term to rectify the base weight, yielding ω_{adj}:

$$\omega_{\text{adj}} = \min\left(\max\left(\omega + \beta \cdot C_{\text{text}}, 0\right), 1\right) \tag{14}$$

where β is a hyperparameter controlling the bias intensity. The final image-text similarity S_{final} is derived via dynamic fusion:

$$S_{\text{final}} = \omega_{\text{adj}} \cdot S_{\text{loc}} + (1 - \omega_{\text{adj}}) \cdot S_{\text{glo}} \tag{15}$$

By employing this strategy, the model can adaptively accommodate texts of varying complexities.

3.5 Loss Function

Following previous methods, we use the bi-direction triplet loss with hard negative mining [13]:

$$\mathcal{L}_{align} = \sum_{(I,T)} \left[\alpha - S(I,T) + S\left(\widehat{I},T\right)\right]_+ + \left[\alpha - S(I,T) + S\left(I,\widehat{T}\right)\right]_+ \tag{16}$$

where α represents a margin parameter, $[x]_+ = \max(x, 0)$, and (I, T) is a positive image-text pair in the minibatch. We represent $\widehat{T} = \text{argmax}_{j \neq T} S(I, j)$ and $\widehat{I} = \text{argmax}_{i \neq I} S(i, I)$ as the hardest negative text and image examples within a mini-batch, respectively.

Furthermore, we constrain the ratio of the selected patches to a predefined value ρ for stable training [32], using a mean squared error loss to supervise the process. Finally, we combine the cross-modal alignment loss $\mathcal{L}_{align}$ with ratio constraint loss $\mathcal{L}_{ratio}$:

$$\mathcal{L} = \mathcal{L}_{align} + \mathcal{L}_{ratio}, \quad \mathcal{L}_{ratio} = \left(\rho - \frac{1}{N} \sum_{i=1}^{N} D_i \right)^2 . \tag{17}$$

At the inference stage, instead of using the Gumbel-Softmax sampling, we directly select constant N_s.

4 Experiments and Analysis

4.1 Datasets and Evaluation Metrics

We evaluate our model on two benchmark datasets: Flickr30K [33] and MS-COCO [34]. Flickr30K consists of 29,000 training, 1,014 validation, and 1,000 test images. MS-COCO comprises 82,738 training, 5,000 validation, and 5,000 test images. Results are reported based on the average of 5-fold cross-validation on the 1K test set and the performance on the full 5K test set. The primary evaluation metrics include Recall@K (R@K), which measures the proportion of correct matches within the top-K retrieved results (for K=1, 5, 10), and rSum, defined as the summation of all R@K scores across both image-to-text and text-to-image retrieval tasks.

4.2 Implementation Details

The experiments were implemented on two NVIDIA L20 GPUs running on an Ubuntu system equipped with 200GB of RAM. We employed the Vision Transformer (ViT-Base) with a patch size of 16×16 as the visual encoder, and BERT as the text encoder. The input images were resized to resolutions of 224×224 or 384×384, resulting in grid patches of 14×14 and 24×24, respectively. To unify the feature spaces, an additional linear layer was introduced on top of the encoders, mapping features to a common dimension of $D = 512$.

The entire framework was trained using the AdamW [35] optimizer with a batch size of 64 for 30 epochs. The initial global learning rate was set to 0.0002. We adopted a multi-step learning rate decay strategy, where the learning rate was decayed to 30% of its current value at epochs 9, 15, 20, and 25. The margin in the loss function was fixed at 0.2. Regarding specific hyperparameters, the weight parameter β was set to 0.8. For the ViT backbone, the selection ratio ρ and the aggregation ratio λ were set to 0.5 and 0.6, respectively.

4.3 Experimental Results

To evaluate the effectiveness of the proposed MSSA model, we conducted an extensive comparative analysis against several mainstream methods [19, 36–48]. As demonstrated in Table 1, MSSA achieves competitive results across various backbone configurations, validating its architectural robustness.

To further verify the performance, we carried out comprehensive experiments on both image-to-text and text-to-image retrieval tasks. For the MS-COCO dataset, we adopted two standard evaluation protocols: the 5-fold 1K test (averaging results over five folds) and the full 5K test set. Specifically, we compared our MSSA model with various state-of-the-art models, including early alignment methods like VSE++ [13] and SCAN [6], as well as recent representative models such as HREM [16], TGDT [49], CHAN [11], SGR [50], and LAPS [18]. The comparative results are presented in Table 2 and Table 3.

Table 1 Performance comparison with state-of-the-art methods on Flickr30K and MS-COCO 1K test sets.

		Flickr30K (1K Test)							MS-COCO (1K Test)						
		Image-to-Text			Text-to-Image				Image-to-Text			Text-to-Image			
Method	Backbone (Img / Txt)	R@1	R@5	R@10	R@1	R@5	R@10	rSum	R@1	R@5	R@10	R@1	R@5	R@10	rSum
VSRN	ResNet / Bi-GRU	71.3	90.6	96.0	54.7	81.8	88.2	482.6	76.2	93.8	98.2	62.8	89.7	95.1	516.8
SCAN	ResNet / Bi-GRU	67.4	90.3	95.8	48.6	77.7	85.2	465.0	73.4	93.8	97.8	57.5	87.3	94.3	504.1
NAAF	ResNet / Bi-GRU	78.3	94.1	97.7	58.9	83.3	89.3	501.3	78.9	96.0	98.7	63.1	90.8	96.5	524.0
DTAN	ResNet / Bi-GRU	79.1	95.1	97.1	61.0	85.9	91.7	510.5	80.7	96.6	98.7	67.0	92.3	96.7	532.0
KIDRR	ResNet / Bi-GRU	80.2	94.9	98.0	61.5	84.5	90.1	509.2	80.9	96.5	99.0	65.0	91.1	96.1	528.6
HREM	ResNet / Bi-GRU	81.4	96.5	98.5	60.9	85.6	91.3	514.3	81.2	96.5	98.9	63.7	90.7	96.0	527.1
CHAN	ResNet / Bi-GRU	79.7	94.5	97.3	60.2	85.3	90.7	507.8	79.7	96.7	98.7	63.8	90.4	95.5	525.0
SDE	ResNet / Bi-GRU	80.9	94.7	97.6	59.4	85.6	91.1	509.3	80.6	96.3	98.8	64.7	91.4	96.2	528.0
GAD	ResNet / Bi-GRU	80.1	96.0	97.9	60.2	85.5	90.4	510.1	80.6	96.4	98.3	64.4	90.8	96.0	526.5
MiC	ResNet / Bi-GRU	81.2	96.7	98.1	60.3	85.5	90.2	512.0	81.5	97.1	98.7	64.9	91.0	96.5	529.7
GLF	ResNet / Bi-GRU	81.7	96.5	98.6	61.8	85.2	90.7	514.4	81.8	96.6	98.9	65.2	90.5	95.7	528.7
BOOM	ResNet / Bi-GRU	82.1	96.4	98.3	61.5	86.0	91.4	515.7	80.5	96.7	99.0	64.5	91.1	96.3	528.1
RANet	ResNet / Bi-GRU	81.1	96.1	97.6	62.5	87.1	92.2	516.5	81.5	96.7	98.7	65.7	91.9	96.5	531.0
AME	ResNet / GloVe	77.1	95.1	97.3	61.2	86.1	91.5	508.3	66.1	78.5	98.7	63.7	90.1	95.6	522.7
ESA	ResNet / GloVe	82.6	95.9	98.1	61.1	85.9	91.1	514.7	79.6	96.5	98.7	63.5	90.9	96.1	525.3
ESL	ResNet / GloVe	83.9	96.4	98.8	62.5	87.3	92.3	521.2	81.3	96.8	99.1	65.3	91.0	96.8	530.2
CORA	ResNet / GloVe	82.3	96.1	98.0	63.0	87.4	92.8	519.6	81.7	96.7	99.0	66.0	92.0	96.7	532.1
MSCA	ResNet / GloVe	84.4	96.0	98.4	70.1	89.2	92.8	530.9	82.9	97.2	98.9	71.2	92.9	96.7	539.7
LAPS	ViT / BERT	79.0	96.0	98.1	67.3	90.5	94.5	525.4	78.6	96.3	98.9	68.0	92.4	96.8	531.0
MSSA	ViT / BERT	**82.1**	**96.4**	**98.4**	**68.9**	**92.0**	**95.7**	**533.5**	**81.8**	**97.3**	**99.3**	**70.3**	**93.4**	**97.3**	**539.2**

As evidenced by the experimental results, the proposed MSSA model consistently outperforms existing state-of-the-art methods across both datasets, underscoring the efficacy and robustness of our multi-scale adaptive fusion architecture. Specifically, on the Flickr30K benchmark using the ViT-Base-224 backbone, MSSA achieves a notable 3.1-point increase in rSum compared to the baseline LAPS. This performance margin further expands to 8.1 points when utilizing the ViT-Base-384 backbone, demonstrating a substantial competitive advantage over contemporary retrieval models.

On the 1K and 5K test sets of MS-COCO, our method achieves the optimal performance regardless of whether ViT-Base-224 or ViT-Base-384 is adopted. For ViT-Base-384, the rSum of our method is 8.2 points higher than that of LAPS on COCO-1K.

Table 2 Comparisons of image-text retrieval performances on Flickr30K and MS-COCO test set. We list the details of feature encoding, image resolution, and the number of obtained regions/patches by visual encoder (e.g. "ViT-Base-224" represents the base-version of Vision Transformer with 224×224 image resolution input, regarding 16×16 pixels as one patch, and getting 14×14 visual patches for one image). FG indicates whether it is the fine-grained cross-modal alignment. Our method is marked bold.

		Flickr30K (1K Test)							MS-COCO (1K Test)						
		Image-to-Text			Text-to-Image				Image-to-Text			Text-to-Image			
Method	FG	R@1	R@5	R@10	R@1	R@5	R@10	rSum	R@1	R@5	R@10	R@1	R@5	R@10	rSum
Faster R-CNN + BERT-Base, 36 pre-computed regions															
HREM	✗	83.3	96.0	98.1	63.5	87.1	92.4	520.4	81.1	96.6	98.9	66.1	91.6	96.5	530.7
TGDT	✗	61.3	86.0	91.4	76.8	93.2	96.4	505.1	65.4	91.8	96.5	78.5	96.4	98.9	527.5
CHAN	✗	80.6	96.1	97.8	63.9	87.5	92.6	518.5	81.4	96.9	98.9	66.5	92.1	96.7	532.6
ViT-Base-224 + BERT-base, 14×14 patches															
VSE++	✗	71.8	92.8	96.5	59.4	84.7	90.9	496.1	75.0	94.6	98.0	62.7	89.4	94.9	514.6
SCAN	✓	69.5	90.9	95.6	56.4	83.1	90.0	485.6	76.0	95.4	98.1	64.5	90.8	95.8	520.6
SGR	✓	69.7	90.8	95.2	59.1	84.1	89.0	488.7	77.2	95.0	98.0	65.1	90.7	95.8	521.8
CHAN	✓	69.2	91.8	95.0	58.4	84.9	90.6	489.9	77.1	95.1	98.1	65.0	91.0	96.0	522.2
LAPS	✓	74.0	93.4	97.4	62.5	87.3	92.7	507.3	78.7	95.5	98.3	66.2	91.3	96.2	526.3
MSSA	✓	**75.8**	**94.2**	**97.2**	**62.4**	**87.6**	**93.1**	**510.4**	**79.3**	**95.5**	**98.6**	**66.1**	**91.9**	**96.6**	**528.8**
ViT-Base-384 + BERT-base, 24×24 patches															
VSE++	✗	77.1	95.7	97.5	65.8	90.2	94.3	520.5	77.0	95.7	98.4	64.6	91.1	96.2	523.0
SCAN	✓	75.4	94.4	96.9	63.6	88.6	93.5	512.5	76.1	95.5	98.5	65.1	91.6	96.3	523.1
SGR	✓	76.9	94.9	98.1	64.2	88.4	93.3	515.8	75.8	95.7	98.6	65.6	92.0	96.5	524.2
CHAN	✓	75.4	94.5	97.6	63.2	88.6	93.1	512.4	78.1	95.8	98.6	66.1	92.1	96.6	527.3
LAPS	✓	79.0	96.0	98.1	67.3	90.5	94.5	525.4	78.6	96.3	98.9	68.0	92.4	96.8	531.0
MSSA	✓	**82.1**	**96.4**	**98.4**	**68.9**	**92.0**	**95.7**	**533.5**	**81.8**	**97.3**	**99.3**	**70.3**	**93.4**	**97.3**	**539.2**

Table 3 Image-text retrieval performance on the MS-COCO 5K test set.

		Image-to-Text			Text-to-Image			
Method	FG	R@1	R@5	R@10	R@1	R@5	R@10	rSum
Faster R-CNN + BERT-Base, 36 pre-computed regions								
HREM	✗	62.3	87.6	93.4	43.9	73.6	83.3	444.1
TGDT	✗	43.3	73.5	83.3	57.5	84.8	91.6	434.0
CHAN	✗	59.8	87.2	93.3	44.9	74.5	84.2	443.9
ViT-Base-224 + BERT-base, 14×14 patches								
VSE++	✗	52.4	80.3	88.8	40.6	70.4	81.1	413.4
SCAN	✓	53.9	81.8	90.0	42.9	72.3	82.5	423.5
SGR	✓	54.9	82.8	90.5	42.8	72.2	82.5	425.8
CHAN	✓	56.3	83.2	90.1	43.0	72.6	82.8	428.8
LAPS	✓	57.5	84.0	90.8	44.5	74.0	83.6	434.4
MSSA	✓	**57.8**	**84.9**	**92.3**	**44.3**	**73.9**	**83.7**	**436.9**
ViT-Base-384 + BERT-base, 24×24 patches								
VSE++	✗	54.9	82.8	90.4	42.4	72.4	82.8	425.8
SCAN	✓	53.3	81.8	90.0	42.6	72.6	82.9	423.1
SGR	✓	53.3	81.0	89.6	42.9	73.1	83.7	423.6
CHAN	✓	55.6	83.8	91.2	43.4	73.6	83.5	431.1
LAPS	✓	57.4	84.9	92.5	46.4	75.8	85.2	442.2
MSSA	✓	**58.8**	**86.0**	**93.5**	**47.8**	**76.8**	**86.2**	**449.3**

Analyzing the experimental results, we observe that VSE++ relies solely on global features for image-text alignment. While it possesses the advantage of computational efficiency, its performance is limited when handling texts with complex semantic relationships due to its failure to capture fine-grained local semantic interactions. The SCAN method introduces a dense cross-attention mechanism, mining local correlations by performing pairwise matching between all image regions and text words. However, its fully connected attention pattern not only introduces substantial background noise but also suffers from low computational efficiency.

HREM, TGDT, SGR, CHAN, and our method all focus on optimizing feature alignment and interaction. However, our approach specifically addresses the core limitations of these preceding methods. Specifically, HREM and TGDT rely heavily on pre-trained object detectors to extract image features. This non-end-to-end modeling approach restricts global context awareness and fails to achieve deep semantic mining through multi-scale features. SGR attempts to model image-text relationships via Graph Neural Networks based on similarity graph reasoning. However, when facing the massive fragmented patches generated by ViT, constructing graph structures increases computational costs and is prone to introducing false associations. Although CHAN achieves efficient retrieval through a "hard alignment" strategy, its aggressive information discarding mechanism often leads to fine-grained semantic discontinuity. LAPS utilizes semantic information to guide a sparse model in selecting image features. However, it is susceptible to interference from variations in text expression, resulting in incomplete retention of effective information and affecting subsequent alignment precision.

In contrast, our method demonstrates significant advantages. The improved sparse module more precisely eliminates redundant image patches. Combined with the HMVTM module, it focuses on the core regions with the highest image-text similarity, effectively mitigating the interference of text expression variations on feature selection. Simultaneously, we introduce coarse- and fine-grained dual-dimensional similarity calculations and utilize the SCGA module to adaptively balance global and local features based on text complexity. Finally, coupled with the MSFI module, our method achieves superior retrieval performance and enhanced model interpretability.

4.4 Impact Analysis of Margin Value

To investigate the influence of the margin value α in the Triplet Loss function, we conducted sensitivity experiments on the Flickr30K dataset using the ViT-Base-224 backbone. By varying α while keeping other components constant, we assessed its impact on the rSum metric. The results are summarized in Table 4.

The margin α in Triplet Loss serves as a critical constraint that governs the relative distances between the anchor, positive, and negative samples, it requires the distance of positive pairs to be smaller than α, while that of negative pairs should be larger. Experimental results indicate that as the value of α increases, the model performance (measured by rSum) initially improves before declining. It achieves optimal performance at $\alpha = 0.2$, after which a continued increase in α leads to a downward trend. Consequently, we set the margin value α for the Triplet Loss to 0.2 in this work.

Table 4 Boundary value impact analysis results.

Boundary Value (α)	rSum
0.0	472.5
0.1	501.3
0.2	510.4
0.3	508.0
0.4	497.7

4.5 Ablation Study

To assess the contribution of the three proposed core modules—MSFI, HMVTM, and SAAF—to the experimental results, we conducted ablation studies by progressively integrating each module into the baseline model. Adopting ViT-Base-224 and BERT as the underlying architecture, the experimental results on the Flickr30K and MS-COCO datasets are presented in Table 5 and Table 6, respectively.

Table 5 Ablation experiments on Flickr30K dataset (%).

Experimental Method	Image to Text Retrieval			Text to Image Retrieval			rSum
	R@1	R@5	R@10	R@1	R@5	R@10	
Baseline Model	74.0	93.4	97.4	62.5	87.3	92.7	507.3
MSFI	72.5	92.3	92.3	61.1	86.3	90.2	496.8
MSFI+HMVTM	76.6	93.7	96.6	61.6	87.5	92.9	508.9
MSFI+HMVTM+SAAF	75.8	94.2	97.2	62.4	87.6	93.1	510.4

Table 6 Ablation experiments on MS-COCO dataset (%).

Experimental Method	Image to Text Retrieval			Text to Image Retrieval			rSum
	R@1	R@5	R@10	R@1	R@5	R@10	
Baseline Model	78.7	95.5	98.3	66.2	91.3	96.2	526.3
MSFI	77.0	94.5	97.8	65.5	90.2	95.6	520.6
MSFI+HMVTM	78.2	95.2	98.2	65.8	91.0	96.0	524.4
MSFI+HMVTM+SAAF	79.3	96.3	98.6	66.1	91.9	96.6	528.8

As shown in Tables 5 and 6, adding only the MSFI to the baseline model leads to a clear decline in retrieval performance and still falls short of the LAPS benchmark. This drop is likely due to the feature distribution distortion caused by simple linear fusion, where multi-scale hierarchical features and sparsely aggregated patches may overlap or conflict without refined alignment.

However, upon integrating the HMVTM, the model's performance exhibits a significant upward trajectory. On the Flickr30K dataset, the rSum score surpasses LAPS

by 1.3 points, benefiting from the hard-assignment mechanism that filters textual expression noise and anchors the most salient visual-textual correspondences. While the improvement on MS-COCO is less pronounced—due to its relatively concise captions providing less granular guidance—the multi-scale cues from MSFI successfully compensate for potential information loss during core-region focusing.

The final integration of the SAAF (Semantic-Aware Adaptive Fusion) module yields the most substantial gains across both datasets, markedly enhancing model robustness and generalization. The core strength of SAAF lies in its ability to dynamically modulate the contributions of global and local similarities based on textual semantic complexity. In scenarios like MS-COCO, where captions may lack exhaustive detail, SAAF adaptively increases the weight of global context to mitigate the limitations of fine-grained alignment. This adaptive synergy between MSFI, HMVTM, and SAAF ensures that MSSA maintains superior performance across diverse linguistic and visual landscapes.

4.6 Visualization and Case Studies

To further validate the rationality and efficacy of our approach, this section selects representative image-text pairs—ranging from simple phrases to complex sentences—for a comparative analysis with the LAPS baseline. We focus on how textual complexity influences weight allocation and the final similarity scores.



1. A man with a gray helmet on skis on a snowy ski slope.

2. Gray-helmeted person skis on snow.

3. Man holding ski poles standing on a snowy surface in a ski resort.

4. A woman gripping ski poles standing on skis in the snow.

5. A snowy area with a group of skiers rests to the foreground of a ski spot with a small cabin and a background of bare trees beneath a bright blue sky.

(a)



1. A man looks at his phone at an indoor table.

2. Some people gather around an indoor dining table — one of them is operating a laptop, and a snack package is also on the table.

3. A group of people cluster by tables in an indoor lounge area: some are looking down at their devices, while others are interacting with those next to them.

4. A woman in a black top and light gray pants is walking through the hallway of the indoor lounge area.

5. In the indoor lounge, people either sit by tables using electronic devices or chat. Hanging chandeliers, metal pillars and an extending corridor are visible in the background, giving the space a lively, busy vibe.

(b)

Fig. 5 On the left are the test images; on the right are their five corresponding descriptions, which vary in both length and focus.

Fig. 5 presents two images, each paired with five descriptive texts of varying lengths and focuses. The experimental results for Fig. 5(a) and 5(b) are shown in Fig. 6 and 7, respectively. Both Fig. 6 and 7 share the same layout: subfigure (a) displays the semantic complexity C_{text} of the five texts; subfigure (b) illustrates the weights assigned to global and local similarity; subfigures (c) and (d) respectively show the

final image-text similarity scores S_{final} obtained by the proposed model MSSA and the baseline model LAPS for the five texts.

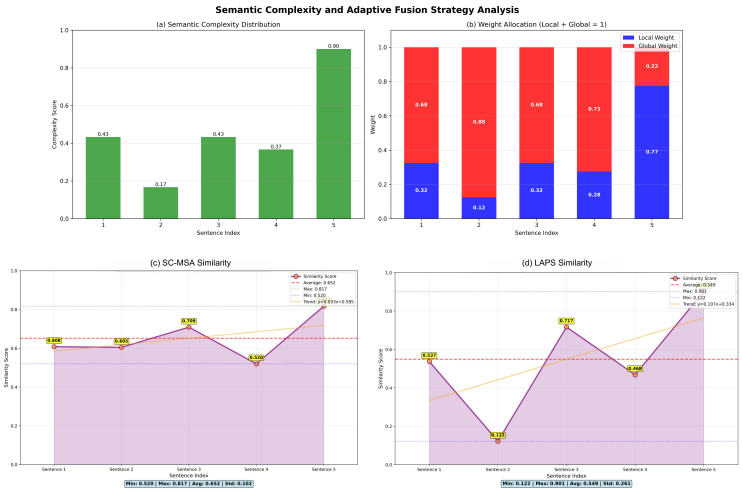


Fig. 6 Experimental Results for the Case in Figure 5(a).

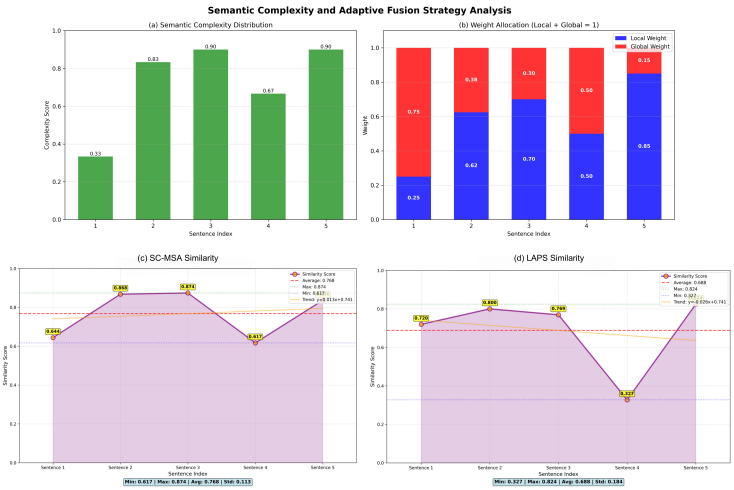


Fig. 7 Experimental Results for the Case in Figure 5(b).

From subfigures (a) and (b) of Fig. 6 and 7, it can be observed that when the textual semantic complexity is high, the corresponding weight for local similarity increases significantly, indicating that the model places greater emphasis on local feature alignment. Conversely, for texts with lower semantic complexity, the weight for

global similarity is correspondingly higher, reflecting a stronger focus on global feature alignment. Subfigures (c) and (d) reveal that MSSA achieves a higher average and lower variance in similarity scores across the five texts, demonstrating its stronger and more robust image-text alignment capability.

Moreover, the performance difference between the two models becomes particularly pronounced when addressing the challenge of "non-salient local objects" (e.g., Text 4 in both experimental cases). Due to the inconspicuous or non-central visual characteristics of the objects described in Text 4, the LAPS model is highly susceptible to failure caused by the excessive filtering of its sparse attention strategy. For instance, in the complex scene of Fig. 5(b), LAPS exhibits a sharp drop in similarity score for Text 4. In contrast, the MSSA maintains strong retrieval performance even on such challenging samples. This advantage can be primarily attributed to the collaborative mechanism of the model: the SAAF sensitively captures the specificity of the text and allocates balanced weights to prevent details from being overlooked; meanwhile, the HMVTM, through its hard-assignment focusing mechanism, compels the model to concentrate on the core regions most relevant to the textual entities, effectively filtering out background noise.

5 Conclusion

We propose the Multi-Scale Semantic Alignment (MSSA) framework for enhanced image-text retrieval. Our approach employs an improved sparse aggregation module to filter out redundant visual patches, while the MSFI module integrates multi-scale image features to enhance the precise capture of key semantic regions in images. The HMVTM focuses on core cross-modal correspondences between image and text. Furthermore, the SAAF dynamically adjusts the weights between global and local features based on textual semantic complexity. This not only reinforces global semantic consistency but also effectively extracts fine-grained local details from the text, ensuring both comprehensiveness and relevance in cross-modal feature representation. Overall, our method effectively balances the trade-off between global semantic matching and local detail alignment, significantly improving the overall performance of image-text retrieval. Ablation studies and visualization experiments validate the effectiveness of each core component and highlight their critical contributions to enhancing retrieval performance.

However, the proposed method also has certain limitations. When the text contains figurative language such as metaphors or personification, the semantic complexity computed by the SAAF may not accurately reflect the true semantic complexity, thereby affecting the rationality of feature weight adjustment. Additionally, the current model relies on pre-trained parameters from ViT and BERT for feature extraction, and its generalization capability on tasks involving low-resolution images or low-resource languages requires further improvement.

Declarations

Funding

1. Hubei Province Key Research and Development Special Project of Science and Technology Innovation Plan(No. 2023BAB087)
2. Wuhan Key Research and Development Projects(No. 2023010402010614)
3. Supported by "the open competition project for selecting the best candidates", Wuhan East Lake High-techDevelopment Zone (No. 2024KJB328)
4. Supported by the Fund for Research Platform of South-Central Minzu University (No. PTZ25001)
5. Central Government Guides Local Funds for Science and Technology Development(No. ZYYD2024QY08)

Conflict of interest/Competing interests

The authors declare that they have no conflict of interest.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Data availability

The datasets used in this study are publicly available: Flickr30K and MS-COCO.

Materials availability

Not applicable.

Code availability

The source code and pre-trained models for the proposed MSSA framework are made available to the research community at: https://github.com/SUZILI7/MSSA.

Author contribution

Haifang Mo: Conceptualization, Methodology, Software, Writing - original draft. Zili Su: Data curation, Validation, Visualization. Huili Zhang: Supervision, Writing - review & editing.

References

[1] Yin S, Fu C, Zhao S, Li K, Sun X, Xu T., Chen E.: A survey on multimodal large language models. arXiv preprint arXiv:2306.13549 (2023)

[2] Liu C, Hou P, Zeng A, Yu H: Transformer-empowered multi-modal item embedding for enhanced image search in e-commerce. arXiv preprint arXiv:2311.17954 (2023) https://doi.org/10.48550/arXiv.2311.17954

[3] Dalu Guo, Chang Xu, Dacheng Tao: Image-question answer synergistic network for visual dialog. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10426–10435 (2019)

[4] Frome A., Corrado G.S., Shlens J., Bengio S., Dean J., Ranzato M., Mikolov T.: Devise: A deep visual-semantic embedding model. In: Advances in Neural Information Processing Systems (NeurIPS), vol. 26, pp. 2121–2129 (2013)

[5] Wang L., Li Y., Lazebnik S.: Learning deep structure-preserving image-text embeddings. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5005–5013 (2016). https://doi.org/10.1109/CVPR.2016.544

[6] Lee K.-H., Chen X., Hua G., Hu H., He, X.: Stacked cross attention for image-text matching. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 201–216 (2018). https://doi.org/10.1007/978-3-030-01225-0_13

[7] Chen Y.-C., Li L., Yu L., Ahmed A. E. K., Gan Z., Cheng Y., Liu J.: Uniter: Universal image-text representation learning. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 1–10 (2019). https://doi.org/10.1109/ICCV.2019.00001

[8] Li J., Selvaraju R. R., Gotmare A., Joty S. R., Xiong C., Hoi S. C.: Align before fuse: Vision and language representation learning with momentum distillation. In: Advances in Neural Information Processing Systems (NeurIPS), pp. 9694–9705 (2021)

[9] Radford A., Kim J. W., Hallacy C., Ramesh A., Goh G., Agarwal S., Sastry G., Askell A., Mishkin P., Clark J., *et al.*: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning (ICML), pp. 8748–8763 (2021)

[10] Rao Y., Zhao W., Liu B., Lu J., Zhou J., Hsieh C. J.: Dynamicvit: Efficient vision transformers with dynamic token sparsification. In: Advances in Neural Information Processing Systems (NeurIPS), pp. 13937–13949 (2021)

[11] Pan Z., Wu F., Zhang B.: Fine-grained image-text matching by cross-modal hard aligning network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 19275–19284 (2023)

[12] Kiros R., Salakhutdinov R., Zemel R. S.: Unifying visual-semantic embeddings with multimodal neural language models. arXiv preprint arXiv:1411.2539 (2014)

[13] Faghri F., Fleet D. J., Kiros J. R., Fidler S.: Vse++: Improving visual-semantic embeddings with hard negatives. arXiv preprint arXiv:1707.05612 (2017)

[14] Qu L., Liu M., Cao D., Nie L., Tian Q.: Context-aware multi-view summarization network for image-text matching. In: Proceedings of the 28th ACM International Conference on Multimedia (MM), pp. 1047–1055. ACM, New York (2020)

[15] Zhang K., Mao Z., Wang Q., *et al.*: Negative-aware attention framework for image-text matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 15661–15670 (2022)

[16] Fu Z., Mao Z., Song Y., Zhang Y.: Learning semantic relationship among instances for image-text matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 15159–15168 (2023)

[17] Jiang D., Ye M.: 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 (CVPR), pp. 15661–15670 (2023)

[18] Fu Z., Zhang L., Xia H., Mao Z.: Linguistic-aware patch slimming framework for fine-grained cross-modal alignment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 15159–15168 (2024)

[19] Zhu L., Han D., Shen X., Chen C., Li K.-C.: Enhancing image-text matching through multi-level semantic consistency alignment. The Visual Computer **41**, 9555–9570 (2025)

[20] Jia C., Yang Y., Xia Y., Chen Y., Parekh Z., Pham H., Le Q.V., Sung Y., Li Z., Duerig T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 15159–15168 (2021)

[21] Li J., Li D., Xiong C., *et al.*: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: Proceedings of the 40th International Conference on Machine Learning (ICML 2023), vol. 202, pp. 18928–18943 (2023)

[22] Li Junnan, Li Dongxu, Xiong Caiming, Hoi Steven: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International Conference on Machine Learning (ICML) (2022)

[23] Qi J., Peng Y., Yuan Y.: Cross-media multi-level alignment with relation attention network. In: Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI), pp. 1–7 (2018)

[24] Zhang Y., Li H.: Deep relation embedding for cross-modal retrieval. IEEE Transactions on Image Processing **30**, 617–627 (2020)

[25] Ji J., Wang H., Han J., *et al.*: Saliency-guided attention network for image-sentence matching. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 5754–5763 (2019)

[26] Ma Y., Wang M., Lu G., Sun Y.: Multi-label semantic sharing based on graph convolutional network for image-to-text retrieval. The Visual Computer **41**, 1827–1840 (2025)

[27] Dosovitskiy A., Beyer L., Kolesnikov A., Weissenborn D., Zhai X., Unterthiner T., Dehghani M., Minderer M., Heigold G., Gelly S., Uszkoreit J., Houlsby N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (ICLR), pp. 1–16 (2020)

[28] Devlin J., Chang M.-W., Lee K., Toutanova K.: 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 (NAACL-HLT), pp. 4171–4186 (2019)

[29] Zong Z., Li K., Song G., Wang Y., Qiao, Y., Leng, B., Liu, Y.: Self-slimmed vision transformer. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 432–448 (2022)

[30] Yiu E., Chien J., Mahowald K.: Linguistic blind spots of large language models: Experiments on lexical density and syntactic complexity. arXiv preprint arXiv:2503.19260 (2025)

[31] McCarthy P. M., Jarvis S.: Mtld, vocd-d, and hd-d: A validation study of sophisticated approaches to lexical diversity assessment. Behavior Research Methods **42**(2), 381–392 (2010)

[32] Selvaraju R. R., Cogswell M., Das A., Vedantam R., Parikh D., Batra D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 618–626 (2017)

[33] Young Peter, Lai Alice, Hodosh Micah, Hockenmaier Julia: 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 (2014)

[34] Lin T., Maire, M., Belongie, S., Hays, J., Perona P., Dollár P., Zitnick C. L.: Microsoft coco: Common objects in context. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 740–755 (2014)

[35] Loshchilov I., Hutter F.: Decoupled weight decay regularization. In: Proceedings of the 7th International Conference on Learning Representations (ICLR), New

Orleans, LA, USA, pp. 1–12 (2019)

[36] Li K., Zhang Y., Li K., Li Y., Fu Y.: Visual semantic reasoning for image-text matching. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 4654–4662 (2019)

[37] Chen C., Wang D., Song B., Tan H.: Inter-intra modal representation augmentation with dct-transformer adversarial network for image-text matching. IEEE Transactions on Multimedia **25**, 8933–8945 (2023)

[38] Xie X., Li Z., Tang Z., Yao D., Ma H.: Unifying knowledge iterative dissemination and relational reconstruction network for image-text matching. Information Processing & Management **60**(1), 103154 (2023)

[39] Kim D., Kim N., Kwak S.: Improving cross-modal retrieval with set of diverse embeddings. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 23422–23431 (2023)

[40] Pu X., Wang Z., Yuan L., Wu Y., Jing L., Gao X.: Gadnet: improving image-text matching via graph-based aggregation and disentanglement. Pattern Recognition **157**, 110900 (2025)

[41] Pu X., Chen Y., Yuan L., Zhang Y., Li H., Jing L., Gao X.: Mic: image-text matching in circles with cross-modal generative knowledge enhancement. Knowledge-Based Systems **289**, 111503 (2024)

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

[43] Li Z., Zhang L., Zhang K., Zhang Y., Mao Z.: Improving image-text matching with bidirectional consistency of cross-modal alignment. IEEE Transactions on Circuits and Systems for Video Technology **34**(7), 6590–6607 (2024)

[44] Xiong G., Meng M., Zhang T., Zhang D., Zhang Y.: Reference aware adaptive network for image-text matching. IEEE Transactions on Circuits and Systems for Video Technology **34**(10), 9678–9691 (2024)

[45] Li J., Niu L., Zhang L.: Action-aware embedding enhancement for image-text retrieval. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), vol. 36, pp. 1323–1331 (2022)

[46] Zhu H., Zhang C., Wei Y., Huang S., Zhao Y.: Esa: external space attention aggregation for image-text retrieval. IEEE Transactions on Circuits and Systems for Video Technology **33**(10), 6131–6143 (2023)

[47] Zhang K., Hu B., Zhang H., Li Z., Mao Z.: Enhanced semantic similarity learning framework for image-text matching. IEEE Transactions on Circuits and Systems

for Video Technology **34**(4), 2973–2988 (2023)

[48] Pham K., Huynh C., Lim S.-N., Shrivastava A.: Composing object relations and attributes for image-text matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 14354–14363 (2024)

[49] Liu C., Zhang Y., Wang H., Chen W., Wang F., Huang Y., Shen Y., Wang L.: Efficient token-guided image-text retrieval with consistent multimodal contrastive training. IEEE Transactions on Image Processing **32**(12), 1234–1247 (2023) https://doi.org/10.1109/TIP.2023.3286710

[50] Diao Haiwen, Zhang Ying, Ma Lingyun, Lu Huchuan: Similarity reasoning and filtration for image-text matching. arXiv preprint arXiv:2101.01368 (2021)