ADAPTIVE MULTI-WORD ANCHORING NETWORK FOR FINE-GRAINED IMAGE-TEXT MATCHING

Anonymous authors
Paper under double-blind review

ABSTRACT

Image-text matching bridges vision and language, which is a fundamental task in multimodal intelligence. The key challenge lies in establishing accurate alignment between two heterogeneous modalities. Existing methods typically rely on either dense soft attention, which suffers from cross-modal noise, or rigid hard assignment, which causes semantic information loss by artificially restricting region-word correspondences to one-to-one mappings. To address the inherent semantic density imbalance between visual and textual modalities, we propose an Adaptive Multi-Word Anchoring Network (MWAN). Specifically, we introduce an Anchored Alignment Mechanism (AAM) as a theoretically motivated adaptive gating approach tailored for cross-modal matching. By utilizing context-aware dynamic thresholds, AAM effectively grounds each image region to multiple semantically relevant words, achieving robust many-to-many alignment. Moreover, to overcome the distinguishability bottleneck in top-ranked retrieval results, we design a Residual Semantic Re-ranking (RSR) module that explicitly isolates and exploits previously suppressed residual information to discriminate among semantically similar candidates. Extensive experiments on Flickr30K and MSCOCO demonstrate that MWAN achieves superior performance. Notably, RSR also serves as a universal plug-and-play refinement module to enhance existing methods.

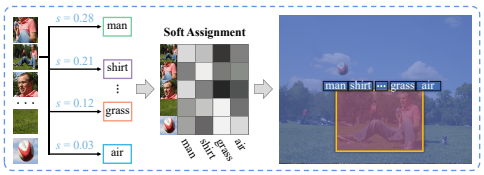
1 INTRODUCTION

Image-text matching has become an essential technique for various applications, such as visual question answering (Liu et al., 2025b; Hong et al., 2026; Ning et al., 2026; Huang et al., 2026), image captioning (Hua et al., 2025; Zhang et al., 2026), cross-modal retrieval (Yin et al., 2026; Chen et al., 2026; Meng et al., 2026) and so forth. Despite considerable development in recent years, establishing accurate visual-semantic alignment remains a significant challenge.

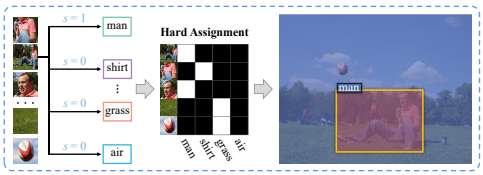
Early approaches typically focused on global alignment, mapping whole images and sentences into a common space. However, these methods often fail to capture fine-grained details. To address this, recent works have shifted focus to fragment-level alignment (Wang et al., 2025b; Chen et al., 2025b; Liu et al., 2026), such as regions in images and words in texts. SCAN (Lee et al., 2018) and its variants (Liu et al., 2025a; Zhang et al., 2025; Tang et al., 2026) are representative of this school of thought. They utilize cross-modal attention mechanisms to aggregate all possible region-word pairs, as shown in Figure 1 (a). While effective, these methods often suffer from redundant alignments, where irrelevant fragments introduce noise into the similarity computation. Conversely, inspired by quantization ideas (Huang et al., 2013), methods such as CHAN (Pan et al., 2023) and its successors (Li et al., 2024b; Pu et al., 2025; Ji & Liu, 2026; Liu et al., 2026) treat text words as a codebook and enforce a rigid hard assignment scheme. As depicted in Figure 1 (b), these methods associate each region with a single word, which effectively reduces noise but limits semantic diversity.

However, bridging the modality gap through fragment alignment remains hindered by two critical limitations: **1) The inherent semantic density imbalance between visual regions and textual words.** Visual regions are often semantically dense compared to individual words. A single region (e.g., "A man in a pink shirt is sitting in the grass") may correspond to multiple words (*man*, *shirt*, *sitting*) simultaneously, as shown in Figure 1 (c). The inherent normalization in cross-modal attention introduces a zero-sum competition mechanism, wherein increasing the weight of one word usually decreases the weights of others. Consequently, this competitive bias makes conventional

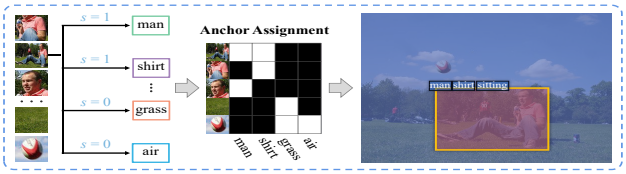
Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



(a) Attention-based alignment method



(b) Saliency-based alignment method



(c) Our MWAN method

Figure 1: Illustration of different fragment alignment methods.

soft-attention suboptimal for dense visual regions containing multiple equally important concepts. **2) The distinguishability bottleneck in retrieval ranking.** Existing models are trained to prioritize dominant signals to maximize the separation between positive and negative pairs. While effective for initial retrieval, this bias becomes a drawback during re-ranking, as top-ranked candidates often share similar semantic information, rendering dominant signals less discriminative.

To address these limitations, we propose an Adaptive **M**ulti-**W**ord **A**nchoring **N**etwork (MWAN). While inspired by standard gated attention, we explicitly tailor this mechanism to cross-modal interactions by designing an **A**nchored **A**lignment **M**echanism (AAM). The AAM introduces a context-aware dynamic gating function that formulates a region-specific threshold modulated simultaneously by local visual features and global textual context, dynamically anchoring each image region to a subset of relevant words. This facilitates a flexible, many-to-many semantic alignment that preserves visual richness and effectively filters out noise. Furthermore, we further introduce a **R**esidual **S**emantic **Re**-ranking (RSR) module to explicitly isolate contextual details that are typically suppressed by dominant salient signals. By leveraging these residual cues as a discriminative "tie-breaker", the RSR module significantly refines the retrieval ranking among semantically similar candidates. We summarize our key contributions as follows:

- We propose an Adaptive Multi-Word Anchoring Network to address the semantic density imbalance in image-text matching. We formulate the Anchored Alignment Mechanism as an adaptive gating approach that achieves context-aware many-to-many semantic alignment, which effectively mitigates semantic loss and filters out noise.
- We design a Residual Semantic Re-ranking module to address the distinguishability bottleneck. By explicitly exploiting residual semantic features, the re-ranking module effectively discriminates among highly similar hard negative candidates.
- Experimental results on Flickr30K and MSCOCO demonstrate the superiority of MWAN. Furthermore, as a universal plug-and-play refinement mechanism, the RSR module can effectively enhance diverse existing models.

2 Related Work

2.1 Attention-based Fragment Alignment

Existing attention-based fragment alignment methods directly learn the semantic alignments between image regions and text words. SCAN (Lee et al., 2018) is the representative work in this direction which has attracted great attention. It introduces cross-attention mechanism to concentrate on significant alignments aiming to minimize the misalignments. Unlike methods devoted to irrelevant alignments removal, Xiong et al. (Xiong et al., 2024) propose a Reference-aware Adaptive Network that jointly leverages a reference attention module to suppress noise and an adaptive aggregation

module to consolidate features, thereby enabling more precise image-text alignment. To address the inherent variations in information density between modalities, Liu et al. (Liu et al., 2025a) propose an Asymmetric Visual Semantic Embedding framework. This approach dynamically selects multi-view image features tailored to different textual inputs and achieves efficient alignment through the optimal matching of segmented meta-semantic embeddings. To mitigate the cross-modal semantic gap and the issue of redundant alignment, Tang et al. (Tang et al., 2026) propose a Fragments Embedding and Aligning Network. This method simultaneously captures global representations and local fragment alignments, incorporating a position weight feature reinforcement module to enrich visual contextual information. Furthermore, it employs a similarity pooling strategy to effectively aggregate multi-level similarity scores, thereby avoiding redundant cross-modal associations. However, due to the intrinsic property of the cross-attention mechanism, the above methods are still affected by redundant or irrelevant alignments, thereby degenerating retrieval accuracy.

2.2 Saliency-based Fragment Alignment

Different from attention-based fragment alignment methods, saliency-based fragment alignment methods learn semantic alignments between salient regions and salient words, and use these alignments to measure the overall semantic association between images and texts. Pan et al. (Pan et al., 2023) propose a simple cross-modal hard aligning network from an information encoding perspective, which can accurately discover the shared semantics of image and text by mining the informative region-word pairs and rejecting the redundant or irrelevant alignments. Li et al. (Li et al., 2024b) propose bidirectional consistency networks for cross-modal alignment, which achieves more accurate cross-modal semantic alignments by imposing explicit consistency constraints in both directions. Similarly, Ji and Liu (Ji & Liu, 2026) propose a unified framework, which employs text-guided contextual reasoning on a spatial graph to explicitly select spatially coherent and semantically relevant structural regions. Furthermore, Liu et al. (Liu et al., 2026) propose a fine-grained alignment approach, which incorporates significance-aware and granularity-aware modeling to explicitly identify salient modality-specific features and suppress redundant information. However, these methods overlook the complex alignment relationships between words and regions, which inevitably results in the loss of semantic information in visual regions. Moreover, they rely heavily on salient regions and keywords, while neglecting the inherent semantic density imbalance between modalities.

3 Method

In this section, we first introduce the Anchored Alignment Mechanism (AAM) (Sec. 3.1). Building upon AAM, we detail the complete Adaptive Multi-Word Anchoring Network (MWAN) pipeline, including the visual representation branch, the textual representation branch, anchor-guided image-text matching (Sec. 3.2), as shown in Figure 2. Finally, we present the Residual Semantic Reranking (RSR) module, which further refines image-text matching results by exploiting residual cues (Sec. 3.3).

3.1 Anchored Alignment Mechanism

We address the inherent semantic density imbalance between visual regions and textual words by reformulating cross-modal alignment as an information-theoretic capacity allocation problem via the AAM. Formally, let $T = \{t_i \in \mathbb{R}^d\}_{i=1}^{L}$ denote the set of textual features, where each t_i represents the embedding of the i-th word in a sentence of length L. Similarly, let $V = \{v_j \in \mathbb{R}^d\}_{j=1}^{K}$ denote the set of visual features, where each v_j corresponds to the representation of the j-th image region, and K is the total number of regions. Both textual and visual features are embedded into a common d-dimensional semantic space. The similarity score $s_{i,j}$ between word t_i and region v_j is computed as $s_{i,j} = \mathcal{S}(t_i, v_j) = \frac{t_i^\top v_j}{\|t_i\| \cdot \|v_j\|}$, where $\mathcal{S}$ denotes the cosine similarity function.

To enable flexible many-to-many semantic alignment and filter out cross-modal noise, we conceptualize cross-modal matching as a constrained variational information allocation problem. Let $m_j = [m_{1,j}, \ldots, m_{L,j}]^\top \in (0,1)^L$ denote the continuous assignment distribution for a given visual region v_j over all textual words. We define an energy functional $\mathcal{E}(m_j)$ that seeks to maximize the expected alignment reward while being penalized by the inherent *Regional Semantic Capacity* of v_j,

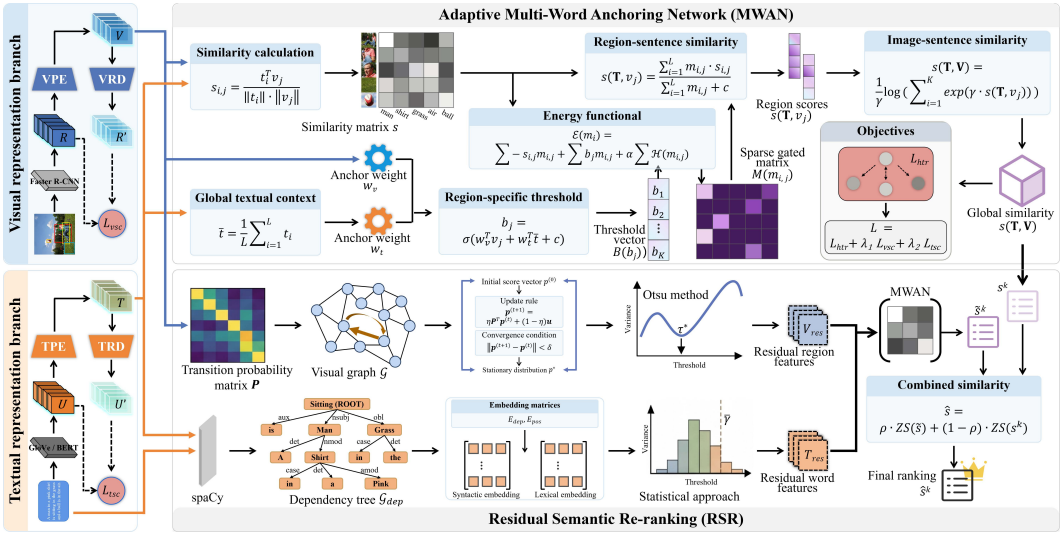


Figure 2: An overview of the proposed MWAN framework.

regularized by binary information entropy (Appendix C.1):

$$\mathcal{E}(m_j) = \sum_{i=1}^{L} \underbrace{-s_{i,j} m_{i,j}}_{\text{Alignment Reward}} + \underbrace{b_j m_{i,j}}_{\text{Capacity Penalty}} + \alpha \underbrace{\mathcal{H}(m_{i,j})}_{\text{Binary Entropy}}, \tag{1}$$

where $\mathcal{H}(m_{i,j}) = m_{i,j} \log m_{i,j} + (1 - m_{i,j}) \log(1 - m_{i,j})$ ensures the smoothness of the allocation, scaled by a temperature coefficient α. Crucially, b_j acts as the *Semantic Capacity Lagrange Multiplier*. By setting the partial derivative $\frac{\partial \mathcal{E}}{\partial m_{i,j}}$ to zero, we derive the analytical closed-form solution:

$$\frac{\partial \mathcal{E}}{\partial m_{i,j}} = 0 \implies m_{i,j} = \frac{1}{1 + \exp\left(-\frac{s_{i,j} - b_j}{\alpha}\right)}. \tag{2}$$

To capture the highly variable semantic complexity of different image regions, we dynamically modulate this context-aware Lagrange multiplier b_j using both the local visual feature v_j and the global textual context $\bar{t}$ (Appendix C.2):

$$b_j = \sigma(w_v^\top v_j + w_t^\top \bar{t} + c), \quad \bar{t} = \frac{1}{L} \sum_{i=1}^{L} t_i, \tag{3}$$

where w_v, w_t are the learnable anchor weight vectors, and c is a scalar bias.

The similarity $s(\mathcal{T}, v_j)$ between the visual region v_j and the entire sentence $\mathcal{T}$ is computed via an anchor-guided weighted aggregation:

$$s(\mathcal{T}, v_j) = \frac{\sum_{i=1}^{L} m_{i,j}\ s_{i,j}}{\sum_{i=1}^{L} m_{i,j} + \varepsilon}, \tag{4}$$

where $\varepsilon > 0$ is a small constant. Consequently, each visual region can adaptively anchor multiple semantically relevant words simultaneously, thereby preserving its intrinsic semantic richness.

3.2 Adaptive Multi-Word Anchoring Network

3.2.1 Visual Representation Branch

For each input image $\mathcal{V}$, we first employ a pre-trained Faster R-CNN to extract the top-K raw region representations $R = \{r_j \in \mathbb{R}^M\}_{j=1}^{K}$. These raw features are then mapped into a semantic embedding space by the Visual Projection Encoder (*VPE*) h_v, which comprises a Fully Connected (FC) layer followed by three self-attention layers:

$$V = h_v(R; \Theta_{h_v}), \tag{5}$$

where Θ_{h_v} denotes the learnable parameters of h_v. To ensure the projected features retain the intrinsic manifold structure of the raw regions, we introduce a Visual Reconstruction Decoder (*VRD*) g_v, consisting of a self-attention layer and two FC layers, to recover the raw features:

$$R' = g_v(V; \Theta_{g_v}), \tag{6}$$

where Θ_{g_v} represents the parameters of g_v. The *VPE* model h_v and *VRD* model g_v are optimized jointly using the **v**isual **s**emantic **c**onsistency loss $\mathcal{L}_{vsc}$ (Appendix C.3):

$$\mathcal{L}_{vsc} = -\mathbb{E}_{r_i \in R}\left[\log \frac{\exp(\mathcal{S}(r_i, r_i')/\tau)}{\sum_{r_j' \in R'} \exp(\mathcal{S}(r_i, r_j')/\tau)}\right]. \tag{7}$$

Here, $\mathcal{L}_{vsc}$ maximizes the mutual information between the projected feature r_i' and the corresponding raw visual feature r_i, while separating it from irrelevant regions. This constraint effectively preserves the core semantic identity of each visual region, preventing semantic drift and representation collapse.

3.2.2 Textual Representation Branch

Following existing works (Wei et al., 2025; Wang et al., 2025a), we investigate two configurations for text representation. For the BiGRU-based formulation, each sentence $\mathcal{T}$ is tokenized into words embedded by pre-trained GloVe vectors, yielding a set of word embeddings $U = \{u_i \in \mathbb{R}^N\}_{i=1}^L$. For the BERT-based formulation, we obtain word-level vectors U from the last layer of a pre-trained BERT. We then map these raw word embeddings into a shared semantic subspace using the Text Projection Encoder (*TPE*) h_t with parameters Θ_{h_t}:

$$T = h_t(U; \Theta_{h_t}), \tag{8}$$

where the architecture of h_t is tailored to the input features. Specifically, for the BiGRU-based formulation, h_t is instantiated as two stacked BiGRU layers to capture sequential dependencies; for the BERT-based formulation, h_t is implemented as an FC layer followed by three self-attention layers to refine contextual semantics.

To preserve the intrinsic information of the raw embeddings U, we employ a Text Reconstruction Decoder (*TRD*) g_t, which consists of a self-attention layer and two FC layers, to perform the inverse mapping:

$$U' = g_t(T; \Theta_{g_t}), \tag{9}$$

where Θ_{g_t} represents the parameters of g_t. The *TPE* model h_t and *TRD* model g_t are optimized jointly via the **t**ext **s**emantic **c**onsistency loss $\mathcal{L}_{tsc}$ (Appendix C.3):

$$\mathcal{L}_{tsc} = -\mathbb{E}_{u_i \in U}\left[\log \frac{\exp(\mathcal{S}(u_i, u_i')/\tau)}{\sum_{u_j' \in U'} \exp(\mathcal{S}(u_i, u_j')/\tau)}\right]. \tag{10}$$

Similar to the visual branch, $\mathcal{L}_{tsc}$ maximizes the mutual information between the projected and raw representations, preventing semantic drift and representation collapse.

3.2.3 Anchor-guided Image-text Matching

Given the projected textual features $T = \{t_i\}_{i=1}^L$ and visual features $V = \{v_j\}_{j=1}^K$, we first apply ℓ_2-normalization to all feature vectors. We then compute the region-sentence similarity $s(\mathcal{T}, v_j)$ for each visual region v_j (Eq. 4). Finally, to derive the global image-text similarity $s(\mathcal{T}, \mathcal{V})$, we aggregate these region-sentence similarities using Log-Sum-Exp (LSE) pooling:

$$s(\mathcal{T}, \mathcal{V}) = \frac{1}{\gamma} \log\left(\sum_{j=1}^K \exp(\gamma \cdot s(\mathcal{T}, v_j))\right), \tag{11}$$

where γ is a scaling factor that magnifies the contribution of the most relevant region-sentence pairs.

3.2.4 Objective Function

Following existing approaches (Zhang et al., 2025; Ji & Liu, 2026), we employ the **h**inge-based **t**riplet **r**anking loss $\mathcal{L}_{htr}$ with online hard negative mining as the primary objective. This loss ensures that matched image-sentence pairs exhibit higher similarity than others. The $\mathcal{L}_{htr}$ is written as:

$$\mathcal{L}_{htr} = \mathbb{E}_{(\mathcal{T},\mathcal{V})\sim\mathcal{D}}[\beta + s(\mathcal{T}, \hat{\mathcal{V}}) - s(\mathcal{T}, \mathcal{V})]_+ + [\beta + s(\hat{\mathcal{T}}, \mathcal{V}) - s(\mathcal{T}, \mathcal{V})]_+, \tag{12}$$

where β is the margin parameter. $(\mathcal{T}, \mathcal{V})$ is a matched image-sentence pair in dataset $\mathcal{D}$ and $[x]_+ \equiv \max(x, 0)$. $\hat{\mathcal{V}} = \arg\max_{\tilde{\mathcal{V}} \neq \mathcal{V}} s(\mathcal{T}, \tilde{\mathcal{V}})$ and $\hat{\mathcal{T}} = \arg\max_{\tilde{\mathcal{T}} \neq \mathcal{T}} s(\tilde{\mathcal{T}}, \mathcal{V})$ denote the hardest image and the hardest sentence within mini-batch, respectively. Therefore, the total objective $\mathcal{L}$ is defined as:

$$\mathcal{L} = \mathcal{L}_{htr} + \lambda_1 \mathcal{L}_{vsc} + \lambda_2 \mathcal{L}_{tsc}, \tag{13}$$

where λ_1 and λ_2 are trade-off factors. By optimizing the loss $\mathcal{L}$, each region can adaptively select multiple semantically relevant words, thereby enabling robust many-to-many semantic alignment and improving image-text matching.

3.3 Residual Semantic Re-ranking Method

Taking the sub-task of image retrieval as an example, given a text query $\mathcal{T}$ and a gallery of C images, MWAN computes an initial similarity vector $s \in \mathbb{R}^C$. By sorting the values of s in descending order, the model identifies the top-k candidates, denoted as $s^k \in \mathbb{R}^k$. To refine this ranking, we compute an additional residual similarity vector $\tilde{s}^k \in \mathbb{R}^k$ derived from the residual feature fragments. Finally, the two similarity vectors are combined using a balancing factor ρ (Appendix C.4):

$$\hat{s}^k = \rho \cdot \text{ZS}(\tilde{s}^k) + (1 - \rho) \cdot \text{ZS}(s^k), \tag{14}$$

where $\text{ZS}(\cdot)$ denotes Z-Score normalization and $\hat{s}^k$ is the new similarity vector used to re-rank the top-k images.

To compute the residual similarity $\tilde{s}^k$, we first extract the full textual features $T = \{t_i\}_{i=1}^L$ and visual features $V^q = \{v_j\}_{j=1}^K$ for the q-th retrieved image (Sec. 3.2). Instead of learning saliency from scratch, we model region relationships as a graph and utilize PageRank centrality to identify objects that are topologically central to the scene. This captures the universal geometric property that semantically important objects tend to interact with many other regions, providing a robust geometric prior (Appendix A). Similarly, our textual module respects the hierarchical nature of language. Guided by dependency grammar, we prioritize syntactic heads over their modifiers. This introduces a strong linguistic inductive bias, compelling the model to focus on the stable *semantic skeleton* of the text (Appendix B). After identifying the salient regions and words, we explicitly exclude these dominant fragments to isolate the residual feature subsets, denoted as $T_{res} \subset T$ and $V^q_{res} \subset V^q$. The q-th element $\tilde{s}^k_q$ of the residual similarity $\tilde{s}^k$ is calculated as:

$$\tilde{s}^k_q = s(T_{res}, V^q_{res}), \quad q = 1, \ldots, k. \tag{15}$$

Note that here we overload the notation $s(\cdot, \cdot)$ to represent the similarity computed between the specified feature subsets rather than the raw inputs.

4 Experiments

4.1 Datasets and Metrics

We evaluate the performance of MWAN on two standard datasets: Flickr30K and MSCOCO. The commonly used evaluation criterions are "R@1", "R@5" and "R@10", i.e., recall rates at the top-1, 5 and 10 results. Following existing works (Zhang et al., 2025; Yin et al., 2026), we use an additional criterion of "Rs" by summing all the recall rates to evaluate the overall performance. Experimental details are provided in Appendix D.

4.2 Comparisons with State-of-the-art Methods

We compare MWAN with recent state-of-the-art methods on Flickr30K and MSCOCO datasets, and report single-model performance. For a fairer comparison, we divide the methods according to their feature extraction backbones. As shown in Table 1, and Table 6, MWAN consistently achieves superior performance on both benchmarks, and the introduction of the RSR module yields further stable improvements across backbones. In particular, MWAN with RSR attains the highest Recall Sum (Rs) of 552.4 on Flickr30K and 549.7 on MSCOCO (1K) under the *Swin+BERT* setting.

These gains can be explained by the complementary effects of the AAM and the RSR module. AAM models fine-grained many-to-many region-word correspondences and suppresses redundant

Table 1: Comparison with state-of-the-art methods on the Flickr30K and MSCOCO 1K test sets. **Bold** denotes the best result, and underline means the second-best result for each metric.

Method	Venue	Flickr30K dataset							MSCOCO (1K) dataset						
		Image Retrieval			Image Annotation			Rs	Image Retrieval			Image Annotation			Rs
		R@1	R@5	R@10	R@1	R@5	R@10		R@1	R@5	R@10	R@1	R@5	R@10	
BUTD + BiGRU															
CHAN (Pan et al., 2023)	CVPR-23	60.2	85.3	90.7	79.7	94.5	97.3	507.8	63.8	90.4	95.8	79.7	96.7	98.7	525.0
TVRN (Pang et al., 2024)	TMM-24	62.3	85.3	90.6	79.7	95.4	97.6	510.9	64.2	90.7	96.1	79.7	96.0	98.6	525.3
IMEB (Li et al., 2024a)	TCSVT-24	60.0	85.9	91.5	80.0	96.0	98.1	511.5	64.0	90.8	95.9	81.0	96.6	98.8	527.1
BOOM (Li et al., 2024b)	TCSVT-24	61.5	86.0	91.4	82.1	96.4	98.3	515.7	64.5	91.1	96.3	80.5	96.7	99.0	528.1
DVSE (Wei et al., 2025)	TMM-25	58.7	84.4	90.7	78.9	94.3	96.9	503.9	63.3	90.3	95.6	79.5	95.9	98.4	523.0
MAMET (Wang et al., 2025a)	TCSVT-25	58.3	83.8	90.0	78.8	95.5	97.9	504.3	63.0	90.7	96.3	79.0	96.4	98.8	524.2
GADNet (Pu et al., 2025)	PR-25	60.2	85.5	90.4	80.1	96.0	97.9	510.1	64.4	90.8	96.0	80.6	96.4	98.4	526.5
SSCT (Chen et al., 2025c)	TMM-25	62.5	86.5	91.5	82.3	96.0	98.6	517.5	66.4	91.5	96.5	81.9	97.1	98.8	532.1
MWAN		62.9	87.1	**92.4**	82.4	**96.9**	**98.8**	520.5	66.7	**92.2**	**96.7**	81.9	**97.3**	**99.1**	533.9
MWAN + RSR		**63.1**	**87.2**	**92.4**	**82.6**	**96.9**	**98.8**	**521.0**	**66.9**	**92.2**	**96.7**	**82.1**	**97.3**	**99.1**	**534.3**
BUTD + BERT															
CHAN (Pan et al., 2023)	CVPR-23	63.9	87.5	92.6	80.6	96.1	97.8	518.5	66.5	92.1	96.7	81.4	96.9	98.9	532.6
TVRN (Pang et al., 2024)	TMM-24	63.9	87.6	92.6	82.1	95.6	98.3	520.1	67.7	92.3	97.1	81.1	96.4	98.8	533.4
IMEB (Li et al., 2024a)	TCSVT-24	64.0	88.0	92.8	84.2	96.7	98.4	524.1	66.7	91.9	96.6	82.4	96.9	99.0	533.5
BOOM (Li et al., 2024b)	TCSVT-24	65.2	88.6	93.1	84.3	96.5	98.5	526.2	67.7	92.6	96.8	82.3	97.2	99.1	535.7
DVSE (Wei et al., 2025)	TMM-25	63.4	87.6	92.3	82.1	96.4	98.3	520.1	65.8	91.3	96.1	80.6	96.8	98.7	529.2
MAMET (Wang et al., 2025a)	TCSVT-25	63.2	87.1	92.2	83.1	97.0	98.6	521.2	67.0	92.2	96.6	82.1	96.7	99.0	533.6
CIMN (Ke et al., 2025)	PR-25	64.8	88.1	92.8	84.4	96.1	98.2	524.4	67.4	92.2	96.6	82.2	97.0	99.0	534.4
PASE (Li et al., 2025)	ESWA-25	63.8	87.9	92.6	**85.8**	96.3	**98.8**	525.2	67.2	91.9	96.5	83.0	97.2	99.0	534.8
FEAN (Tang et al., 2026)	EAAI-26	62.9	86.6	92.1	81.7	95.4	97.9	516.6	65.7	91.6	96.6	81.4	97.0	99.1	531.4
MWAN		65.6	88.9	93.3	85.7	**97.2**	**98.8**	529.5	68.2	**92.9**	**97.2**	83.2	**97.4**	**99.2**	538.1
MWAN + RSR		**65.8**	**89.0**	**93.4**	**85.8**	**97.2**	**98.8**	**530.0**	**68.4**	**92.9**	**97.2**	**83.3**	**97.4**	**99.2**	**538.4**
ViT + BERT															
CHAN (Pan et al., 2023)	CVPR-23	58.4	84.9	90.6	69.2	91.8	95.0	489.9	65.0	91.0	96.0	77.1	95.1	98.1	522.2
LAPS (Fu et al., 2024)	CVPR-24	62.5	87.3	92.7	74.0	93.4	97.4	507.3	66.2	91.3	96.2	78.7	95.5	98.3	526.3
AVSE (Liu et al., 2025a)	AAAI-25	62.7	88.4	93.1	76.0	94.6	97.5	512.3	67.0	91.5	96.3	79.8	95.6	98.3	528.5
MG-Net (Ji & Liu, 2026)	AAAI-26	63.4	87.5	92.7	78.9	94.8	97.7	515.0	66.7	90.8	95.6	80.5	96.1	98.4	528.1
GRM (Liu et al., 2026)	AAAI-26	64.6	88.5	93.5	77.2	94.5	**97.9**	516.2	67.6	91.8	96.7	80.8	96.7	98.9	532.5
MWAN		65.4	88.8	**93.7**	79.2	**94.9**	**97.9**	519.9	68.3	**92.2**	**96.9**	81.3	**96.9**	**99.0**	534.6
MWAN + RSR		**65.8**	**88.9**	**93.7**	**79.4**	**94.9**	**97.9**	**520.6**	**68.7**	**92.2**	**96.9**	**81.6**	**96.9**	**99.0**	**535.3**
Swin + BERT															
CHAN (Pan et al., 2023)	CVPR-23	68.5	90.6	94.5	81.4	97.0	98.6	530.6	70.6	93.7	97.6	81.6	97.2	99.3	539.8
LAPS (Fu et al., 2024)	CVPR-24	70.0	91.7	95.4	82.4	97.4	99.5	536.3	72.1	93.7	97.3	84.0	97.6	99.3	544.1
AVSE (Liu et al., 2025a)	AAAI-25	70.0	92.4	95.6	83.9	97.4	99.4	538.7	72.1	94.0	97.4	84.9	98.0	99.3	545.7
DH-Set (Zhang et al., 2025)	CVPR-25	75.4	93.1	96.2	88.3	98.5	99.6	551.0	73.2	94.0	97.7	86.0	98.0	99.4	548.2
MG-Net (Ji & Liu, 2026)	AAAI-26	72.8	92.6	95.7	87.8	98.6	**99.7**	547.2	73.1	93.3	97.2	84.4	97.9	99.5	545.4
GRM (Liu et al., 2026)	AAAI-26	73.0	92.9	96.2	85.6	**98.7**	99.5	546.0	73.2	93.9	**97.8**	85.2	97.8	**99.6**	547.5
MWAN		75.8	**93.4**	**96.3**	88.5	**98.7**	99.4	552.1	73.5	**94.2**	**97.8**	86.2	**98.1**	99.5	549.3
MWAN + RSR		**76.0**	**93.4**	**96.3**	**88.6**	**98.7**	99.4	**552.4**	**73.7**	**94.2**	**97.8**	**86.4**	**98.1**	99.5	**549.7**

Table 2: The re-ranking results of RSR on the Flickr30K and MSCOCO datasets.

Method	Flickr30K dataset							MSCOCO dataset						
	Image Retrieval			Image Annotation			Rs	Image Retrieval			Image Annotation			Rs
	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 + BERT														
CHAN	58.4	84.9	90.6	69.2	91.8	**95.0**	489.9	65.0	91.0	96.0	77.1	95.1	**98.1**	522.2
CHAN + **RSR**	**59.3**	**85.3**	**90.7**	**70.0**	**92.0**	**95.0**	**492.3**	**65.7**	**91.3**	**96.1**	**77.7**	**95.2**	**98.1**	**524.1**
LAPS	62.5	87.3	92.7	74.0	93.4	**97.4**	507.3	66.2	91.3	96.2	78.7	95.5	**98.3**	526.3
LAPS + **RSR**	**63.3**	**87.6**	**92.8**	**74.7**	**93.5**	**97.4**	**509.3**	**66.9**	**91.6**	**96.3**	**79.3**	**95.6**	**98.3**	**528.0**
GRM	64.6	88.5	93.5	77.2	94.5	**97.9**	516.2	67.6	91.8	**96.7**	80.8	96.7	**98.9**	532.5
GRM + **RSR**	**65.2**	**88.8**	**93.6**	**77.8**	**94.6**	**97.9**	**517.9**	**68.1**	**92.1**	**96.7**	**81.3**	**96.8**	**98.9**	**533.9**
Swin + BERT														
CHAN	68.5	90.6	94.5	81.4	**97.0**	**98.6**	530.6	70.6	93.7	97.6	81.6	**97.2**	**99.3**	539.8
CHAN + **RSR**	**69.2**	**90.8**	**94.6**	**81.7**	**97.0**	**98.6**	**531.9**	**71.1**	**93.9**	**97.7**	**81.8**	**97.2**	**99.3**	**541.0**
LAPS	70.0	91.7	**95.4**	82.4	**97.4**	**99.5**	536.3	72.1	93.7	97.3	84.0	97.6	**99.3**	544.1
LAPS + **RSR**	**70.6**	**91.9**	**95.4**	**82.6**	**97.4**	**99.5**	**537.4**	**72.6**	**93.8**	**97.4**	**84.2**	**97.7**	**99.3**	**545.0**
GRM	73.0	92.9	**96.2**	85.6	**98.7**	**99.5**	546.0	73.2	93.9	**97.8**	85.2	**97.8**	**99.6**	547.5
GRM + **RSR**	**73.4**	**93.0**	**96.2**	**85.8**	**98.7**	**99.5**	**546.6**	**73.5**	**94.0**	**97.8**	**85.4**	**97.8**	**99.6**	**548.1**
Foundation VLMs														
CLIP	65.4	87.2	91.7	85.4	**97.1**	**98.7**	525.5	35.3	60.0	70.1	55.2	78.7	86.7	386.0
CLIP + **RSR**	**66.2**	**87.5**	**91.8**	**85.8**	**97.1**	**98.7**	**527.1**	**36.5**	**60.6**	**70.5**	**55.8**	**79.2**	**86.9**	**389.5**
ALBEF	59.9	84.8	90.5	78.2	95.5	**97.9**	506.9	40.2	68.4	78.9	62.4	85.9	92.1	428.3
ALBEF + **RSR**	**60.9**	**85.1**	**90.8**	**78.8**	**95.6**	**97.9**	**509.1**	**41.4**	**68.8**	**79.2**	**63.0**	**86.2**	**92.3**	**430.9**

cross-modal noise, allowing each visual region to be represented by multiple semantically relevant words. This design alleviates both over-alignment caused by irrelevant associations and semantic under-matching induced by rigid hard alignment, thereby preserving the multi-faceted semantic nature of visual content and yielding more accurate similarity estimation. Moreover, the RSR mod-

ule further enhances retrieval by exploiting residual semantic cues that are often overshadowed by dominant salient fragments. The consistent improvements observed across different backbones and evaluation metrics demonstrate the effectiveness of the proposed framework. Furthermore, MWAN achieves these performance gains while offering a highly favorable efficiency-accuracy trade-off (Appendix E.8).

4.3 Compatibility with Existing Methods

To verify the general applicability of the RSR module, we apply it as a post hoc re-ranking stage to a diverse set of image-text matching models, including *BUTD+BiGRU*, *BUTD+BERT*, *ViT+BERT*, and *Swin+BERT* architectures, as well as pretrained foundation Vision-Language Models (VLMs). The original and re-ranked results are reported in Table 2 and Table 7. While RSR provides steady but marginal gains for our proposed MWAN, it yields substantial improvements for foundation VLMs. This discrepancy is highly intuitive. MWAN's AAM already establishes high-quality many-to-many fine-grained alignments, leaving a very tight residual margin. In contrast, standard VLMs optimized via global contrastive learning systematically suffer from the distinguishability bottleneck, as they aggressively discard fine-grained residual semantics in favor of dominant global features. By explicitly exploiting these suppressed residual cues, RSR mitigates subtle semantic ambiguities among hard negatives. This demonstrates that RSR's true potential lies not merely as a component of MWAN, but as a universal, model-agnostic refinement plugin that efficiently enhances ranking precision across architectures.

4.4 Ablation Study

Additional ablation results are provided in Appendix E.

4.4.1 Loss Ablation Analysis

To evaluate the contribution of each loss component, we conduct an ablation study on the consistency loss weights λ_1 and λ_2, as summarized in Table 3. The results indicate that MWAN optimized solely with the hinge-based triplet ranking loss $\mathcal{L}_{htr}$ yields the weakest performance across all metrics on both Flickr30K and MSCOCO. Introducing either the visual semantic consistency loss $\mathcal{L}_{vsc}$ or the textual semantic consistency loss $\mathcal{L}_{tsc}$ leads to clear and consistent improvements. Jointly optimizing all three losses achieves the best overall performance with a peak Rs of 520.5 on Flickr30K. These results indicate that $\mathcal{L}_{vsc}$ and $\mathcal{L}_{tsc}$ serve as essential constraints for semantic mutual information maximization. By preserving the unique semantic identity of each fragment, these losses effectively mitigate semantic collapse during projection, yielding aligned representations that support more stable and fine-grained cross-modal alignment.

Table 3: Ablation study of the loss function components on the Flickr30K and MSCOCO datasets.

$\mathcal{L}_{htr}$	$\mathcal{L}_{vsc}$	$\mathcal{L}_{tsc}$	Flickr30K dataset							MSCOCO dataset						
			Image Retrieval			Image Annotation			Rs	Image Retrieval			Image Annotation			Rs
			R@1	R@5	R@10	R@1	R@5	R@10		R@1	R@5	R@10	R@1	R@5	R@10	
√			60.6	85.4	90.7	79.6	94.8	96.4	507.5	63.6	90.0	94.5	79.2	95.0	96.9	519.2
√	√		62.2	86.6	92.1	82.0	96.6	98.7	518.2	65.5	91.9	96.4	81.4	97.1	98.8	531.1
√		√	61.8	86.4	92.1	81.7	96.4	98.6	517.0	65.3	91.7	96.3	81.3	97.0	98.8	530.4
√	√	√	62.9	87.1	92.4	82.4	96.9	98.8	520.5	66.7	92.2	96.7	81.9	97.3	99.1	533.9

4.4.2 The Re-ranking Fusion Weight ρ

The hyperparameter ρ controls the relative contribution of the primary similarity s^k and the residual similarity $\tilde{s}^k$ in the re-ranking stage, thus playing a critical role in balancing dominant semantic alignment and residual semantic refinement. As shown in Table 4, performance on both Flickr30K and MSCOCO improves steadily as ρ increases from 0 to 0.15, reaching the highest Rs of 521.0 and 534.3, respectively. This trend demonstrates that residual information effectively complements primary matching signals, serving as subtle and discriminative cues to resolve ambiguities among top-ranked candidates. However, when ρ is further increased to 0.2, performance slightly degrades, suggesting that overemphasizing residual fragments can dilute the salient semantic context and introduce undesirable noise. Ultimately, these results confirm that residual cues are most beneficial

when deployed as a calibrated tie-breaker rather than a replacement for the primary similarity, with a moderate ρ achieving the optimal integration of salient and residual semantics.

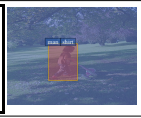
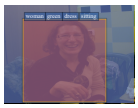
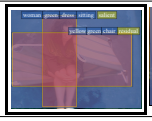
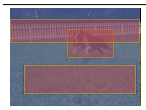
Table 4: Sensitivity analysis of the re-ranking fusion weight ρ on both datasets.

ρ	Flickr30K dataset							MSCOCO dataset						
	Image Retrieval			Image Annotation			Rs	Image Retrieval			Image Annotation			Rs
	R@1	R@5	R@10	R@1	R@5	R@10		R@1	R@5	R@10	R@1	R@5	R@10	
0	62.9	87.1	92.4	82.4	96.9	98.8	520.5	66.7	92.2	96.7	81.9	97.3	99.1	533.9
0.1	63.0	87.2	92.4	82.5	96.9	98.8	520.8	66.7	92.2	96.7	82.1	97.3	99.1	534.1
0.15	63.1	87.2	92.4	82.6	96.9	98.8	521.0	66.9	92.2	96.7	82.1	97.3	99.1	534.3
0.2	62.9	87.0	92.6	82.4	97.1	98.9	520.9	66.7	92.3	96.8	81.8	97.4	99.1	534.1

4.5 Robustness on Hard Negatives

The visualization of retrieval results on hard negative cases, shown in Table 5, further demonstrates the critical role of the RSR module in refining cross-modal matching. The MWAN occasionally ranks visually similar yet semantically mismatched images at top positions due to an over-reliance on dominant salient cues, such as the generic presence of a *man* or *woman*. In contrast, RSR promotes the ground-truth image to the top-1 position by exploiting informative residual cues. Specifically, RSR encourages the model to discriminate among competing candidates by attending to subtle but decisive details, such as the presence of *air* in the first example or the *yellow* chair attributes in the second, which are typically suppressed by dominant signals in standard alignment paradigms. These qualitative results confirm that the RSR module effectively leverages residual semantic cues as a robust 'tie-breaker', enabling more precise semantic matching.

Table 5: Examples of the top-3 images/texts retrieved by MWAN for given text/image queries. Groundtruth matched images are marked by black bounding boxes, which can be re-ranked higher by the proposed RSR. Similarly, the groundtruth text is annotated using the symbol "**(GT)**".

Query text/image	Returned top-3 images/texts by MWAN			Reranked top-3 images/texts by MWAN+RSR		
A man in a pink shirt is sitting in the grass and a ball is in the air.						
A woman in a green dress sitting in a broken green and yellow chair.						
	Four **white dogs** are **running** and jumping with masks over their **mouth**.	A **dog runs** on the **green grass** near a wooden fence. **(GT)**	Two **dogs** jump for a ball on a **green field**.	A **dog runs** on the **green grass** near a **wooden fence**. **(GT)**	Four **white dogs** are **running** and **jumping** with masks over their **mouth**.	Two **dogs jump** for a ball on a **green field**.
	A **black man** leans over as a **white man jumps** on his back backwards.	Two football **players leap** into the air as a **player** on the opposing team **moves** toward them. **(GT)**	Two American football **players** in blue are tackling a football **player** in **red**.	Two **football players leap** into the **air** as a **player** on the opposing **team moves** toward them. **(GT)**	A **black man leans** over as a **white man jumps** on his back backwards.	Two American **football players** in **blue** are tackling a **football player** in **red**.

5 Conclusion

In this paper, we propose the Adaptive Multi-Word Anchoring Network to overcome the inherent semantic density imbalance in fine-grained image-text matching. Specifically, we design the Anchored Alignment Mechanism to enable adaptive many-to-many semantic alignment, effectively preserving the intrinsic semantic richness of visual regions. Furthermore, we introduce the Residual Semantic Re-ranking module to address the distinguishability bottleneck by exploiting residual cues for rank refinement. Extensive experiments on the Flickr30K and MSCOCO datasets demonstrate the superiority of our proposed approach in image-text matching.

AI USE STATEMENT

We use LLMs solely to assist in checking grammatical correctness. After the initial check by the model, we further refine and correct any remaining grammatical issues manually. Therefore, the role of LLMs in this work is limited.

REFERENCES

Da Chen, Yangtao Wang, Yanzhao Xie, Siyuan Chen, Weilong Peng, Maobin Tang, Meie Fang, CL Philip Chen, Ping Li, and Wensheng Zhang. Intra-modal consistency for image-text retrieval through soft-label distillation. *Pattern Recognition*, pp. 112817, 2025a.

Junyu Chen, Yihua Gao, Mingyuan Ge, and Mingyong Li. Ambiguity-aware and high-order relation learning for multi-grained image–text matching. *Knowledge-Based Systems*, 316:113355, 2025b.

Ran Chen, Taiyi Su, Hanli Wang, and Zhangkai Ni. Similarity shuffled criss-cross transformer with angle loss for image-text matching. *IEEE Transactions on Multimedia*, 2025c.

Zihao Chen, Chenyang Bu, Shengwei Ji, and Xindong Wu. Dual-branch multi-granularity network with structured contrastive ranking for cross-modal retrieval. In *Proceedings of the ACM Web Conference 2026*, pp. 1959–1970, 2026.

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

Yuyang Hong, Jiaqi Gu, Qi Yang, Lubin Fan, Yue Wu, Ying Wang, Kun Ding, Shiming Xiang, and Jieping Ye. Knowledge-based visual question answer with multimodal processing, retrieval and filtering. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. URL https://openreview.net/forum?id=h0LzGQq6uO.

Hang Hua, Qing Liu, Lingzhi Zhang, Jing Shi, Soo Ye Kim, Zhifei Zhang, Yilin Wang, Jianming Zhang, Zhe Lin, and Jiebo Luo. Finecaption: Compositional image captioning focusing on wherever you want at any granularity. In *Proceedings of the computer vision and pattern recognition conference*, pp. 24763–24773, 2025.

Yongzhen Huang, Zifeng Wu, Liang Wang, and Tieniu Tan. Feature coding in image classification: A comprehensive study. *IEEE transactions on pattern analysis and machine intelligence*, 36(3): 493–506, 2013.

Zhiqi Huang, Huanjia Zhu, Xiangwen Deng, Zhong Qinghao, and Bingzhi Chen. Bayesvqa: Energy-guided bayesian debiasing for language-bias-robust visual question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 5221–5229, 2026.

Linlin Ji and Li Liu. Multi-view differential mixing and graph-guided structural region selection for cross-modal alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 22200–22208, 2026.

Xiao Ke, Baitao Chen, Xiong Yang, Yuhang Cai, Hao Liu, and Wenzhong Guo. Cross-modal independent matching network for image-text retrieval. *Pattern Recognition*, 159:111096, 2025.

Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 201–216, 2018.

Mingyong Li, Yihua Gao, Honggang Zhao, Ruiheng Li, and Junyu Chen. Progressive semantic aggregation and structured cognitive enhancement for image–text matching. *Expert Systems with Applications*, 274:126943, 2025.

Zhe Li, Lei Zhang, Kun Zhang, Yongdong Zhang, and Zhendong Mao. Fast, accurate, and lightweight memory-enhanced embedding learning framework for image-text retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(7):6542–6558, 2024a.

Zhe Li, Lei Zhang, Kun Zhang, Yongdong Zhang, and Zhendong Mao. Improving image-text matching with bidirectional consistency of cross-modal alignment. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(7):6590–6607, 2024b.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pp. 740–755. Springer, 2014.

Jiale Liu, Haoming Zhou, Yishu Liu, Bingzhi Chen, and Yuncheng Jiang. Cross modal fine-grained alignment via granularity-aware and region-uncertain modeling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 23703–23711, 2026.

Yang Liu, Mengyuan Liu, Shudong Huang, and Jiancheng Lv. Asymmetric visual semantic embedding framework for efficient vision-language alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 5676–5684, 2025a.

Yishu Liu, Jiawei Zhu, Congcong Wen, Guangming Lu, Hui Lin, and Bingzhi Chen. Towards robust visual question answering via prompt-driven geometric harmonization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 5721–5729, 2025b.

GuangHao Meng, Jinpeng Wang, Qian-Wei Wang, XuDong Ren, and Dan Zhao. Imagine with layout and sketch: Enhancing vision-language retrieval with dual-stream multi-modal query refinement. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 7972–7980, 2026.

Shan Ning, Longtian Qiu, and Xuming He. Wiki-r1: Incentivizing multimodal reasoning for knowledge-based VQA via data and sampling curriculum. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=TH1Tgbjkm7.

Zhengxin Pan, Fangyu Wu, and Bailing Zhang. Fine-grained image-text matching by cross-modal hard aligning network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19275–19284, 2023.

Shanmin Pang, Yueyang Zeng, Jiawei Zhao, and Jianru Xue. A mutually textual and visual refinement network for image-text matching. *IEEE Transactions on Multimedia*, 26:7555–7566, 2024.

Xiao Pu, Zhiwen Wang, Lin Yuan, Yu Wu, Liping Jing, and Xinbo Gao. Gadnet: Improving image–text matching via graph-based aggregation and disentanglement. *Pattern Recognition*, 157:110900, 2025.

Xianlun Tang, Lin Jiang, Yu Xia, Wuquan Deng, Bo Tang, Tianyu Xiang, and Hao Zhu. Fean: A fragments embedding and aligning network for image-text matching. *Engineering Applications of Artificial Intelligence*, 163:113009, 2026.

Pengzhe Wang, Lei Zhang, Zhendong Mao, Nenan Lyu, and Yongdong Zhang. Matryoshka learning with metric transfer for image-text matching. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025a.

Yaxiong Wang, Lianwei Wu, Lechao Cheng, Zhun Zhong, Yujiao Wu, and Meng Wang. Beyond general alignment: Fine-grained entity-centric image-text matching with multimodal attentive experts. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 792–802, 2025b.

Wenzhang Wei, Zhipeng Gui, Changguang Wu, Anqi Zhao, Dehua Peng, and Huayi Wu. Dynamic visual semantic sub-embeddings and fast re-ranking for image-text retrieval. *IEEE Transactions on Multimedia*, 27:3781–3796, 2025.

Guoxin Xiong, Meng Meng, Tianzhu Zhang, Dongming Zhang, and Yongdong Zhang. Reference-aware adaptive network for image-text matching. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(10):9678–9691, 2024.

Laiguo Yin, Yixin Zhang, YUQING SUN, and Lizhen Cui. Multimodal aligned semantic knowledge for unpaired image-text matching. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=d3CISVVO6v.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 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.

Jusheng Zhang, Kaitong Cai, Jing Yang, Jian Wang, Chengpei Tang, and Keze Wang. Top-down semantic refinement for image captioning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 12591–12599, 2026.

Kun Zhang, Jingyu Li, Zhe Li, and S Kevin Zhou. Dh-set: Improving vision-language alignment with diverse and hybrid set-embeddings learning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 24993–25003, 2025.

A Graph-based Visual Saliency Mining

Since the accuracy of image-text matching is heavily contingent on dominant semantic fragments, identifying these salient visual components is a prerequisite for our residual mechanism. Instead of relying on external object detectors' heuristic confidence scores or cross-modal attention maps, we propose an graph-theoretic approach modulated by spatial priors to mine the intrinsic structural importance of each region. We conceptualize the set of visual features $V = \{v_i\}_{i=1}^{K}$ alongside their spatial coordinates $C = \{c_i\}_{i=1}^{K}$ as nodes in a fully connected undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where the edges represent the spatially-penalized semantic affinity between visual tokens. We posit that semantically central regions, defined as those exhibiting high visual affinity and spatial cohesion with other regions, constitute the visual core of the image.

Visual graph construction. We first construct an affinity matrix $\mathbf{A} \in \mathbb{R}^{K\times K}$ to model the pairwise relationships between visual features. To filter out noise and weak correlations, we apply a rectified cosine similarity measure with a spatial Gaussian decay kernel:

$$\mathbf{A}_{ij} = \max\left(0, \frac{v_i^\top v_j}{\|v_i\| \cdot \|v_j\|}\right) \cdot \exp\left(-\lambda \|c_i - c_j\|_2^2\right), \tag{16}$$

where $c_i \in [0,1]^2$ represents the normalized 2D spatial coordinates of the i-th visual token. To ensure architectural generalization across diverse visual backbones, c_i is explicitly instantiated as follows:

- Region-based backbones: c_i is defined as the normalized geometric center coordinate of the extracted bounding box.
- Patch-based backbones: The output features naturally form a 2D spatial grid (e.g., $H \times W$). We map the 1D sequence index of the i-th patch back to its 2D grid position (x, y), and define $c_i = (\frac{x+0.5}{W}, \frac{y+0.5}{H})$ as its normalized spatial center.

Here, $\lambda > 0$ is a hyperparameter controlling the strength of the spatial penalty. The ReLU-like operation ensures that negative semantic correlations do not contribute to the centrality propagation, while the geometric regularization explicitly suppresses spurious edges between highly similar but spatially disjoint background patches. To establish a valid probabilistic transition matrix for the random walk, we perform row-wise normalization on $\mathbf{A}$:

$$\mathbf{P}_{ij} = \frac{\mathbf{A}_{ij}}{\sum_{k=1}^{K} \mathbf{A}_{ik}}, \tag{17}$$

where $\mathbf{P}_{ij}$ denotes the transition probability from node i to node j. It is noteworthy that we strictly retain the self-loops in the visual graph. Since $\|c_i - c_i\|_2^2 = 0$, the spatial penalty evaluates to 1, yielding $A_{ii} = \max(0, \frac{v_i^\top v_i}{||v_i||\cdot||v_i||}) \cdot 1 = 1$. Consequently, the row sum $\sum_{k=1}^{K} A_{ik}$ is strictly bounded below by 1. Furthermore, this ensures that the transition matrix $\mathbf{P}$ is strictly row-stochastic, guaranteeing the existence and uniqueness of the PageRank stationary distribution.

VisualRank centrality. We employ the PageRank algorithm to compute the stationary distribution of the random walk on graph $\mathcal{G}$. Let $\mathbf{p}^{(t)} \in \mathbb{R}^K$ denote the importance score vector at iteration t. The recursive update rule is formulated as:

$$\mathbf{p}^{(t+1)} = \eta \mathbf{P}^\top \mathbf{p}^{(t)} + (1-\eta)\mathbf{u}, \tag{18}$$

where η is the damping factor controlling the probability of continuing the walk, and $\mathbf{u} = [\frac{1}{K}, \ldots, \frac{1}{K}]^\top$ represents the uniform prior, ensuring that the random surfer can jump to any node with equal probability to avoid sinks. The iteration converges when $\|\mathbf{p}^{(t+1)} - \mathbf{p}^{(t)}\| < \delta$, yielding the final visual saliency map $\mathbf{p}^*$.

Adaptive thresholding via variance maximization. The stationary distribution $\mathbf{p}^*$ provides a continuous measure of importance. To partition the regions into *salient* (V_{sal}) and *residual* (V_{res}) subsets, we employ Otsu's method to determine the optimal threshold τ^* that maximizes the inter-class variance between the two groups. Let $\omega_0(\tau)$ and $\omega_1(\tau)$ be the cumulative probabilities of the residual and salient classes for a given threshold τ, and $\mu_0(\tau), \mu_1(\tau)$ be their respective mean importance scores. The optimization objective is defined as:

$$\tau^* = \arg\max_{\tau} \left(\omega_0(\tau)\omega_1(\tau)[\mu_0(\tau) - \mu_1(\tau)]^2\right). \tag{19}$$

Based on τ^*, we explicitly separate the dominant signals from the fine-grained cues:

$$V_{sal} = \{v_i \mid \mathbf{p}_i^* \geq \tau^*\}, \quad V_{res} = \{v_i \mid \mathbf{p}_i^* < \tau^*\}. \tag{20}$$

This partitioning ensures that V_{res} captures the subtle, contextual details necessary for the second-stage re-ranking.

B Syntax-aware Textual Saliency Mining

While visual saliency is derived from geometric and semantic centrality, textual saliency is intrinsically governed by linguistic structure. According to dependency grammar, a sentence is not merely a linear sequence of words but a hierarchical structure where syntactic heads govern their dependents. This hierarchy reveals the *semantic energy* of each token, wherein words functioning as roots or core arguments typically convey the core semantic content, while modifiers and function words primarily provide contextual support. To preserve this vital linguistic inductive bias, we propose a learnable, syntax-aware dynamic scoring mechanism that explicitly incorporates dependency and part-of-speech (POS) attributes into a continuous representation space.

Dependency tree modeling. Let a sentence $\mathcal{T}$ be represented as a sequence of L tokens $\mathcal{T} = \{w_1, \ldots, w_L\}$. We employ a pre-trained dependency parser to parse $\mathcal{T}$ into a Directed Acyclic Graph (DAG), specifically a tree structure $\mathcal{G}_{dep} = (\mathcal{V}, \mathcal{E})$, where vertices $\mathcal{V}$ correspond to tokens and edges $\mathcal{E}$ represent typed dependency relations (e.g., `nsubj`, `dobj`, `det`).

Learnable syntactic and lexical embeddings. Instead of projecting linguistic tags into a fixed scalar space via heuristic mapping functions, we parameterize these attributes to enable data-driven optimization. Let $\mathcal{D}$ be the vocabulary of all dependency relation types and $\mathcal{P}$ be the vocabulary of POS tags. We introduce two learnable embedding matrices: $E_{dep} \in \mathbb{R}^{|\mathcal{D}| \times d_{syn}}$ and $E_{pos} \in \mathbb{R}^{|\mathcal{P}| \times d_{syn}}$, where d_{syn} is the embedding dimension. For each token w_j, its syntactic and lexical representations are extracted as $e^j_{dep} = E_{dep}(\text{type}(w_j \to \text{head}(w_j)))$ and $e^j_{pos} = E_{pos}(\text{tag}(w_j))$, respectively.

Dynamic saliency scoring. To quantify the structural importance of each token contextually, we design a dynamic scoring network. The syntactic embedding e^j_{dep} and lexical embedding e^j_{pos} are concatenated with the token's contextual semantic feature $t_j \in \mathbb{R}^d$. The composite feature is then fed into a Multi-Layer Perceptron (MLP) followed by a Sigmoid activation to generate a continuous saliency score $s_j \in (0, 1)$:

$$s_j = \sigma(\text{MLP}([t_j \parallel e^j_{pos} \parallel e^j_{dep}])), \tag{21}$$

where $\parallel$ denotes the concatenation operation. This formulation allows the network to adaptively infer the semantic contribution of each token by jointly optimizing linguistic priors and cross-modal alignment objectives.

Adaptive salience partitioning. To dynamically separate salient words from residual ones based on the predicted scores, we employ a statistical thresholding approach. Let μ_s and σ_s denote the mean and standard deviation of the saliency scores $\{s_j\}_{j=1}^{L}$ within the sentence:

$$\mu_s = \frac{1}{L}\sum_{j=1}^{L} s_j, \quad \sigma_s = \sqrt{\frac{1}{L}\sum_{j=1}^{L}(s_j - \mu_s)^2}. \tag{22}$$

We initially define an adaptive threshold $\bar{\gamma} = \mu_s + \kappa \cdot \sigma_s$, where κ controls the selectivity. However, relying purely on global variance is vulnerable to degenerate cases, such as extremely short queries where σ_s diminishes, potentially resulting in an empty residual set and causing numerical instability during similarity computation.

To robustify the partitioning against such distributional bottlenecks, we introduce a cardinality-aware fallback mechanism. We first define a preliminary residual set $T_{res}^{pre} = \{t_j \mid s_j < \bar{\gamma}\}$. If the statistical thresholding yields an empty set ($|T_{res}^{pre}| = 0$), we gracefully fallback to a rank-based percentile approach, forcing the token with the lowest saliency score into the residual subset to guarantee minimum cardinality:

$$T_{res} = \begin{cases} T_{res}^{pre}, & \text{if } |T_{res}^{pre}| \geq 1 \\ \{t_k\}, \text{where } k = \arg\min_j(s_j), & \text{otherwise} \end{cases}. \tag{23}$$

The salient set is then simply defined as the relative complement $T_{sal} = T \setminus T_{res}$. This formulation ensures that T_{sal} retains the skeleton of the sentence's intent, while T_{res} safely and consistently captures the subtle modifiers utilized in our re-ranking module, entirely circumventing the failure modes of small-sample variance estimation.

By parameterizing the dependency and POS attributes, our method fundamentally transitions from rigid hard-coding to a data-driven optimization paradigm. Acting as a strong structural inductive bias, this explicit integration of linguistic embeddings compels the model to focus on the semantic hierarchy without overfitting to dataset-specific biases or suffering from the severe error propagation of external text parsers. Consequently, the model naturally learns to elevate core arguments and leverage residual semantics for image-text matching.

C Theoretical Motivation and Intuition

In this section, we provide the conceptual intuition and information-theoretic motivation behind the architectural designs of our Anchored Alignment Mechanism (AAM).

C.1 Information-Theoretic Perspective on the Energy Functional $\mathcal{E}(m_j)$

We conceptually frame the many-to-many semantic matching between a visual region v_j and a set of textual words $\{t_i\}_{i=1}^{L}$ as a probabilistic allocation process. Let $Z_j = \{z_{1,j}, z_{2,j}, \ldots, z_{L,j}\}$ be a set of unobservable independent binary latent variables, where $z_{i,j} \in \{0, 1\}$ denotes whether the i-th word is anchored to the j-th region. The primary objective is to maximize the expected similarity reward $\sum_{i=1}^{L} s_{i,j} m_{i,j}$, where $m_{i,j} = P(z_{i,j} = 1)$.

A single visual region contains rich but finite semantic information. To prevent unbounded alignment, we introduce a regional semantic capacity constraint C_j, such that $\sum_{i=1}^{L} m_{i,j} \leq C_j$. Using the method of Lagrange multipliers, we incorporate this constraint via a region-specific dual variable $b_j \geq 0$, yielding the constrained utility function:

$$\widetilde{U}(m_j) = \sum_{i=1}^{L} s_{i,j} m_{i,j} - b_j \left(\sum_{i=1}^{L} m_{i,j} - C_j \right). \tag{24}$$

In fine-grained image-text matching, hard deterministic assignments inevitably lead to representation collapse. Following the maximum entropy principle, the optimal probabilistic allocation should maximize the information entropy to preserve structural uncertainty. For independent Bernoulli

variables, the joint Shannon entropy $S(Z_j)$ is:

$$S(Z_j) = \sum_{i=1}^{L} \Big[- m_{i,j} \log m_{i,j} - (1 - m_{i,j}) \log(1 - m_{i,j}) \Big]. \tag{25}$$

Notice that the binary information entropy term defined in our main paper is exactly the negative entropy, i.e., $-S(Z_j) = \sum_{i=1}^{L} \mathcal{H}(m_{i,j})$.

We formulate this heuristically as a variational free energy minimization problem, defining the energy functional $\mathcal{E}$ as the expected energy minus the temperature-scaled entropy:

$$\mathcal{E}(m_j) = -\widetilde{U}(m_j) - \alpha S(Z_j), \tag{26}$$

where $\alpha > 0$ acts as the thermodynamic temperature controlling the exploration-exploitation trade-off. Substituting the terms and dropping the constant $b_j C_j$, we recover the exact functional proposed in our network:

$$\mathcal{E}(m_j) = \sum_{i=1}^{L} -s_{i,j} m_{i,j} + \sum_{i=1}^{L} b_j m_{i,j} + \alpha \sum_{i=1}^{L} \mathcal{H}(m_{i,j}). \tag{27}$$

Discussion and testable predictions: Setting the partial derivative to zero naturally yields the Sigmoid gating function $m_{i,j} = 1/(1 + \exp(-\frac{s_{i,j} - b_j}{\alpha}))$. More importantly, this formulation yields a crucial, testable design constraint. It reveals that the temperature hyperparameter α fundamentally regulates the intrinsic thermodynamic uncertainty of the embedding space. Consequently, α is completely independent of the dataset semantic density.

C.2 Intuition for the Context-aware Dynamic Threshold b_j

To provide intuition for the dynamic threshold b_j, we draw inspiration from Bayesian decision theory. Let $z_{i,j} \in \{0, 1\}$ denote whether the i-th word t_i and the j-th region v_j are semantically aligned. Assuming the L_2-normalized features lie on a unit hypersphere, their distributions can be heuristically modeled using von Mises-Fisher (vMF) distributions.

Conditioned on the region v_j, we assume the aligned words ($z_{i,j} = 1$) and unaligned words ($z_{i,j} = 0$) are distributed around the mean direction v_j with different concentration parameters κ_1 and κ_0 (where $\kappa_1 > \kappa_0 \geq 0$):

$$P(t_i \mid v_j, z_{i,j} = 1) = C_d(\kappa_1) \exp(\kappa_1 s_{i,j}), \tag{28}$$

$$P(t_i \mid v_j, z_{i,j} = 0) = C_d(\kappa_0) \exp(\kappa_0 s_{i,j}), \tag{29}$$

where $s_{i,j} = v_j^\top t_i$ is the cosine similarity, and $C_d(\kappa)$ acts as the vMF normalization constant.

Consequently, the log-likelihood ratio of a word being aligned versus unaligned is strictly proportional to their cosine similarity:

$$\log \frac{P(s_{i,j} \mid z_{i,j} = 1)}{P(s_{i,j} \mid z_{i,j} = 0)} = (\kappa_1 - \kappa_0) s_{i,j} + \log \frac{C_d(\kappa_1)}{C_d(\kappa_0)}. \tag{30}$$

To determine the optimal decision boundary b_j^*, this likelihood must be balanced against a prior probability of alignment, such that the log-odds sum to zero. Using a Generalized Linear Model (GLM) framework, we can heuristically parameterize the log-odds of this prior using the region's local semantic capacity (v_j) and the sentence's global semantic distribution ($\bar{t}$):

$$\log \frac{P(z_{i,j} = 1 \mid v_j, \bar{t})}{P(z_{i,j} = 0 \mid v_j, \bar{t})} \approx \theta_v^\top v_j + \theta_t^\top \bar{t} + \theta_0. \tag{31}$$

By equating the log-likelihood ratio with this context-aware prior, the optimal unconstrained threshold b_j^* emerges as a linear combination of v_j and $\bar{t}$. Absorbing the scalar constants into learnable affine weights yields a linear projection:

$$b_j^* = W_v^\top v_j + W_t^\top \bar{t} + C. \tag{32}$$

Applying a standard logistic link function to map this unbounded boundary back to a valid probability space $(0, 1)$ naturally motivates our Sigmoid-activated threshold:

$$b_j = \sigma(W_v^\top v_j + W_t^\top \bar{t} + C). \tag{33}$$

Discussion: This conceptual bridge highlights that computing b_j via a linear projection and Sigmoid activation is deeply rooted in directional statistics. It confirms that b_j acts as an adaptive decision boundary shifted dynamically by both local visual semantics and global context, rather than a purely arbitrary architectural choice.

C.3 Information-Theoretic Perspective on the Semantic Consistency Losses ($\mathcal{L}_{vsc}, \mathcal{L}_{tsc}$)

In Section 3.2.1, we introduce the symmetric semantic consistency losses $\mathcal{L}_{vsc}$ and $\mathcal{L}_{tsc}$. Rather than arbitrary regularizers, these objectives are motivated by mutual information maximization principles. Here, we briefly outline the information-theoretic perspective that guides this dual-loss design.

First, we can view the visual semantic consistency loss $\mathcal{L}_{vsc}$ (Eq. 7) through the lens of InfoNCE. Let $r \in R$ be a raw region feature, $v \in V$ be its projected semantic embedding, and $r' \in R'$ be the reconstructed feature. The encoding and decoding pipeline fundamentally constitutes a Markov chain:

$$r \xrightarrow{h_v(\cdot;\Theta_{h_v})} v \xrightarrow{g_v(\cdot;\Theta_{g_v})} r'. \tag{34}$$

According to the Data Processing Inequality (DPI), $I(R; V) \geq I(R; R')$. By leveraging the standard Noise-Contrastive Estimation (NCE) framework, the InfoNCE objective computed over the K reconstructed regions naturally maximizes a lower bound on this mutual information:

$$-\mathcal{L}_{vsc} = \mathbb{E}\left[\log \frac{f(r_i, r_i')}{f(r_i, r_i') + \sum_{j \neq i} f(r_i, r_j')}\right], \tag{35}$$

where $f(r, r') = \exp(\mathcal{S}(r, r')/\tau)$. Under the optimal critic assumption, applying Jensen's Inequality allows us to establish the classic variational bound:

$$I(R; V) \geq \log K_v - \mathcal{L}_{vsc}. \tag{36}$$

Building upon this intra-modal bound, we can further demonstrate why a *symmetric* constraint on both modalities is mathematically optimal for cross-modal matching. Let R and U denote the raw paired visual and textual features, and $V = h_v(R)$, $T = h_t(U)$ denote their projections. Our primary objective $\mathcal{L}_{htr}$ aims to maximize the cross-modal mutual information $I(V; T)$.

Using the chain rule of mutual information and the deterministic nature of the projections ($I(V; T \mid R, U) = 0$), we have:

$$I(R; U) = I(V; T) + I(R; U \mid V, T). \tag{37}$$

By expanding the conditional mutual information $I(R; U \mid V, T)$ and applying the property that conditioning reduces entropy, we can bound the information loss:

$$I(R; U \mid V, T) \leq H(R \mid V) + H(U \mid T). \tag{38}$$

Substituting this back and using $H(X \mid Y) = H(X) - I(X; Y)$, we obtain:

$$I(V; T) \geq I(R; V) + I(U; T) + [I(R; U) - H(R) - H(U)]. \tag{39}$$

Since the term $\Delta = I(R; U) - H(R) - H(U)$ is a constant dependent only on the raw intrinsic dataset properties, substituting the InfoNCE bounds for both modalities yields the ultimate optimization landscape:

$$I(V; T) \geq -(\mathcal{L}_{vsc} + \mathcal{L}_{tsc}) + (\log K_v + \log K_t + \Delta). \tag{40}$$

Discussion and design constraints: This inequality mathematically settles the necessity of our symmetric loss formulation. It demonstrates that the cross-modal alignment capacity $I(V; T)$ is additively decoupled into two intra-modal information preservation terms. Although visual and textual data originate from distinct distributions, their information degradation during projection behaves completely symmetrically in the lower bound. This provides a rigorous design constraint: employing identical InfoNCE loss structures for both $\mathcal{L}_{vsc}$ and $\mathcal{L}_{tsc}$ is not a heuristic redundancy, but the mathematically optimal strategy to simultaneously lift the lower bound of overall cross-modal matching performance.

C.4 Statistical Perspective on the Residual Semantic Re-ranking

In Section 3.3, we introduce the Residual Semantic Re-ranking (RSR) module to resolve the distinguishability bottleneck among top-k retrieved candidates. Here, we provide a simplified statistical signal-processing perspective to illustrate how isolating residual features inherently helps to disambiguate hard negatives.

Let $\mathcal{T}$ be a given text query. Let $\mathcal{V}^+$ denote the ground-truth image and $\mathcal{V}^-$ denote a hard negative image retrieved in the top-k candidates. We define the primary ranking margin M_{orig} as a random variable:

$$M_{orig} = s(\mathcal{T}, \mathcal{V}^+) - s(\mathcal{T}, \mathcal{V}^-). \tag{41}$$

A ranking error occurs if $M_{orig} < 0$. To build intuition, we conceptually model this margin using a first-order Gaussian approximation, $M_{orig} \sim \mathcal{N}(\mu_{orig}, \sigma^2_{orig})$. By the definition of the distinguishability bottleneck, hard negatives share highly similar semantic information with the true positive. Consequently, the primary similarities are nearly indistinguishable, meaning the expected margin approaches zero, and the variance is dominated by noise ($\mu_{orig} \to 0^+$, $SNR_{orig} \to 0$). The probability of a ranking error using only the primary similarity thus roughly approaches 0.5.

To resolve this, our RSR module extracts the residual textual features $\mathcal{T}_{res}$ and visual features $\mathcal{V}_{res}$ to compute the residual similarity $\tilde{s}$. We define the residual ranking margin as:

$$M_{res} = \tilde{s}(\mathcal{T}_{res}, \mathcal{V}^+_{res}) - \tilde{s}(\mathcal{T}_{res}, \mathcal{V}^-_{res}). \tag{42}$$

Because the dominant salient features have been explicitly masked out, the residual features isolate the fine-grained mismatched cues. This provides a strongly positive expected margin $\mu_{res} \gg 0$ with variance σ^2_{res}.

According to Eq. 14, the re-ranked similarity is a convex combination $\hat{s} = (1-\rho)s + \rho\tilde{s}$. While strict statistical independence is unattainable in the highly non-linear feature spaces of deep neural networks, our explicit masking strategy heavily decorrelates the residual cues from the salient ones. As a conceptual approximation, assuming they provide roughly additive evidence, the new re-ranked margin M_{new} follows:

$$M_{new} \sim \mathcal{N}((1-\rho)\mu_{orig} + \rho\mu_{res}, (1-\rho)^2\sigma^2_{orig} + \rho^2\sigma^2_{res}). \tag{43}$$

Given the bottleneck condition $\mu_{orig} \approx 0$, the new Signal-to-Noise Ratio (SNR) conceptually simplifies to:

$$SNR_{new} \approx \frac{\rho\mu_{res}}{\sqrt{(1-\rho)^2\sigma^2_{orig} + \rho^2\sigma^2_{res}}}. \tag{44}$$

Instead of claiming a strict mathematical bound, we use this SNR to qualitatively understand the ranking error probability, which decays monotonically as the SNR increases.

Discussion and testable predictions: This simplified statistical model is not meant to provide a strict theoretical guarantee, but rather to yield a clear, testable prediction regarding the fusion weight ρ. The equation for SNR_{new} mathematically demonstrates that while injecting residual semantics ($\rho > 0$) increases the signal, an excessively large ρ will cause the residual variance ($\rho^2\sigma^2_{res}$) to overwhelm the denominator, thereby degrading the SNR. This conceptually predicts that the model's performance should exhibit an inverted-U trend with respect to ρ.

D Experimental Setting

D.1 Dataset

Following existing works (Chen et al., 2025a; Zhang et al., 2025; Wang et al., 2025a; Yin et al., 2026), Flickr30K and MSCOCO remain the standard benchmarks in this research domain. The specific datasets are described below.

Flickr30K (Young et al., 2014) consists of 31,783 images collected from the Flickr website. Each image has 5 human annotated texts. We use the public training, validation and testing splits, which contain 29,783, 1,000 and 1,000 images, respectively.

Table 6: Comparison with state-of-the-art methods on the MSCOCO 5K test set.

Method	MSCOCO (5K) dataset							Method	MSCOCO (5K) dataset						
	Image Retrieval			Image Annotation			Rs		Image Retrieval			Image Annotation			Rs
	R@1	R@5	R@10	R@1	R@5	R@10			R@1	R@5	R@10	R@1	R@5	R@10	
BUTD + BiGRU								***BUTD + BERT***							
AME	40.1	70.2	80.5	54.0	82.1	90.7	417.6	AME	42.2	71.7	82.0	57.1	83.5	91.6	428.1
MV-VSE	40.3	70.6	81.6	56.7	84.1	91.4	424.6	MV-VSE	42.5	72.8	83.1	59.1	86.3	92.5	436.3
SGRAF+LG	43.0	71.8	82.2	59.1	84.9	92.0	432.9	HREM	43.9	73.6	83.3	62.3	87.6	93.4	444.1
DRCE	40.3	69.5	80.6	56.2	83.0	90.9	420.5	CHAN	44.9	74.5	84.2	59.8	87.2	93.3	443.9
HREM	40.0	70.6	81.2	58.9	85.3	92.1	428.1	USER	44.8	73.4	82.7	63.7	87.4	93.5	445.5
CHAN	41.7	71.5	81.7	60.2	85.9	92.4	433.4	TVRN	45.5	75.0	84.8	61.1	86.3	92.5	445.2
SDE	42.6	73.1	83.1	60.4	86.2	92.4	437.8	IMEB	44.9	74.6	84.0	62.8	87.8	93.5	447.6
TVRN	42.5	71.8	82.1	59.2	84.6	91.6	431.8	BOOM	45.2	75.5	85.7	63.1	87.7	93.4	450.6
IMEB	41.8	72.1	82.2	60.4	86.3	**92.6**	435.4	RVSE++	43.6	73.4	83.6	59.8	85.8	92.3	438.5
BOOM	43.4	72.1	82.5	60.5	86.2	**92.6**	437.3	DVSE	43.8	73.8	83.2	60.5	86.3	92.9	440.5
RANet	42.3	72.7	83.2	57.9	85.3	92.4	433.8	MAMET	45.6	74.9	84.1	63.0	87.7	**93.9**	449.2
RVSE++	39.6	70.1	81.1	55.3	84.5	91.3	421.9	CIMN	45.7	75.3	84.6	62.5	87.6	93.4	449.1
MAMET	41.4	70.9	81.6	57.7	85.6	92.0	429.2	FEAN	43.9	73.3	83.6	61.1	87.1	93.4	442.4
MWAN	43.5	73.5	**83.3**	61.2	**86.5**	92.5	440.5	**MWAN**	46.2	75.9	**85.8**	64.0	**88.0**	93.5	453.4
MWAN + RSR	**43.8**	**73.6**	**83.3**	**61.4**	**86.5**	92.5	**441.1**	**MWAN + RSR**	**46.5**	**76.0**	**85.8**	**64.2**	**88.0**	93.5	**454.0**
ViT + BERT								***Swin + BERT***							
VSE++	40.6	70.4	81.1	52.4	80.3	88.8	413.4	VSE++	48.1	77.1	86.1	60.7	86.6	93.2	451.8
SCAN	42.9	72.3	82.5	53.9	81.8	90.0	423.5	SCAN	49.9	78.0	86.6	64.0	88.2	94.2	460.9
SGR	42.8	72.2	82.5	54.9	82.8	90.5	425.8	SGR	48.6	77.2	86.3	61.0	86.7	93.2	453.1
CHAN	43.0	72.6	82.8	56.3	83.2	90.1	428.0	CHAN	49.1	77.3	86.1	64.1	87.9	93.5	458.0
LAPS	44.5	74.0	83.6	57.5	84.0	90.8	434.4	LAPS	51.6	78.9	87.2	64.5	89.2	94.4	465.8
AVSE	45.1	74.3	83.9	58.8	84.3	91.0	437.4	AVSE	51.7	79.2	87.3	66.2	89.8	94.7	468.9
GRM	44.4	74.5	84.1	60.1	86.7	93.2	443.0	GRM	52.1	79.8	87.6	66.3	90.1	94.9	470.8
MWAN	45.5	74.9	**84.2**	61.2	**87.1**	**93.3**	446.2	**MWAN**	52.9	80.2	**87.7**	67.2	**90.4**	**95.0**	473.4
MWAN + RSR	**45.8**	**75.0**	**84.2**	**61.5**	**87.1**	**93.3**	**446.9**	**MWAN + RSR**	**53.2**	**80.3**	**87.7**	**67.5**	**90.4**	**95.0**	**474.1**

MSCOCO (Lin et al., 2014) consists of 123,287 images, each of which is associated with 5 texts. We use the public training, validation and testing splits, with 113,287, 5,000 and 5,000 images, respectively.

D.2 Implementation Details

Following the standard practice in fine-grained image-text matching, we employ the pre-trained object detection model Faster R-CNN [1] to obtain raw region representations, setting the number of detected regions to I=36 and the dimensionality of each region representation to M=2048. For text processing, word embeddings are initialized using the pre-trained word vectors glove-840B-300d [2], and syntactic parsing is implemented via the `spaCy` library. During training, all models are optimized using the Adam optimizer for 50 epochs with a batch size of 128. The initial learning rate is set to 5e-4 and decays by 10% every 10 epochs. The trade-off factors λ_1 and λ_2 are set to 0.05 and 0.01, respectively. The dynamic temperature α is initialized to 2.0 and undergoes exponential decay with a rate of 0.95. The balancing factor ρ is fixed to 0.15. All experiments are implemented in PyTorch and executed on a single NVIDIA L40 GPU.

E Additional Experiments

E.1 Generalization Capability for Cross-Dataset Adaptation

Generalization is one crucial and practical capability for cross-modal retrieval. To assess the generalization ability of MWAN, we conduct bidirectional cross-dataset evaluations by training the model on the MSCOCO training set and testing it on the Flickr30K test set, and vice versa. As reported in Table 8, while the inherent linguistic and visual domain gap between the two datasets inevitably causes an absolute performance drop across all evaluated methods, MWAN consistently outperforms state-of-the-art baselines. Notably, MWAN + RSR achieves a peak Rs of 493.9. This indicates that our model exhibits a much stronger relative robustness under distribution shifts.

This mitigation of cross-domain degradation can be attributed primarily to the AAM. Conventional rigid alignment methods tend to overfit to dataset-specific co-occurrence statistics and annotation styles, leading to a severe performance collapse when evaluated on unseen datasets. By introducing region-specific dynamic thresholds conditioned on both local visual features and global textual

[1] https://github.com/MILVLG/bottom-up-attention.pytorch

[2] https://nlp.stanford.edu/projects/glove/

Table 7: The re-ranking results of RSR on additional backbones.

Method	Flickr30K dataset							MSCOCO dataset						
	Image Retrieval			Image Annotation			Rs	Image Retrieval			Image Annotation			Rs
	R@1	R@5	R@10	R@1	R@5	R@10		R@1	R@5	R@10	R@1	R@5	R@10	
BUTD + BiGRU														
AME	58.9	84.7	90.2	74.9	93.5	**97.0**	499.2	62.8	89.4	**95.1**	77.1	**95.4**	**98.3**	518.1
AME + **RSR**	**60.0**	**85.1**	**90.4**	**75.5**	**93.6**	**97.0**	**501.6**	**63.6**	**89.7**	**95.1**	**77.6**	**95.4**	**98.3**	**519.7**
CHAN	60.2	85.3	90.7	79.7	**94.5**	**97.3**	507.8	63.8	90.4	95.8	79.7	**96.7**	**98.7**	525.0
CHAN + **RSR**	**61.1**	**85.8**	**90.8**	**80.5**	**94.5**	**97.3**	**510.0**	**64.4**	**90.7**	**95.9**	**80.4**	**96.7**	**98.7**	**526.8**
BOOM	61.5	86.0	91.4	82.1	96.4	**98.3**	515.7	64.5	91.1	**96.3**	80.5	96.7	**99.0**	528.1
BOOM + **RSR**	**62.2**	**86.4**	**91.6**	**82.7**	**96.5**	**98.3**	**517.7**	**65.1**	**91.4**	**96.3**	**80.9**	**96.8**	**99.0**	**529.5**
BUTD + BERT														
AME	62.1	86.8	91.9	78.4	95.4	**97.8**	512.4	64.2	90.3	**95.7**	78.6	96.0	**98.6**	523.4
AME + **RSR**	**62.9**	**87.2**	**92.1**	**79.1**	**95.5**	**97.8**	**514.6**	**65.0**	**90.7**	**95.7**	**79.2**	**96.2**	**98.6**	**525.4**
CHAN	63.9	87.5	92.6	80.6	96.1	**97.8**	518.5	66.5	92.1	**96.7**	81.4	**96.9**	**98.9**	532.6
CHAN + **RSR**	**64.6**	**87.9**	**92.8**	**80.9**	**96.2**	**97.8**	**520.2**	**67.1**	**92.4**	**96.7**	**81.9**	**96.9**	**98.9**	**533.9**
BOOM	65.2	88.6	93.1	84.3	**96.5**	**98.5**	526.2	67.7	92.6	**96.8**	82.3	**97.2**	**99.1**	535.7
BOOM + **RSR**	**65.7**	**88.8**	**93.2**	**84.6**	**96.5**	**98.5**	**527.3**	**68.1**	**92.8**	**96.8**	**82.6**	**97.2**	**99.1**	**536.6**

Table 8: Results of cross-dataset generalizability between Flickr30K and MSCOCO.

Method	MSCOCO → Flickr30K							Flickr30K → MSCOCO						
	Image Retrieval			Image Annotation			Rs	Image Retrieval			Image Annotation			Rs
	R@1	R@5	R@10	R@1	R@5	R@10		R@1	R@5	R@10	R@1	R@5	R@10	
SCAN	38.4	65.0	74.4	49.8	77.8	86.0	391.4	18.5	42.0	54.2	22.1	47.5	59.8	244.1
CVSE	39.9	68.6	77.2	56.4	83.0	89.0	414.1	19.8	44.8	57.5	23.5	50.8	64.1	260.5
SGRAF	48.1	73.9	81.9	65.7	87.2	93.4	450.2	22.5	48.9	61.2	26.2	56.4	70.2	285.4
DIME	49.7	76.1	83.9	63.5	86.9	93.1	453.2	23.0	49.4	61.8	26.6	56.8	70.2	287.8
GPO	50.0	77.0	84.9	68.0	89.2	93.7	462.8	24.2	50.8	63.1	28.1	58.2	69.9	294.3
ESA	51.5	77.9	85.7	69.5	89.1	93.8	467.5	25.1	51.6	63.8	29.5	58.8	69.7	298.5
HREM	52.9	78.7	86.2	70.3	90.7	95.2	474.0	28.2	52.8	63.5	31.4	57.8	68.5	302.2
CHAN	55.6	81.2	87.5	68.7	91.5	95.3	479.8	29.5	54.2	64.8	32.6	59.2	70.1	310.4
IMEB	54.8	80.0	86.9	72.3	90.2	95.1	479.3	29.3	54.0	64.6	32.4	59.0	70.0	309.3
ESL	55.6	80.3	87.1	73.1	93.5	95.6	485.3	30.5	55.4	65.8	34.1	61.0	71.4	318.2
BOOM	57.4	82.2	88.6	72.0	91.9	96.0	488.1	31.5	56.8	67.2	35.2	62.5	73.0	326.2
MAMET	54.9	80.5	87.2	71.7	91.0	94.8	480.1	29.8	54.6	65.1	33.1	59.8	70.8	313.2
MWAN	57.8	82.5	**88.8**	73.9	**93.8**	**96.3**	493.1	32.5	58.2	68.5	36.8	64.4	74.5	334.9
MWAN + RSR	**58.1**	**82.7**	**88.8**	**74.2**	**93.8**	**96.3**	**493.9**	**32.7**	**58.4**	**68.6**	**37.0**	**64.7**	**74.6**	**336.0**

context, the AAM enables content-aware many-to-many associations. While it does not entirely eliminate the fundamental domain gap, this adaptive strategy effectively discourages the model's over-reliance on source-domain spurious correlations. Furthermore, allowing each region to flexibly anchor multiple attributes, such as *color*, *material*, and *action*, preserves intrinsic visual semantics that remain fundamentally stable across different datasets. Finally, the RSR module complements this robust alignment by refining rankings using residual cues, providing additional gains under cross-dataset evaluation.

E.2 Comparison with Capacity-Matched Sparse Alignment Alternatives

To validate the superiority of the proposed Anchored Alignment Mechanism (AAM) over capacity-matched sparse alternatives, we compared its performance against various Top-c word-selection baselines. In this context, selecting the Top-c most relevant words for each region serves as a strict capacity-constrained proxy for sparse attention mechanisms, which shares functional similarities with fixed-sparsity methods such as sparsemax, forcing the region to align with a predetermined number of words. As shown in Table 9, the performance of these static, capacity-matched strategies exhibits a clear non-monotonic trend, revealing a critical trade-off between semantic coverage and noise suppression inherent in globally fixed sparsity levels. Specifically, accuracy initially increases when expanding the alignment capacity from $c = 1$ to $c = 3$ words, with the Rs improving from 507.8 to 514.1 on Flickr30K. This initial gain reinforces our core argument that visual regions inherently possess rich semantics, demonstrating that relying on a single word through hard assignment is inadequate and causes information loss. However, indiscriminately expanding the selection ca-

pacity to $c = 10$ words or all words introduces severe noise from weakly correlated words, causing a drastic performance collapse.

Table 9: Comparison with capacity-matched sparse alignment alternatives on the Flickr30K and MSCOCO datasets.

Alignment Mechanism	Flickr30K dataset							MSCOCO dataset						
	Image Retrieval			Image Annotation			Rs	Image Retrieval			Image Annotation			Rs
	R@1	R@5	R@10	R@1	R@5	R@10		R@1	R@5	R@10	R@1	R@5	R@10	
AAM	62.9	87.1	92.4	82.4	96.9	98.8	520.5	66.7	92.2	96.7	81.9	97.3	99.1	533.9
$c = 1$	60.2	85.3	90.7	79.7	94.5	97.3	507.8	63.8	90.4	95.8	79.7	96.7	98.7	525.0
$c = 3$	61.5	86.4	91.6	80.9	95.7	98.0	514.1	65.1	91.3	96.3	80.8	97.0	98.9	529.4
$c = 5$	58.5	83.9	89.5	77.2	93.1	96.5	498.7	62.5	89.6	95.3	78.1	96.1	98.6	520.2
$c = 10$	54.2	81.5	87.8	73.5	91.8	96.2	485.0	60.9	89.0	95.1	75.6	95.5	98.5	514.6
$c = all$	48.6	77.7	85.2	67.4	90.3	95.8	465.0	58.8	88.4	94.8	72.7	94.8	98.4	507.9

In contrast to these rigid "one-size-fits-all" capacity constraints, our proposed AAM bypasses the limitations of globally fixed sparse thresholds. By leveraging adaptive, context-conditioned gating, AAM dynamically determines the optimal sparse alignment set for each region. This flexibility allows AAM to consistently outperform even the best capacity-matched static variant by a significant margin, achieving peak scores of 520.5 and 533.9, respectively. Ultimately, this demonstrates that AAM effectively preserves semantic richness while filtering out noise, ensuring robust and precise semantic alignment without requiring manually tuned sparsity capacities.

E.3 The Temperature Parameter α

The hyperparameter α plays a pivotal role in the AAM by controlling the steepness of the gating function, thereby modulating the selectivity of word-region associations. As shown in Table 10, the retrieval performance exhibits a clear inverted-U trend as α increases, peaking consistently at $\alpha = 2.0$ across both the Flickr30K and MSCOCO datasets. This empirical behavior perfectly corroborates the theoretical formulation of α as the thermodynamic temperature in the variational free energy functional (Appendix C.1). A very small α causes the Sigmoid function to behave like a rigid step function, which not only strictly limits multi-word anchoring but also suffers from gradient vanishing during backpropagation. Conversely, an excessively large α makes the gating too smooth, failing to filter out noisy alignments and degrading into non-discriminative average pooling. Notably, the optimal temperature remains remarkably stable at $\alpha = 2.0$ for both datasets. This cross-dataset consistency directly validates our theoretical prediction that because α fundamentally regulates the intrinsic thermodynamic uncertainty and information entropy of the shared embedding space, its optimal value is completely independent of dataset-specific semantic density or co-occurrence statistics. This universal characteristic endows MWAN with strong robustness and cross-dataset generalization capabilities, circumventing the need for exhaustive, per-dataset hyperparameter tuning.

Table 10: Sensitivity analysis of the initial temperature α for the annealing curriculum on Flickr30K and MSCOCO datasets.

α	Flickr30K dataset							MSCOCO dataset						
	Image Retrieval			Image Annotation			Rs	Image Retrieval			Image Annotation			Rs
	R@1	R@5	R@10	R@1	R@5	R@10		R@1	R@5	R@10	R@1	R@5	R@10	
0.1	59.4	85.4	91.5	77.8	94.5	97.6	506.2	64.7	91.3	96.4	79.0	96.2	98.8	526.4
0.2	59.9	85.6	91.7	78.7	94.9	97.8	508.6	64.8	91.3	96.4	79.1	96.2	98.8	526.6
0.5	60.9	86.3	91.9	79.8	95.5	98.0	512.4	65.5	91.6	96.5	80.2	96.7	98.9	529.4
1.0	62.0	86.9	92.2	81.3	96.2	98.5	517.1	66.1	91.9	96.6	81.1	97.0	99.0	531.7
2.0	62.9	87.1	92.4	82.4	96.9	98.8	520.5	66.7	92.2	96.7	81.9	97.3	99.1	533.9
5.0	61.1	86.2	91.9	79.9	95.6	98.0	512.7	65.9	91.8	96.6	80.7	96.8	99.0	530.8

E.4 The Re-ranking Depth k

To determine the optimal trade-off between accuracy and latency, we analyzed the performance and added latency as a function of the re-ranking depth k on both Flickr30K and MSCOCO datasets.

As shown in Table 11, performance on both datasets improves rapidly and saturates at $k = 20$, stabilizing at 521.0 Rs for Flickr30K and 534.3 Rs for MSCOCO. Crucially, increasing k further to 100 yields negligible additional gains while quintupling the computational cost (from +6ms to +30ms). This consistent trend confirms that RSR functions as a highly efficient precision filter, effectively resolving the most ambiguous hard negatives within a localized candidate pool, rendering expensive large-scale re-ranking unnecessary. Consequently, we selected $k = 20$ as the default setting, as it achieves the maximum performance boost while incurring a marginal +6ms latency overhead.

Table 11: Sensitivity analysis of the re-ranking depth k on the Flickr30K and MSCOCO datasets.

k	Flickr30K dataset							MSCOCO dataset							Added Latency
	Image Retrieval			Image Annotation			Rs	Image Retrieval			Image Annotation			Rs	
	R@1	R@5	R@10	R@1	R@5	R@10		R@1	R@5	R@10	R@1	R@5	R@10		
0	62.9	87.1	92.4	82.4	96.9	98.8	520.5	66.7	92.2	96.7	81.9	97.3	99.1	533.9	+0ms
5	63.0	87.2	92.4	82.5	96.9	98.8	520.8	66.8	92.2	96.7	82.0	97.3	99.1	534.1	+1.5ms
10	63.0	87.2	92.4	82.6	96.9	98.8	520.9	66.8	92.2	96.7	82.0	97.3	99.1	534.1	+3ms
20	63.1	87.2	92.4	82.6	96.9	98.8	521.0	66.9	92.2	96.7	82.1	97.3	99.1	534.3	+6ms
50	63.1	87.2	92.4	82.6	96.9	98.8	521.0	66.9	92.2	96.7	82.1	97.3	99.1	534.3	+15ms
100	63.1	87.2	92.4	82.6	96.9	98.8	521.0	66.9	92.2	96.7	82.1	97.3	99.2	534.4	+30ms

E.5 Comparison with Alternative Gating Mechanisms

AAM is not merely applying a Sigmoid gate, but formulating a context-aware dynamic threshold b_j conditioned on both the specific region feature v_j and global text $\bar{t}$ (Eq. 3). To isolate this contribution and address comparisons with standard gated attention alternatives, we replace our region-specific thresholding with two standard variants:

- Fixed Scalar: $b_j = \gamma$ (constant, tuned via grid search).
- Global Learnable: $b_j = b_{global}$ (a learnable parameter shared across all regions in an image).

As shown in Table 12, the disentangled experiments demonstrate that while learning a global threshold leads to a significant improvement over the fixed scalar baseline, it still lags behind our proposed AAM by 3.7 Rs on Flickr30K and 3.8 Rs on MSCOCO. This performance gap directly validates our theoretical claims. In naive thresholded attention, forcing a constant or global threshold implicitly assumes a uniform alignment prior across the entire image. However, real-world images exhibit extreme spatial variance in semantic density. Visually complex regions inherently possess a higher prior probability of anchoring multiple words and therefore logically require a lower decision boundary to accommodate many-to-many semantic alignments. Conversely, simple background patches necessitate a higher threshold to strictly reject cross-modal noise. By conditioning b_j on both the local region feature v_j and the global textual context $\bar{t}$, AAM correctly models this localized semantic granularity. Ultimately, the inability of heuristic global thresholds to match our performance confirms that the observed gains stem not from the mere introduction of a selection mechanism, but from the rigorous, context-aware mathematical adaptivity that fundamentally distinguishes AAM from naive thresholded attention paradigms.

Table 12: Comparison with alternative gating mechanisms on the Flickr30K and MSCOCO datasets.

Gating Mechanism	Variant	Description	Rs (Flickr30K)	Rs (MSCOCO)
Fixed Scalar	$\gamma = 0.1$	Constant bias (Loose)	504.2	516.5
	$\gamma = 0.2$	Constant bias (Best)	512.4	525.5
	$\gamma = 0.3$	Constant bias (Strict)	509.8	522.5
	$\gamma = 0.4$	Constant bias (Very Strict)	506.1	518.5
Global Learnable	b_{global}	Image-level adaptive bias	516.8	530.1
Region-Specific (Ours)	b_j (Eq. 3)	Region-level adaptive bias	520.5	533.9

E.6 Pooling Strategy Analysis

To analyze the effect of different pooling strategies for aggregating region-sentence similarities, we compare several representative pooling methods, with the results reported in Table 13. LSE-pooling consistently achieves the best performance across all evaluation metrics on both Flickr30K and MSCOCO, reaching the highest Rs values of 520.5 and 533.9, respectively. Conversely, Average-Pooling yields the weakest results. This occurs because Average-Pooling tends to over-smooth the similarity distribution, diluting salient cross-modal correspondences with irrelevant background noise. Meanwhile, Max-Pooling focuses exclusively on a single dominant region, ignoring informative yet complementary cues. Although intermediate strategies like Sum-Pooling and Softmax-Pooling allow for multi-region participation and thus outperform the aforementioned extremes, they still fall short of optimal dynamic range control. In contrast, LSE-Pooling acts as a smooth approximation of the maximum operation, effectively enabling multiple semantically relevant regions to contribute to the final similarity estimation. By emphasizing informative region-sentence interactions and suppressing non-discriminative fragments, LSE-Pooling aligns most naturally with the AAM, producing a more robust and expressive global similarity representation.

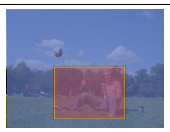
Table 13: Ablation study of various pooling strategies on the Flickr30K and MSCOCO datasets.

Pooling Types	Flickr30K dataset							MSCOCO dataset						
	Image Retrieval			Image Annotation			Rs	Image Retrieval			Image Annotation			Rs
	R@1	R@5	R@10	R@1	R@5	R@10		R@1	R@5	R@10	R@1	R@5	R@10	
Max-Pooling	60.8	86.0	91.9	80.5	95.8	98.3	513.3	64.9	91.3	96.2	79.9	96.4	98.6	527.3
Average-Pooling	58.0	84.4	91.2	77.6	94.3	97.7	503.2	62.7	90.1	95.7	77.6	95.3	98.1	519.5
Sum-Pooling	61.8	86.5	92.1	81.2	96.3	98.5	516.4	65.8	91.7	96.4	80.9	96.9	98.8	530.5
Softmax-Pooling	62.5	86.9	92.2	82.2	96.8	98.7	519.3	66.4	92.0	96.6	81.5	97.2	99.0	532.7
LSE-Pooling	62.9	87.1	92.4	82.4	96.9	98.8	520.5	66.7	92.2	96.7	81.9	97.3	99.1	533.9

E.7 Visualization and Case Study

To provide intuitive insight into the differences between MWAN and existing rigid alignment methods, we visualize the region-word alignment results for representative image-text pairs. These qualitative comparisons, shown in Table 14, reveal striking distinctions in alignment behavior. Conventional methods consistently suffer from semantic under-matching by constraining a visually complex region to a single dominant word. By rigidly anchoring a region solely to terms like *man* or *person*, these approaches entirely ignore associated attributes and contextual states. In contrast, MWAN leverages the AAM to establish robust many-to-many associations, enabling a single visual region to simultaneously anchor a comprehensive set of relevant words, such as *man*, *shirt*, and *sitting*. This adaptive mechanism preserves the multi-faceted nature of visual semantics, yielding significantly richer regional representations. In addition, the RSR module complements dominant alignments by incorporating subtle residual cues, such as *air* or *rope*, which provide additional contextual evidence for disambiguating visually similar hard negatives, ultimately enabling more nuanced and semantically complete cross-modal alignment.

Table 14: Visual comparison between MWAN and standard rigid alignment baseline. The bold tokens denote words that are semantically grounded in localized visual regions.

Text query	Existing method	Ours	Image query	Existing method	Ours
A man in a pink shirt is sitting in the grass and a ball is in the air.				A **man** in a pink shirt is sitting in the grass and a ball is in the air.	A **man** in a pink **shirt** is **sitting** in the grass and a ball is in the air.
The person has a striped shirt on and is holding on to a rope on a mountain.				The **person** has a striped shirt on and is holding on to a rope on a mountain.	The **person** has a striped **shirt** on and is holding on to a rope on a mountain.

E.8 Retrieval Latency Analysis

To evaluate the feasibility of practical deployment, which necessitates a strict balance between accuracy and efficiency, we analyze the model's retrieval latency using a single NVIDIA L40 GPU. In real-world cross-modal retrieval, the gallery of images is typically static, allowing for asymmetric computation. To provide complete transparency and address potential concerns regarding computational overhead, we strictly decouple the pipeline into offline pre-computation and online inference:

Offline pre-computation. For all images in the gallery, the computationally expensive extraction of region proposals, the projection of visual embeddings, and the construction of the visual affinity graphs $\mathcal{G}$ are performed entirely offline. Consequently, these operations contribute zero overhead to the online query latency.

Table 15: Comparison of retrieval latency on the Flickr30K and MSCOCO datasets.

Method	Interaction Type	Time per Query	Speedup vs. SCAN	Rs (Flickr30K)	Rs (MSCOCO)
SCAN	Iterative Attention	45ms	1.00 ×	465.0	507.9
SGRAF	Graph Reasoning	58ms	0.78 ×	499.6	524.3
TVRN	Iterative Refinement	66ms	0.68 ×	510.9	525.3
RANet	Adaptive Aggregation	51ms	0.88 ×	516.5	531.0
MWAN	Parallel Dense Alignment	32ms	1.40 ×	520.5	533.9
MWAN+RSR	Re-ranking	38ms	1.18 ×	521.0	534.3

Online inference. During online retrieval, the latency strictly consists of text query processing, initial dense retrieval, and the RSR top-k refinement. The precise timing breakdown is as follows:

- Text query parsing and embedding: The syntactic parsing of the input text query using the `spaCy` dependency parser takes approximately $1.5 \sim 2.0$ ms per sentence. Crucially, this operation is executed only *once* per query, regardless of the gallery size.
- Initial dense retrieval: Unlike iterative cross-attention methods, our proposed AAM utilizes highly efficient, parallelizable matrix operations to compute the global similarity between the single query and the entire image gallery, taking approximately 30.0 ms.
- RSR refinement ($k = 20$): Since the RSR module only refines the top-20 retrieved candidates, the masking of semantic features and the recalculation of residual similarities $\tilde{s}^k$ are performed strictly on a localized subset. These highly parallelized operations take merely ~ 4.0 ms.

Comparison with existing paradigms. As presented in Table 15, complex reasoning models such as SGRAF, TVRN, and RANet suffer from significant computational overhead due to dense cross-modal interactions, resulting in reduced speeds ranging from $0.68\times$ to $0.88\times$ relative to the standard SCAN baseline. In contrast, MWAN achieves the lowest retrieval latency (32 ms) and a substantial $1.40\times$ speedup. Furthermore, even with the inclusion of the RSR module, the overall retrieval latency (38 ms) remains highly competitive.