PACE: MITIGATING AUTOREGRESSIVE SCORING BIASES FOR ROBUST IMAGE-TEXT RETRIEVAL

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal large language models (MLLMs) offer rich cross-modal reasoning capabilities for image-text retrieval, but their high inference cost and biased generative likelihood limit their scalability. Global dual-encoder scores also struggle to distinguish highly similar negative samples. We propose **PACE** (**P**rogressive **A**daptive **C**ross-modal **E**nhancement), a training-free coarse-to-fine framework that selectively introduces MLLM-based evidence into efficient retrieval. PACE first employs SigLIP2 to construct a compact candidate pool and uses a margin-based guard to route only ambiguous cases to MLLM refinement. For large-scale reranking, we calibrate image-conditioned likelihood with a text-only prior using pointwise mutual information (PMI), reducing the influence of language priors. For fine-grained matching, we combine Position-Discounted Autoregressive Scoring (PDAS), which discounts later token contributions, with a Contrastive Entailment Margin (CEM) based on visual Yes/No verification. The resulting framework separates efficient candidate retrieval from fine-grained multimodal verification, allowing MLLM computation to be allocated where it is most informative. Experiments across image-text retrieval and compositional benchmarks demonstrate improved fine-grained matching while substantially reducing unnecessary MLLM inference.

1 INTRODUCTION

Multimodal large language models (MLLMs) provide rich cross-modal reasoning capabilities for image–text retrieval by evaluating candidate descriptions conditioned on visual input (Lin et al., 2025; Liu et al., 2025). However, directly using MLLMs as retrieval scorers faces two fundamental challenges. First, exhaustive evaluation requires expensive multimodal computation for every candidate, making it difficult to scale to large galleries (Miech et al., 2021; Geigle et al., 2022). Second, autoregressive likelihood reflects both image-conditioned evidence and language priors, allowing linguistically plausible but visually unsupported descriptions to receive high scores (Lin et al., 2023; Chung et al., 2025). Thus, effective MLLM-based retrieval requires not only deciding when multimodal reasoning is necessary, but also extracting visual evidence from generative scores reliably.

Dual encoders such as CLIP, ALIGN, and SigLIP provide an efficient complement to MLLMs by independently mapping images and text into a shared embedding space (Radford et al., 2021; Jia et al., 2021; Zhai et al., 2023). Their representations can be precomputed, enabling fast similarity-based retrieval over large galleries (Sun et al., 2021). However, global embedding similarity may overlook fine-grained correspondences between images and descriptions (Lee et al., 2018). In particular, descriptions that share the same objects can still differ in attributes, relations, or compositional structure, resulting in hard negatives with similar global similarity scores (Thrush et al., 2022; Yuksekgonul et al., 2023; Hsieh et al., 2023). These observations suggest a coarse-to-fine retrieval paradigm in which a dual encoder efficiently narrows the candidate space, while an MLLM is selectively introduced to resolve ambiguous matches.

Based on this observation, we propose **PACE** (**P**rogressive **A**daptive **C**ross-modal **E**nhancement), a training-free framework that combines efficient dual-encoder retrieval with selective MLLM refinement. PACE first employs SigLIP2 to construct a compact candidate pool and uses a margin-based guard to identify ambiguous cases. Candidates with a sufficiently large top-two similarity margin

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.

retain the efficient first-stage ranking, whereas ambiguous cases are routed to MLLM-based refinement. The refinement strategy is adapted to the retrieval setting. For large-scale image-to-text retrieval, we calibrate the image-conditioned likelihood with a precomputed text-only prior using pointwise mutual information (PMI) (Lin et al., 2023), reducing the influence of candidate-specific language preferences. For fine-grained matching, we combine Position-Discounted Autoregressive Scoring (PDAS), which reduces the contribution of later token predictions, with a Contrastive Entailment Margin (CEM) that compares affirmative and negative verification logits to provide explicit visual consistency evidence (Lin et al., 2024).

PACE therefore separates three complementary aspects of retrieval, including efficient candidate construction, adaptive allocation of MLLM computation, and fine-grained interpretation of multimodal evidence. It requires no additional training and preserves the scalability of dual-encoder retrieval while introducing MLLM reasoning only for ambiguous cases. Across the 11 ARO and SugarCrepe subsets, PACE improves the mean top-1 accuracy of Qwen3.5-9B from 86.9% to 94.8%. On large-scale retrieval, it improves image-to-text R@1 over the SigLIP2 first-stage retrieval from 67.1% to 69.44% on MS-COCO and from 92.3% to 93.6% on Flickr30K.

2 Related Work

2.1 Vision-Language Retrieval

Vision-language retrieval has traditionally relied on dual-encoder models that independently encode images and texts into a shared embedding space, enabling efficient similarity-based search over large candidate sets (Radford et al., 2021; Jia et al., 2021). Subsequent models improve this paradigm through stronger pre-training objectives and larger-scale data. SigLIP introduces a pairwise sigmoid loss (Zhai et al., 2023), while SigLIP2 further improves visual and semantic representations through a revised pre-training recipe (Tschannen et al., 2025).

Despite their efficiency, dual encoders provide only a global similarity signal and can struggle with fine-grained compositionality, including subtle differences in attributes, relations, and word order (Thrush et al., 2022; Yuksekgonul et al., 2023; Hsieh et al., 2023). These limitations motivate candidate-level interaction beyond embedding similarity. Our approach therefore uses a dual encoder for efficient candidate retrieval and an MLLM for subsequent fine-grained scoring.

2.2 MLLMs for Image-Text Matching

Recent work has explored MLLMs to introduce richer semantic interaction into vision-language retrieval. Some methods incorporate visual information into retrieval-oriented representations while retaining an embedding-based retrieval stage (Saito et al., 2023; Bao et al., 2025), whereas others jointly model visual and textual inputs beyond independent embeddings (Huynh et al., 2025).

MLLMs can also directly evaluate image-text pairs for retrieval or reranking (Lin et al., 2025). Reasoning-based approaches further introduce intermediate inference or verification steps (Tang et al., 2025; Yang et al., 2026). Although these methods enable deeper multimodal interaction, applying MLLMs exhaustively over large candidate sets remains computationally expensive. Our approach instead restricts MLLM scoring to a compact candidate pool and selectively refines ambiguous candidates using conditional likelihood and verification signals.

2.3 Likelihood-Based Scoring and Candidate Verification

Autoregressive likelihood provides a natural signal for image-text compatibility, but it is also affected by the intrinsic likelihood of candidate texts. Common or linguistically probable descriptions may therefore receive high scores independent of visual compatibility. We address this bias in large-scale retrieval using Pointwise Mutual Information (PMI), which contrasts image-conditioned likelihood with a precomputed text prior. Autoregressive scoring can also become less grounded toward later tokens, while existing methods mitigate fine-grained confusion through semantic parsing or token-level matching (Jiang et al., 2024; Khattab & Zaharia, 2020; Faysse et al., 2025).

Rather than introducing additional representation learning, our framework directly modifies the scoring process. Position-Discounted Autoregressive Scoring (PDAS) emphasizes visually informative tokens, while Contrastive Entailment Margin (CEM) provides an explicit verification signal for fine-grained matching. Together, they form the PACE scoring mechanism. For large-scale retrieval,

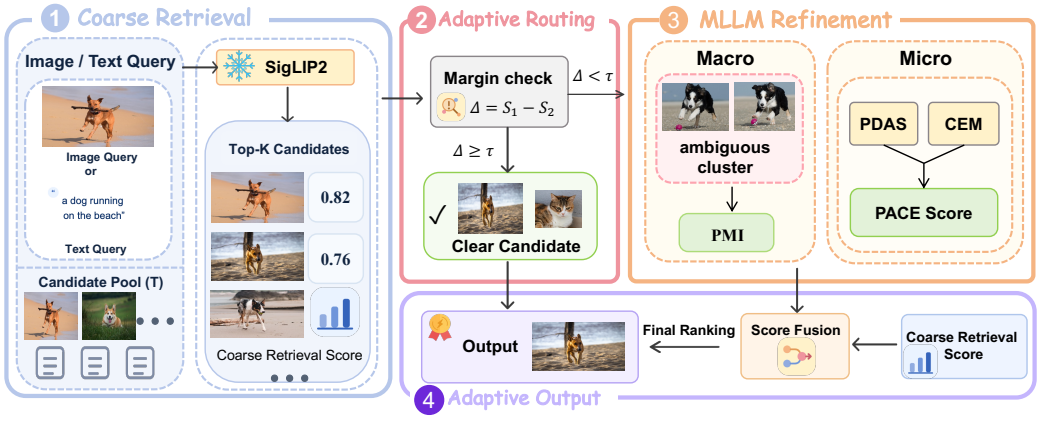


Figure 1: Overview of PACE. PACE progressively combines efficient coarse retrieval with selective MLLM refinement. SigLIP2 first constructs a compact candidate pool, followed by adaptive routing based on retrieval confidence or candidate similarity. For macro-level retrieval, ambiguous candidates are refined using PMI-calibrated likelihood for image-to-text retrieval, while text-to-image retrieval uses a fixed top-4 reranking window due to the absence of an image-side text prior. For micro-level matching, PDAS and CEM provide complementary generative and visual verification signals.

a confidence margin guard further limits MLLM reranking to uncertain candidates, preserving retrieval efficiency.

3 Methodology

3.1 Problem Formulation

We consider cross-modal retrieval between images and text. Let $\mathcal{I}$ and $\mathcal{T}$ denote the image and text domains, respectively. Given a query from one modality, the goal is to rank candidates from the other modality according to their cross-modal compatibility.

Our framework follows a coarse-to-fine design. An efficient dual-encoder model first retrieves a small candidate set from the full database. We then use an MLLM to provide finer-grained evidence for the selected candidates. For large-scale retrieval, we calibrate the MLLM likelihood with a text-only prior and invoke the expensive reranking stage only when the first-stage retrieval is ambiguous. For fine-grained matching tasks with a small candidate set, we further combine autoregressive likelihood with an explicit Yes/No consistency signal. We refer to this hybrid scoring mechanism as Progressive Adaptive Cross-modal Enhancement (PACE).

3.2 Overview

PACE follows a progressive coarse-to-fine pipeline that separates efficient candidate retrieval from selective multimodal refinement. As illustrated in Figure 1, SigLIP2 first constructs a compact candidate pool from the full gallery. An adaptive routing stage then determines whether further MLLM evaluation is necessary based on the confidence margin for macro-level retrieval or candidate similarity for micro-level matching. Ambiguous candidates are refined using task-specific multimodal evidence, including PMI-calibrated likelihood for large-scale retrieval and the combination of PDAS and CEM for fine-grained matching. The resulting scores are fused with the coarse retrieval signal to produce the final ranking.

3.3 Progressive Adaptive Cross-modal Enhancement

Fine-grained image-text matching requires more than global embedding similarity. In particular, hard negatives often differ from a correct description only in a small attribute, object, or relation. We therefore use an MLLM to evaluate candidate descriptions from two complementary perspectives. The first measures how well the candidate text is supported by the image through autoregressive likelihood. The second directly asks whether the description is visually consistent with the image.

3.3.1 POSITION-DISCOUNTED AUTOREGRESSIVE SCORING

Given an image I and a candidate description $T = (y_1, \ldots, y_N)$, an autoregressive MLLM assigns token-level conditional log-likelihoods:

$$\ell_t(I, T) = \log P_\theta(y_t \mid y_{<t}, I). \tag{1}$$

A uniform average treats all tokens equally. We instead assign larger weights to earlier positions and gradually discount later tokens using:

$$w_t = \frac{1}{\log_2(t+1)}. \tag{2}$$

The resulting Position-Discounted Autoregressive Scoring (PDAS) function is as follows:

$$S_{\mathrm{PDAS}}(I, T) = \frac{\sum_{t=1}^{N} w_t \ell_t(I, T)}{\sum_{t=1}^{N} w_t}. \tag{3}$$

The weighting reduces the contribution of later tokens while retaining the original image-conditioned autoregressive likelihood as the underlying signal. PDAS is particularly useful when candidate descriptions contain additional textual details whose likelihood may be less directly tied to the visual evidence.

3.3.2 CONTRASTIVE ENTAILMENT MARGIN

Autoregressive likelihood provides a generative compatibility signal, but it does not explicitly distinguish visual consistency from contradiction. We therefore introduce a second signal based on a binary visual verification prompt.

For an image-text pair (I, T), the MLLM is prompted to answer whether T accurately describes I without introducing incorrect or unsupported attributes. Let $z(v)$ denote the logit of a candidate answer token v. We collect the token sets corresponding to affirmative and negative responses as $\mathcal{V}_{\mathrm{yes}}$ and $\mathcal{V}_{\mathrm{no}}$, respectively. We define the Contrastive Entailment Margin (CEM) as

$$S_{\mathrm{CEM}}(I, T) = \log \sum_{v \in \mathcal{V}_{\mathrm{yes}}} \exp z(v) - \log \sum_{v \in \mathcal{V}_{\mathrm{no}}} \exp z(v). \tag{4}$$

This score directly measures the relative preference of the MLLM for visual consistency over contradiction at the verification position. Unlike sequence-level likelihood, it does not accumulate evidence over the full candidate sequence.

3.3.3 ADAPTIVE VERIFICATION ROUTING AND HYBRID SCORING

Fine-grained compositional counterfactuals are typically formed through minimal lexical edits on genuine captions. When candidate descriptions share nearly identical phrasing, standard autoregressive likelihood often fails to identify the altered local semantics. Conversely, when candidates exhibit substantial textual divergence, generative likelihood already provides a reliable ranking signal, making explicit verification redundant.

To allocate verification compute efficiently, we route candidates using the Normalized Syntactic Edit Distance across the candidate set. For a candidate pool $\mathcal{T} = \{T_1, \ldots, T_C\}$ under image I, we compute the average pairwise word-level edit distance:

$$\mathrm{NSED}(\mathcal{T}) = \frac{2}{C(C-1)} \sum_{1 \le i < j \le C} \frac{\mathcal{D}_{\mathrm{edit}}(T_i, T_j)}{\max(|T_i|, |T_j|)}, \tag{5}$$

where $\mathcal{D}_{\mathrm{edit}}$ represents the Levenshtein distance across word sequences and $|\cdot|$ denotes sentence length in words. We then define the binary verification gate as:

$$g(\mathcal{T}) = \mathbb{I}\left(\mathrm{NSED}(\mathcal{T}) \le \gamma_{\mathrm{nsed}}\right), \tag{6}$$

where $\gamma_{\mathrm{nsed}} = 0.40$ serves as the trigger threshold. When $g(\mathcal{T}) = 1$, the descriptions differ only by local modifications, which activates the Contrastive Entailment Margin S_{CEM}. We align S_{CEM} and S_{PDAS} to a common scale using pool-level min-max normalization:

$$\Phi(s_i) = \frac{s_i - \min(\mathbf{s})}{\max(\mathbf{s}) - \min(\mathbf{s}) + \epsilon}. \tag{7}$$

The final hybrid PACE score is computed as:

$$S_{\mathrm{PACE}}(I,T) = (1-\beta)\Phi\left(S_{\mathrm{CEM}}(I,T)\right) + \beta\Phi\left(S_{\mathrm{PDAS}}(I,T)\right). \tag{8}$$

When $g(\mathcal{T}) = 0$, candidates are sufficiently distinct in vocabulary, and ranking is determined entirely by S_{PDAS}, bypassing unnecessary multimodal verification prompts.

3.3.4 Pairwise Bias Decoupling

Fine-grained matching can be affected by modality-specific biases in autoregressive scoring. In particular, different images may induce different overall likelihood scales, while some descriptions may receive higher scores due to their linguistic properties rather than their visual compatibility. Such biases are especially relevant in paired evaluation settings such as Winoground (Thrush et al., 2022), where a small set of images and descriptions forms a closed matching problem. To separate these modality-specific effects from the actual image-text interaction, we construct the pairwise score matrix $\mathbf{S} \in \mathbb{R}^{M\times N}$ for each evaluation group and perform two-way centering:

$$\bar{S}_{ij} = S_{ij} - \frac{1}{N}\sum_{k=1}^{N} S_{ik} - \frac{1}{M}\sum_{k=1}^{M} S_{kj} + \frac{1}{MN}\sum_{k=1}^{M}\sum_{l=1}^{N} S_{kl}. \tag{9}$$

The row and column means capture image-specific and text-specific scoring biases, respectively, while the global mean restores the common score level. The resulting residual therefore emphasizes the relative compatibility between an image and a description after removing their marginal effects.

This operation is applied only when the evaluation protocol provides a closed pairwise matching set. For standard retrieval, where candidates are independently ranked within a query-specific pool, we retain the original scoring procedure. This gives a unified interpretation of the scoring pipeline while avoiding additional dataset-specific parameters or learned components.

3.4 Guarded Large-Scale Retrieval

While PACE provides fine-grained multimodal evidence, applying MLLM scoring to every candidate is prohibitively expensive for large-scale retrieval. We therefore use a hierarchical retrieval pipeline that first narrows the search space with a dual encoder and then invokes the MLLM only for ambiguous queries.

3.4.1 Dual-Encoder Candidate Retrieval

We use SigLIP2-So400m-patch14-384 as the first-stage retriever. Image and text inputs are independently encoded into a shared embedding space. Given normalized embeddings $\mathbf{e}_I$ and $\mathbf{e}_T$, their retrieval similarity is:

$$S_{\mathrm{enc}}(I,T) = \cos(\mathbf{e}_I, \mathbf{e}_T). \tag{10}$$

For each query, we retain the top-K candidates according to S_{enc}. This stage reduces the full retrieval problem to a small candidate pool while preserving the efficiency of dual-encoder retrieval.

3.4.2 Confidence Margin Guard

The first-stage retriever is often sufficiently confident when the leading candidate is clearly separated from the remaining candidates. In such cases, invoking the MLLM is unnecessary and may introduce ranking changes to an already well-separated result. Let s_1 and s_2 denote the similarities of the top-1 and top-2 candidates. We define the retrieval margin as:

$$\delta = s_1 - s_2. \tag{11}$$

If $\delta > \tau_{\mathrm{guard}}$, we preserve the original dual-encoder ranking. Otherwise, the query is considered ambiguous and is passed to the MLLM reranking stage.

For image-to-text retrieval, we rerank the top-M candidates within the larger top-K pool, where $M \ll K$. This selective routing reduces unnecessary MLLM computation while retaining the original retrieval result for queries with a sufficiently large first-stage margin.

3.4.3 PMI-BASED LIKELIHOOD CALIBRATION

For ambiguous image-to-text queries, the MLLM provides an image-conditioned sequence likelihood for each candidate. In the large-scale retrieval setting, we use the sequence-level log likelihood:

$$S_{\mathrm{MLLM}}(I,T) = \sum_{t=1}^{N} \log P_\theta(y_t \mid y_{<t}, I). \tag{12}$$

However, this quantity reflects not only image-text compatibility but also the intrinsic likelihood of the candidate text under the language model. Common or linguistically probable descriptions may therefore receive high scores even without strong visual support. To account for this effect, we precompute a text-only likelihood for each candidate:

$$S_{\mathrm{prior}}(T) = \sum_{t=1}^{N} \log P_\theta(y_t \mid y_{<t}, \emptyset), \tag{13}$$

where $\emptyset$ denotes the absence of visual input. We then define the calibrated score as:

$$S_{\mathrm{PMI}}(I,T) = S_{\mathrm{MLLM}}(I,T) - S_{\mathrm{prior}}(T). \tag{14}$$

This subtraction removes the candidate-specific textual component that is independent of the queried image. The resulting score therefore emphasizes the additional evidence provided by the image for the candidate description. Since $S_{\mathrm{prior}}(T)$ depends only on the candidate text, it can be computed offline once for the retrieval database.

3.4.4 CANDIDATE-LEVEL RERANKING

For an ambiguous query, we combine the dual-encoder similarity with the PMI-calibrated MLLM score. Because the two scores are on different numerical scales, we normalize them within the reranking pool using the same operator $\Phi(\cdot)$. The final image-to-text reranking score is:

$$S_{\mathrm{final}}(I,T) = \alpha\Phi\left(S_{\mathrm{enc}}(I,T)\right) + (1-\alpha)\Phi\left(S_{\mathrm{PMI}}(I,T)\right), \tag{15}$$

where α controls the relative contribution of the first-stage retriever and the MLLM-based calibration. Only the selected top-M candidates are reordered according to S_{final}. Candidates outside this reranking window retain their original dual-encoder order. The resulting pipeline combines efficient global retrieval with more detailed candidate-level multimodal scoring.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

We evaluate our framework on Winoground (Thrush et al., 2022), SugarCrepe (Hsieh et al., 2023), and ARO (Yuksekgonul et al., 2023) for micro-level compositionality, alongside MS-COCO (Karpathy 5K) and Flickr30k (Karpathy 1K) for macro-level retrieval, reporting standard Recall@K and top-1 accuracy. Baselines include dual encoders (CLIP-ViT-L/14-336, SigLIP2-So400m), multi-vector late interaction (ColPali v1.3 (Faysse et al., 2025)), and MLLMs (LLaVA-1.5-7B, Qwen2.5-VL-7B, Qwen3.5-9B). By default, we adopt SigLIP2-So400m as the initial retriever and Qwen3.5-9B as the multimodal backbone at 392×392 resolution in bfloat16. System hyperparameters are set to $K = 64$, $\tau_{\mathrm{guard}} = 0.02$, $M_{\mathrm{I2T}} = 5$ and $M_{\mathrm{T2I}} = 4$, $\alpha = 0.60$, $\beta = 0.40$, and $\gamma_{\mathrm{nsed}} = 0.40$. Additional comparative results with InternVL2.5 and detailed sensitivity analyses across hyperparameter choices are deferred to Appendix A. All experiments run on a single NVIDIA H100 GPU without parameter tuning.

4.2 MAIN RESULTS

Fine-Grained Compositional Reasoning. As shown in Table 1, PACE improves overall compositional matching performance across all three MLLM backbones. Averaged over the 11 ARO and SugarCrepe subsets, average accuracy rises from 89.1% to 90.7% for LLaVA-1.5-7B, 89.1% to 94.9% for Qwen2.5-VL-7B, and 86.9% to 94.8% for Qwen3.5-9B, showing pronounced gains on hard object and relation replacements. This validates the synergy of our design: PDAS suppresses tail language priors, while CEM builds an explicit verification boundary for fine-grained

Table 1: Fine-grained compositional evaluation on ARO and SugarCrepe (Top-1 Accuracy %). Improvements over the corresponding MLLM baseline are in **bold**.

Method	ARO				SugarCrepe							Avg
	Attr	Rel	C-Ord	F-Ord	Add-A	Add-O	Rep-A	Rep-O	Rep-R	Swp-A	Swp-O	
Dual-Encoder / Multi-Vector Baselines												
SigLIP2	100.0	100.0	100.0	100.0	79.1	91.4	85.0	95.1	72.9	71.0	67.4	87.4
SigLIP2 + ComCLIP	100.0	100.0	100.0	100.0	78.6	91.2	85.0	95.1	73.3	70.6	68.6	87.5
ColPali v1.3	82.2	56.4	67.9	75.3	50.9	19.4	76.1	91.6	65.6	81.4	66.1	66.6
LLaVA-1.5-7B												
Likelihood	100.0	100.0	100.0	100.0	**87.4**	73.7	82.9	83.8	80.6	**88.6**	83.3	89.1
+ PACE (Ours)	100.0	100.0	100.0	100.0	82.1	**78.7**	**87.6**	**93.6**	**84.0**	87.4	**84.1**	**90.7**
Qwen2.5-VL-7B												
Likelihood	100.0	100.0	100.0	100.0	86.3	61.2	87.1	85.4	81.7	**94.3**	83.7	89.1
+ PACE (Ours)	100.0	100.0	100.0	100.0	**96.2**	**91.5**	**93.3**	**96.6**	**88.6**	93.2	**84.1**	**94.9**
Qwen3.5-9B												
Likelihood	94.2	68.8	99.2	99.3	**93.1**	75.7	81.9	82.1	82.0	93.1	86.1	86.9
+ PACE (Ours)	**100.0**	**100.0**	**100.0**	**100.0**	91.8	**93.7**	**91.1**	**95.2**	**87.8**	**94.3**	**88.6**	**94.8**

Table 2: Diagnostic evaluation on Winoground benchmark (%). The best score within each MLLM backbone is highlighted in **bold**.

Method	Text Score	Image Score	Group Score
Dual-Encoder / Multi-Vector Baselines			
SigLIP2	38.50	19.75	17.00
SigLIP2 + ComCLIP	37.75	18.50	15.50
ColPali v1.3	35.50	17.25	15.00
LLaVA-1.5-7B Backbone			
Likelihood (Baseline)	0.00	78.50	0.00
+ PACE (Ours)	**80.00**	**80.00**	**80.00**
Qwen2.5-VL-7B Backbone			
Likelihood (Baseline)	0.00	80.25	0.00
+ PACE (Ours)	**11.75**	**88.00**	**11.75**
Qwen3.5-9B Backbone			
Likelihood (Baseline)	37.50	40.25	24.00
+ PACE (Ours)	**89.00**	**89.00**	**89.00**

entity swaps. In contrast, ComCLIP yields marginal improvements ($+0.05\%$) over SigLIP2, and ColPali v1.3 lags significantly on relational reasoning (66.6% average), confirming the necessity of calibrated MLLM scoring.

Diagnostic Evaluation on Winoground. Table 2 reports diagnostic results on counterfactual pairs. Due to strong language priors, uniform likelihood collapses to 0.00% Text and Group scores on LLaVA-1.5-7B and Qwen2.5-VL-7B. Combining bi-directional bias decoupling with CEM verification fundamentally restores cross-modal sensitivity: Group scores reach 80.00% on LLaVA-1.5-7B and 89.00% on Qwen3.5-9B ($+65.00\%$ over its baseline). This confirms that extracting the second-order interaction residual isolates true visual-textual compatibility from modality-specific main effects.

Macro-Scale Bidirectional Retrieval. Table 3 summarizes bidirectional retrieval on MS-COCO and Flickr30k. We evaluate two distinct autoregressive baselines: *Full-Pool Baseline* directly evaluates generative likelihood across the unconstrained gallery without dual-encoder filtering, whereas + *Likelihood* reranks only the compact candidate pool retrieved by SigLIP2. As shown, unconstrained full-pool evaluation severely degrades text retrieval (e.g., dropping to 45.1% R@1 on Qwen3.5-9B) while requiring exhaustive pairwise forward passes across the entire gallery. This quadratic inference cost becomes computationally intractable on LLaVA-1.5-7B, where full-pool evaluation is omitted due to extreme latency. Conversely, while dual-encoder filtering bounds computation,

Table 3: Bidirectional image-text retrieval on MS-COCO (5K) and Flickr30k (1K) (%). Top-performing reranking results within each MLLM backbone that surpass the first-stage SigLIP2 baseline in **R@1** are highlighted in **bold**.

Method	MS-COCO (5K)						Flickr30k (1K)					
	I2T (Text)			T2I (Image)			I2T (Text)			T2I (Image)		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
Dual-Encoder Baselines (First-Stage)												
SigLIP2	67.1	87.7	93.2	52.5	75.7	83.5	92.3	99.1	99.9	79.3	94.3	96.9
SigLIP2 + ComCLIP	67.6	88.1	93.3	52.9	76.0	83.7	93.4	99.4	99.9	79.8	94.4	97.0
LLaVA-1.5-7B												
Full-Pool Baseline	—	—	—	—	—	—	—	—	—	—	—	—
+ Likelihood	59.9	87.7	93.2	**58.9**	75.7	83.5	87.5	99.1	99.9	**85.0**	94.3	96.9
+ PACE (Ours)	**69.5**	87.7	93.2	56.3	75.7	83.5	**93.1**	99.1	99.9	82.9	94.3	96.9
Qwen2.5-VL-7B												
Full-Pool Baseline	47.0	70.0	75.1	74.7	88.3	91.5	60.5	80.0	86.5	81.4	90.8	93.3
+ Likelihood	63.8	87.7	93.2	54.8	75.7	83.5	90.1	99.1	99.9	80.1	94.3	96.9
+ PACE (Ours)	**69.5**	87.7	93.2	**55.7**	75.7	83.5	**93.1**	99.1	99.9	**82.2**	94.3	96.9
Qwen3.5-9B												
Full-Pool Baseline	45.1	70.0	79.1	85.1	96.4	98.4	65.5	81.0	88.0	92.6	97.9	98.8
+ Likelihood	62.8	87.7	93.2	**55.8**	75.7	83.5	90.2	99.1	99.9	82.4	94.3	96.9
+ PACE (Ours)	**69.44**	87.7	93.2	55.7	75.7	83.5	**93.6**	99.1	99.9	**82.5**	94.3	96.9

Note: Full-pool baseline for LLaVA-1.5-7B is omitted (denoted as —) due to prohibitive computational cost constraints without first-stage filtering.

standard candidate-level reranking (+ *Likelihood*) still triggers negative transfer due to generic caption bias, dropping LLaVA from 67.1% to 59.9% on MS-COCO. In contrast, our PMI calibration explicitly discounts unconditional text priors, allowing PACE to surpass both raw likelihood and native SigLIP2 across all models (reaching 69.5% on MS-COCO and 93.6% on Flickr30k). For image retrieval, the confidence margin guard exempts well-separated candidates from redundant scoring while refining ambiguous clusters, boosting R@1 by up to $+3.88\%$ on MS-COCO and $+3.56\%$ on Flickr30k over SigLIP2. Together, these results demonstrate that our guarded, calibrated pipeline scales MLLM reasoning to large retrieval pools without compromising fidelity or efficiency.

4.3 Analysis and Ablation

Table 4 shows the contribution of PMI calibration and the margin guard. Removing PMI reduces I2T R@1 from 69.44% to 67.84%, confirming that text-prior calibration is important for reliable likelihood-based reranking. Removing the margin guard has almost no effect on accuracy, with I2T R@1 changing from 69.44% to 69.40%, as its primary role is to avoid unnecessary MLLM evaluation rather than modify the ranking score. The small variation in T2I results comes from our asymmetric implementation. Since PMI and the image-conditioned PACE score cannot be directly applied to image candidates, we fix the reranking window to the top four images within the top ten SigLIP2 candidates. The resulting T2I fluctuations are therefore mainly caused by this fixed candidate selection and remain small across configurations.

Table 5 further demonstrates the complementary roles of PDAS and CEM in fine-grained matching. CEM alone achieves 86.58% micro-average accuracy, while adding PDAS improves it to 87.12%. In contrast, removing CEM and using PDAS alone reduces the score to 79.25%. These results show that position-aware generative evidence and explicit visual entailment provide complementary signals for distinguishing fine-grained compositional differences.

5 Conclusion

We presented PACE, a training-free coarse-to-fine framework for cross-modal retrieval that combines efficient dual-encoder retrieval with selective MLLM refinement. PACE uses SigLIP2 to construct a compact candidate pool and a margin-based guard to allocate MLLM computation to ambiguous cases. To improve the reliability of MLLM scoring, PMI calibration reduces the influence of

Table 4: Component ablation on MS-COCO (5K) using Qwen3.5-9B (%). Best R@1 are in **bold**.

Variant / Configuration	Text Retrieval (I2T)			Image Retrieval (T2I)			R@1 Avg
	R@1	R@5	R@10	R@1	R@5	R@10	
Ours (Full: PMI + Guard)	**69.44**	87.68	93.18	**55.70**	75.72	83.49	**62.57**
w/o PMI (Raw Visual Likelihood)	67.84	87.68	93.18	55.63	75.72	83.49	61.74
w/o Margin Guard (Force Rerank)	69.40	87.68	93.18	55.68	75.72	83.49	62.54
Reference Baselines							
Pure SigLIP2 Baseline	67.14	87.68	93.18	52.45	75.72	83.49	59.79
Pure MLLM Likelihood Baseline	50.88	86.02	93.18	55.20	77.73	83.49	53.04

Table 5: Component ablations of PACE using Qwen3.5-9B (%). (a) SugarCrepe (Micro R@1); (b) MS-COCO 5K (Macro retrieval).

(a) Micro-Level Component Ablation (PACE on SugarCrepe)

Variant / Configuration	Rep-Obj	Rep-Rel	Micro-Avg
Ours (Full: CEM + PDAS)	**93.16**	**81.08**	**87.12**
w/o PDAS (Pure CEM)	92.86	80.30	86.58
w/o CEM (Pure PDAS)	80.75	77.74	79.25

(b) Macro-Level System Ablation (Retrieval on MS-COCO)

Configuration	TR@1	IR@1	Macro-Avg
Ours (Full: PMI + Guard)	**69.44**	**55.70**	**62.57**
w/o PMI Calibration	67.84	55.63	61.74
w/o Margin Guard	69.40	55.68	62.54
Pure SigLIP2 Baseline	67.14	52.45	59.79
Pure MLLM Likelihood	50.88	55.20	53.04

text-specific language priors in large-scale retrieval, while PDAS and CEM provide complementary generative and visual entailment evidence for fine-grained matching. Experiments across compositional and large-scale retrieval benchmarks show that PACE consistently improves fine-grained discrimination and image-to-text retrieval over direct likelihood-based scoring while preserving the efficiency of a dual-encoder first stage. These results suggest that MLLMs are most effective in retrieval when used as selective sources of fine-grained cross-modal evidence rather than as exhaustive retrieval scorers. Future work will explore more symmetric MLLM scoring for image and text candidates and further reduce the cost of multimodal refinement.

AI USE STATEMENT

In this work, we used generative AI tools exclusively for code completion and grammatical polishing of the manuscript. Generative AI tools were not used for formulating research hypotheses, designing the mathematical formulations, or drafting the core scientific arguments. All experimental execution pipelines and evaluation scripts were independently implemented, run on hardware, and verified for numerical correctness by the authors. We take full responsibility for the final scientific conclusions, experimental figures, and technical artifacts presented in this work.

ETHICS STATEMENT

This work studies training-free scoring mechanisms and hierarchical routing for vision-language retrieval. All evaluations are conducted entirely on established, publicly available academic benchmarks without introducing proprietary personal information or human subjects. The models used in this study are publicly released open-source checkpoints. While multimodal systems can inherit data-driven biases present in their pre-training corpora, our proposed framework focuses on calibration and contrastive alignment without promoting harmful or sensitive content. To the best of our knowledge, this work poses no direct negative societal impact.

REPRODUCIBILITY STATEMENT

To ensure full reproducibility of our findings, all mathematical formulations, decision rules, and evaluation protocols are thoroughly specified. Specifically, the definitions of Position-Discounted Autoregressive Scoring (PDAS), the Contrastive Entailment Margin (CEM), Normalized Syntactic Edit Distance (NSED) routing, and adaptive bias decoupling are formally derived in Section 3. Complete hardware specifications, model checkpoint identifiers, prompt templates, and hyperparameter choices (such as the reranking window M, confidence margin guard threshold τ_{guard}, and combination weights α and β) are detailed in Section 4 and A. All evaluations are carried out on standard public benchmarks (ARO, SugarCrepe, Winoground, MS-COCO, and Flickr30k) using deterministic decoding configurations. The full codebase, configuration scripts, and precomputed prior logs are provided in the supplementary material and will be made publicly available upon publication.

REFERENCES

Tong Bao, Che Liu, Derong Xu, Zhi Zheng, and Tong Xu. Mllm-i2w: Harnessing multimodal large language model for zero-shot composed image retrieval. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 1839–1849, 2025.

Jiwan Chung, Seungwon Lim, Sangkyu Lee, and Youngjae Yu. Mass: Overcoming language bias in image-text matching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 2591–2599, 2025.

Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre Colombo. Colpali: Efficient document retrieval with vision language models. In *International Conference on Learning Representations*, volume 2025, pp. 61424–61449, 2025.

Gregor Geigle, Jonas Pfeiffer, Nils Reimers, Ivan Vulić, and Iryna Gurevych. Retrieve fast, rerank smart: Cooperative and joint approaches for improved cross-modal retrieval. *Transactions of the Association for Computational Linguistics*, 10:503–521, 2022.

Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. Sugarcrepe: Fixing hackable benchmarks for vision-language compositionality. *Advances in neural information processing systems*, 36:31096–31116, 2023.

Chuong Huynh, Jinyu Yang, Ashish Tawari, Mubarak Shah, Son Tran, Raffay Hamid, Trishul Chilimbi, and Abhinav Shrivastava. Collm: A large language model for composed image retrieval. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3994–4004. IEEE, 2025.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pp. 4904–4916. PMLR, 2021.

Kenan Jiang, Xuehai He, Ruize Xu, and Xin Wang. Comclip: Training-free compositional image and text matching. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 6639–6659, 2024.

Omar Khattab and Matei Zaharia. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pp. 39–48, 2020.

Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *European conference on computer vision*, pp. 212–228. Springer, 2018.

Sheng-Chieh Lin, Chankyu Lee, Mohammad Shoeybi, Jimmy Lin, Bryan Catanzaro, and Wei Ping. Mm-embed: Universal multimodal retrieval with multimodal llms. In *International Conference on Learning Representations*, volume 2025, pp. 44215–44234, 2025.

Zhiqiu Lin, Xinyue Chen, Deepak Pathak, Pengchuan Zhang, and Deva Ramanan. Revisiting the role of language priors in vision-language models. *arXiv preprint arXiv:2306.01879*, 2023.

Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation. In *European Conference on Computer Vision*, pp. 366–384. Springer, 2024.

Yikun Liu, Yajie Zhang, Jiayin Cai, Xiaolong Jiang, Yao Hu, Jiangchao Yao, Yanfeng Wang, and Weidi Xie. Lamra: Large multimodal model as your advanced retrieval assistant. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4015–4025. IEEE, 2025.

Antoine Miech, Jean-Baptiste Alayrac, Ivan Laptev, Josef Sivic, and Andrew Zisserman. Thinking fast and slow: Efficient text-to-visual retrieval with transformers. *arXiv preprint arXiv:2103.16553*, 2021.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Kuniaki Saito, Kihyuk Sohn, Xiang Zhang, Chun-Liang Li, Chen-Yu Lee, Kate Saenko, and Tomas Pfister. Pic2word: Mapping pictures to words for zero-shot composed image retrieval. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19305–19314. IEEE, 2023.

Siqi Sun, Yen-Chun Chen, Linjie Li, Shuohang Wang, Yuwei Fang, and Jingjing Liu. Lightningdot: Pre-training visual-semantic embeddings for real-time image-text retrieval. In *Proceedings of the 2021 conference of the north american chapter of the association for computational linguistics: human language technologies*, pp. 982–997, 2021.

Yuanmin Tang, Jue Zhang, Xiaoting Qin, Jing Yu, Gaopeng Gou, Gang Xiong, Qingwei Lin, Saravan Rajmohan, Dongmei Zhang, and Qi Wu. Reason-before-retrieve: One-stage reflective chain-of-thoughts for training-free zero-shot composed image retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14400–14410, 2025.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5228–5238. IEEE, 2022.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

Zhongyu Yang, Wei Pang, and Yingfang Yuan. Xr: Cross-modal agents for composed image retrieval. In *Proceedings of the ACM Web Conference 2026*, pp. 2071–2082, 2026.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=KRLUvxh8uaX.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11941–11952. IEEE, 2023.

A Supplementary Analyses and Additional Backbones

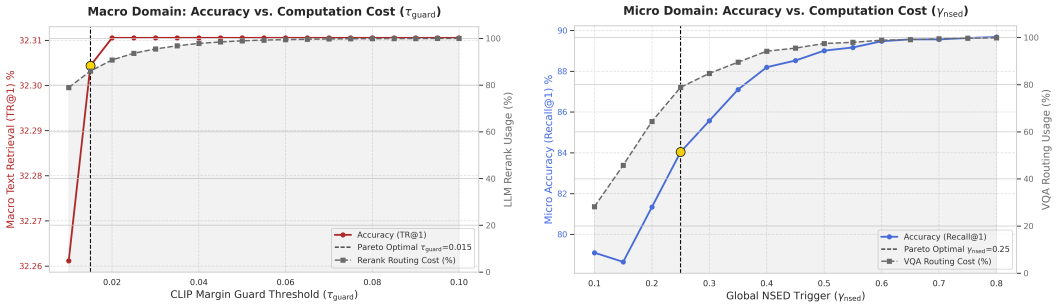
A.1 Efficiency and Sensitivity Analysis

To evaluate the stability of PACE and understand the trade-offs in our design, we conduct extensive sensitivity experiments across the macro- and micro-retrieval domains. Specifically, we investigate two categories of hyperparameters comprising four critical variables: the computational gating thresholds (τ_{guard} and γ_{nsed}) in Figure 2, and the late-fusion interpolation weights (α and β) in Figure 3.

Routing Thresholds vs. Computational Overhead. Figure 2 illustrates the trade-off between retrieval accuracy and MLLM invocation cost across varying gating thresholds. In the macro domain (Figure 2a), increasing the CLIP margin guard threshold τ_{guard} routes more ambiguous candidates to MLLM reranking until accuracy saturates. Although the theoretical Pareto-optimal inflection point occurs at $\tau_{\text{guard}} = 0.015$ (invoking approximately 85% reranking compute), we adopt $\tau_{\text{guard}} = 0.020$ in our final system. This trade-off slightly increases MLLM invocation to ensure that near-boundary hard false positives are thoroughly refined, maximizing the top-1 retrieval performance.

Similarly, in the micro domain (Figure 2b), the NSED trigger γ_{nsed} controls the activation of the compute-intensive Contrastive Entailment Margin (CEM). While the Pareto-optimal knee is located at $\gamma_{\text{nsed}} = 0.25$ (activating explicit VQA on roughly 72% of cases), fine-grained counterfactuals often exhibit subtle lexical variations that fall just beyond this strict threshold. We therefore select $\gamma_{\text{nsed}} = 0.40$, trading a moderate increase in VQA routing cost (reaching $\sim$90%) for a substantial gain in fine-grained accuracy (from $\sim$84% to over 88%). This choice prioritizes absolute compositional discrimination over conservative compute savings.

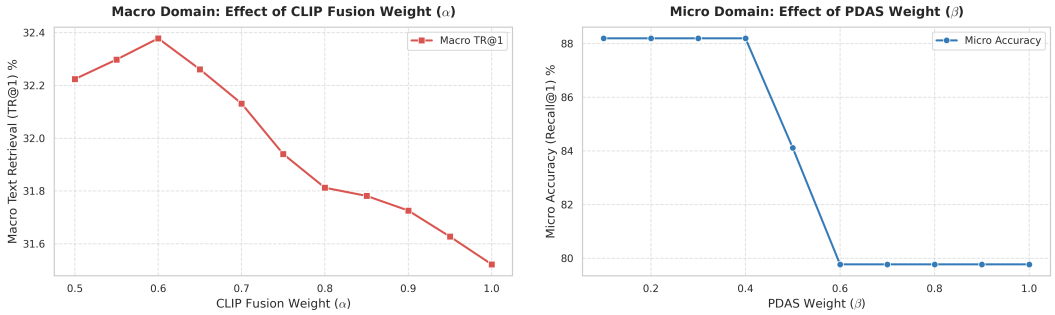
Sensitivity of Fusion Weights. Figure 3 examines the influence of scoring mixture weights across the two evaluation regimes. For macro retrieval (Figure 3a), the dual-encoder weight α balances global representation retrieval against the PMI-calibrated MLLM likelihood increment. Accuracy exhibits a clear unimodal curve that peaks sharply at $\alpha = 0.60$. Over-relying on the first-stage retriever ($\alpha \to 1.0$) causes severe degradation by discarding multimodal reasoning, whereas excessively weighting raw MLLM scores ($\alpha \leq 0.50$) reintroduces variance from non-global visual features.



(a) Macro Domain (τ_{guard}) (b) Micro Domain (γ_{nsed})

Figure 2: **Trade-off between retrieval accuracy and computational cost across domains.** Yellow circles indicate the Pareto-optimal thresholds: (a) CLIP margin guard threshold $\tau_{\text{guard}} = 0.015$ in the macro domain, and (b) NSED trigger $\gamma_{\text{nsed}} = 0.25$ in the micro domain.

In the micro domain (Figure 3b), the parameter β modulates the balance between the generative likelihood (PDAS) and the discriminative margin (CEM). The compositional accuracy remains remarkably plateaued and optimal across $\beta \in [0.1, 0.4]$, confirming that explicit visual verification provides the dominant, robust signal for resolving local entity edits. However, as β surpasses 0.40 and approaches pure autoregressive scoring ($\beta \to 1.0$), performance sharply drops by over 8%, corroborating that unconstrained likelihood alone cannot effectively discriminate minimal counterfactual swaps. We accordingly fix $\beta = 0.40$ to preserve the beneficial sequence-level regularization while retaining full discriminative margin power.



(a) Macro Domain (α) (b) Micro Domain (β)

Figure 3: **Sensitivity analysis of fusion weights across domains.** (a) Effect of CLIP fusion weight α on Macro text retrieval accuracy (TR@1), peaking around $\alpha = 0.6$. (b) Effect of PDAS weight β on Micro retrieval accuracy (Recall@1), demonstrating stability for $\beta \leq 0.4$ before degrading.

Impact of Reranking Window Size. Table 6 evaluates the sensitivity of the candidate reranking window M on MS-COCO. Expanding M from 2 to 4 yields substantial performance gains, with the overall macro average jumping from 59.79% to 62.47%. This confirms that evaluating fine-grained multimodal interactions over a small set of top candidates effectively rectifies ambiguity in the initial dual-encoder ranking.

However, enlarging the window further exhibits clear diminishing marginal returns. Increasing M from 5 to 10 requires double the multimodal forward passes yet produces only a minor gain of +0.67% in Macro-Avg (from 62.80% to 63.47%). Moreover, asymmetric evaluation indicates that text retrieval benefits from slightly broader candidate coverage ($M = 5$), whereas image retrieval achieves competitive accuracy with a tighter set ($M = 4$) while mitigating visual drift. We therefore adopt the asymmetric default setting (TR $= 5$, IR $= 4$), which captures the vast majority of retrieval gains while strictly bounding inference latency.

Table 6: Sensitivity analysis of the candidate reranking window M on MS-COCO (5K split) using Qwen3.5-9B (%). Performance exhibits diminishing marginal returns beyond the default setting.

Window Setting	**TR@1**	**TR@5**	**IR@1**	**IR@5**	**Macro-Avg**
Symmetric $M = 2$	67.14	87.68	52.45	75.72	59.79
Symmetric $M = 3$	68.62	87.68	55.02	75.72	61.82
Symmetric $M = 4$	69.30	87.68	55.63	75.72	62.47
Symmetric $M = 5$	69.44	87.68	56.15	75.72	62.80
Default (TR=5, IR=4)	**69.44**	87.68	55.70	75.72	62.57
Extended Windows					
Symmetric $M = 8$	69.88	88.98	56.64	77.91	63.26
Symmetric $M = 10$	70.00	89.30	56.94	78.34	63.47

A.2 Additional Results on InternVL2.5

Table 7: Fine-grained compositional evaluation using InternVL2.5 backbone (Top-1 Accuracy %). Improvements over the corresponding MLLM baseline are highlighted in **bold**.

Method	**ARO**				**SugarCrepe**							**Avg**
	Attr	**Rel**	**C-Ord**	**F-Ord**	**Add-A**	**Add-O**	**Rep-A**	**Rep-O**	**Rep-R**	**Swp-A**	**Swp-O**	
First-Stage Retriever												
CLIP-ViT-L/14	100.0	100.0	100.0	100.0	74.0	80.9	81.0	94.7	66.9	62.3	64.9	84.1
InternVL2.5 Backbone												
Likelihood	100.0	100.0	100.0	100.0	89.9	68.2	87.3	86.2	83.6	93.4	87.8	90.6
+ PACE (Ours)	100.0	100.0	100.0	100.0	**97.5**	**92.3**	**95.8**	**97.5**	**90.8**	**94.6**	87.8	**96.0**

Table 8: Diagnostic evaluation on Winoground using InternVL2.5 (%). Best scores within the backbone are highlighted in **bold**.

Method	Text Score	Image Score	Group Score
First-Stage Retriever			
CLIP-ViT-L/14	27.25	12.00	8.25
InternVL2.5 Backbone			
Likelihood	0.00	79.50	0.00
+ PACE (Ours)	**89.25**	**89.25**	**89.25**

Fine-Grained Compositional Evaluation. Table 7 presents the compositional evaluation results when deploying InternVL2.5 as the multimodal backbone. Across both ARO and SugarCrepe, PACE demonstrates substantial gains over standard likelihood, increasing the overall macro average from 90.6% to 96.0%. Consistent with our findings on other backbones, the performance improvement is particularly pronounced on complex entity and attribute manipulations: top-1 accuracy rises from 68.2% to 92.3% on Add-Object ($+24.1\%$), from 86.2% to 97.5% on Replace-Object ($+11.3\%$), and from 87.3% to 95.8% on Replace-Attribute ($+8.5\%$). While dual-encoder models like CLIP-ViT-L/14 achieve competitive performance on coarse entity recognition, they struggle significantly on relational and attribute bindings (e.g., 62.3% on Swap-Attribute). By adaptively combining position-discounted likelihood with contrastive entailment verification via NSED gating, PACE reliably disentangles subtle counterfactual swaps without relying on external task-specific fine-tuning.

Counterfactual Diagnostic Verification on Winoground. Table 8 reports the diagnostic matching results on the challenging Winoground benchmark. Due to unconditioned linguistic priors, the naive autoregressive likelihood baseline completely collapses to 0.00% on both Text Score and Group Score, despite achieving a high Image Score of 79.50%. This discrepancy exposes the vulnerability of generative likelihood to language frequency shifts when contrasting paired counterfactual descriptions. By decoupling bidirectional modality biases and integrating contrastive verification, PACE achieves a balanced score of 89.25% across all three evaluation dimensions, drastically outperforming the initial CLIP-ViT-L retriever (8.25% Group Score). This robust improvement confirms that our two-way centering operation isolates genuine cross-modal interaction residuals, enabling reliable disambiguation even on closed counterfactual test beds.

Table 9: Bidirectional retrieval with InternVL2.5 on MS-COCO (5K) and Flickr30k (1K) (%). Reranking results surpassing the initial CLIP baseline in **R@1** with the top score in the group are highlighted in **bold**.

Method	MS-COCO (5K)						Flickr30k (1K)					
	I2T (Text)			T2I (Image)			I2T (Text)			T2I (Image)		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
First-Stage Retriever												
CLIP-ViT-L/14	58.0	81.0	88.0	37.3	62.0	72.0	87.8	98.6	99.4	67.0	89.1	93.6
InternVL2.5 Backbone												
Full-Pool Baseline	54.6	81.0	85.8	84.9	96.6	98.0	48.9	74.1	83.3	82.5	94.6	97.0
+ Likelihood	55.3	81.0	88.0	**49.7**	62.0	71.9	82.7	98.7	99.4	**80.0**	89.0	93.6
+ PACE (Ours)	**62.3**	81.0	88.0	42.5	62.0	71.9	**90.9**	98.7	99.4	73.3	89.0	93.6

Macro-Scale Bidirectional Retrieval. Table 9 summarizes bidirectional retrieval on MS-COCO and Flickr30k using the InternVL2.5 backbone alongside the CLIP-ViT-L/14 first-stage retriever. Evaluating generative likelihood directly across the full, unconstrained gallery without dual-encoder filtering (*Full-Pool Baseline*) yields suboptimal text retrieval accuracy, reaching only 54.6% on MS-COCO and 48.9% on Flickr30k while incurring prohibitive latency that scales quadratically with database size. Within the scalable two-stage pipeline, standard reranking with raw image-conditioned likelihood (+ *Likelihood*) still causes noticeable negative transfer in text retrieval, dropping from 58.0% to 55.3% on MS-COCO and from 87.8% to 82.7% on Flickr30k, because the model over-rewards linguistically generic captions. Our PMI calibration resolves this failure mode

by explicitly discounting unconditional text priors, allowing PACE to achieve 62.3% and 90.9% R@1, comfortably outperforming both the naive likelihood baseline and the initial CLIP retriever.

In the text-to-image (T2I) direction, performance gains show moderate variance across reranking variants. Unlike text candidates, visual inputs cannot be calibrated through simple unconditional prior subtraction, leaving the MLLM to rely on raw visual likelihood. Moreover, to constrain multimodal inference latency, candidate reranking is restricted to a tight window of the top four visual candidates ($M = 4$). This truncation introduces score fluctuations whenever the target image sits near the boundary of the initial pool. Even with this trade-off, our margin-guarded scoring effectively preserves retrieval fidelity, consistently improving R@1 over the native CLIP baseline across both benchmarks.

B Implementation Details

B.1 Prompts

Our framework uses two prompts for different levels of candidate verification. For macro-level scoring, we use the following prompt for both image-conditioned likelihood estimation and the text-only prior:

```
Provide a caption that accurately describes this
image.
```

Given a candidate caption t, the model is evaluated with the candidate provided as the target assistant response. We compute its conditional log-likelihood from the model's forward-pass logits rather than generating a response.

For micro-level verification, we use a binary question-answering prompt:

```
Does the following description accurately describe the
image without any extra or incorrect attributes?
Description: '{candidate}'
Answer strictly 'Yes' or 'No'.
Answer:
```

The model again performs a forward pass and does not generate text. The resulting Yes/No probabilities are used to determine whether the candidate description is consistent with the image.

B.2 Candidate Routing

To reduce unnecessary multimodal computation, we first estimate the semantic similarity among candidate descriptions using Normalized String Edit Distance. When the average Normalized String Edit Distance is no greater than 0.4, the candidate set is considered sufficiently similar and is passed to the micro-level VQA verification stage.

For text-to-image retrieval, an image-side text-only prior is unavailable. Therefore, PMI cannot be directly computed for image candidates. In this setting, we use a fixed candidate window and force the top four SigLIP candidates into the multimodal reranking stage. This provides a consistent reranking budget while avoiding an undefined image-side prior.