PURIFYING GENERIC VISUAL SCENE GRAPHS TO ENABLE CAPTION-ALIGNED IMAGE-TEXT RETRIEVAL

Anonymous authors
Paper under double-blind review

ABSTRACT

Scene graph matching methods for image-text retrieval represent images and captions with structured entities and relations, enabling fine-grained cross-modal matching beyond isolated region-word alignment. However, existing methods typically rely on generic scene graph generators, which produce Visual Scene Graphs (VSGs) containing caption-irrelevant semantics. This creates a semantic gap with the sparse Textual Scene Graphs (TSGs) parsed from captions. To bridge this gap, we propose PSGR, a retrieval-oriented framework that purifies generic visual scene graphs prior to matching. PSGR uses GPT-5-generated caption-guided supervision to adapt a Large Vision-Language Model (LVLM) with SFT and GRPO, then applies Rank-Preserved Candidate Sampling (RPCS) to select informative and diverse triplets from refined and generic graph candidates. The resulting Purified Scene Graphs (PSGs) are compared with TSGs through local matching and global GPO-based graph matching. The LVLM refiner is adapted using Flickr30K training supervision and then frozen to generate offline graph inputs, while the retrieval matchers are trained separately for each benchmark. Under this protocol, PSGR obtains 558.9 and 462.0 Rsum on Flickr30K and MS-COCO 5K, respectively, outperforming the existing scene graph matching methods. With the visual backbone and matching architecture held fixed, LVLM-based graph refinement and RPCS jointly improve Rsum by 13.0 points on Flickr30K and 6.6 points on MS-COCO 5K over generic VSG top-rank sampling. Our curated data and code are available at `https://anonymous.4open.science/r/PSGR-DA6D`

1 INTRODUCTION

Image–Text Retrieval (ITR) aims to align matched visual and textual representations in a shared embedding space, serving as the bedrock for various downstream vision-language tasks (Kim et al., 2025; Kuang et al., 2025; Cao et al., 2025; Yuan et al., 2025). Recently, scene graphs provide a structured approach to facilitate fine-grained ITR by encoding rich semantic information into structured entities and relations (Wang et al., 2020b; Zeng et al., 2021; Liu et al., 2024; Nguyen et al., 2021). Standard graph-based ITR frameworks typically integrate local detail matching with global graph alignment (Nguyen et al., 2021; Liu et al., 2024). Specifically, local-level matching captures fine-grained correspondences across entities and relations, while global-level matching aggregates these local features into a holistic representation to measure cross-modal similarity in the joint space.

To model this paradigm, existing methods focus on devising specialized networks for different matching granularities, typically employing Graph Convolutional Networks (GCNs) (Kipf & Welling, 2016) to refine local features and pooling mechanisms for global fusion. For instance, SEMScene (Liu et al., 2024) aligns hierarchical multi-level components (entities, triplets, and holistic graphs), while CSCC (Zeng et al., 2021) simultaneously explores local conceptual and global syntactical alignments. Despite their success, the performance of these methods is heavily constrained by the input Visual Scene Graphs (VSGs), which act as a fundamental bottleneck. While Textual Scene Graphs (TSGs) can be parsed directly from captions with negligible variations, constructing a VSG is a highly challenging, two-stage pipeline requiring sequential entity grounding and relation prediction. Consequently, whether the VSG can express the specific semantic focus outlined by the caption strongly influences the performance upper bound of scene graph matching.

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.

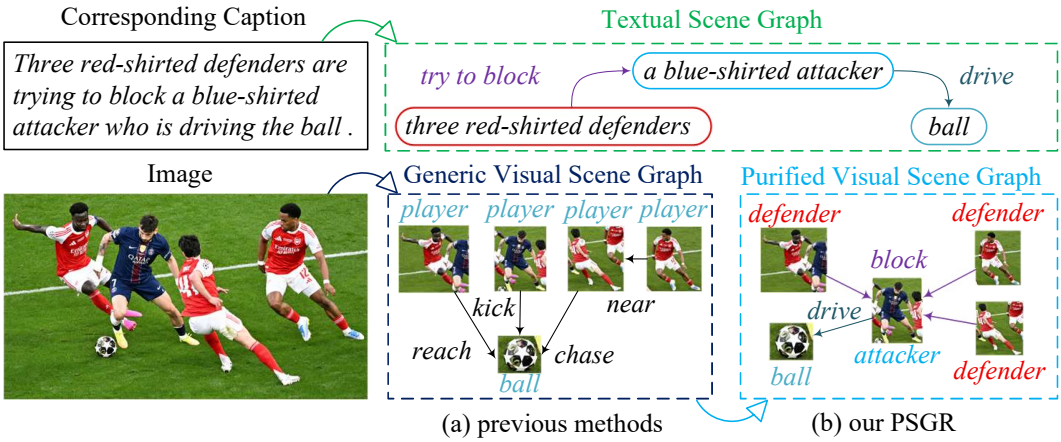

Figure 1: Motivation of our proposed PSGR for ITR. PSGR bridges the semantic gap between generic Visual Scene Graphs (VSGs) and Textual Scene Graphs (TSGs) by refining an overly dense visual graph into a caption-aligned Purified Scene Graph (PSG). Compared with the generic VSG, PSG is more semantically aligned with the TSG extracted from the caption.

As shown in Fig. 1(a), previous methods rely on generic Scene Graph Generation models (SGGs) (Zellers et al., 2018; Tang et al., 2019) to obtain VSGs. However, these generic VSGs are not optimized for caption-level semantic alignment. Although they may correctly ground visible objects, their generic category labels often mismatch the specific semantics in the caption (e.g., generic "player" and "near" vs. caption-specific "defender/attacker" and "block/drive"). To bridge this semantic gap, we propose **PSGR**, a retrieval-oriented scene graph purification framework that **P**urifies the generic visual **S**cene **G**raph to enable image-text **R**etrieval. Specifically, we first curate a Caption-Guided VSG (CG-VSG) dataset using GPT-5 (Singh et al., 2025), and adapt a Large Vision-Language Model (LVLM) via Supervised Fine-Tuning (SFT) and Group Relative Policy Optimization (GRPO) (Shao et al., 2024) to generate Refined Scene Graphs (RSGs). We then devise a Rank-Preserved Candidate Sampling (RPCS) strategy to extract informative triplets from both RSGs and generic VSGs, producing highly optimized Purified Scene Graphs (PSGs). Finally, PSGR combines local entity and relation matching with GPO-based global graph matching, integrating the two similarity scores through min-max normalization. The LVLM refiner is adapted using Flickr30K training supervision and then frozen to generate offline graph inputs for both benchmarks, while the retrieval matchers are trained separately on each benchmark's training split. Under this protocol, detailed in Section 4, PSGR achieves 558.9 Rsum on Flickr30K (Young et al., 2014) and 462.0 Rsum on MS-COCO 5K (Lin et al., 2014). With the backbone and matching architecture fixed, replacing generic VSG top-rank sampling with RPCS-extracted PSGs improves Rsum by 13.0 on Flickr30K and 6.6 on MS-COCO 5K, showing that graph purification contributes gains complementary to the visual backbone. Moreover, PSGR fuses self-supervised DINOv2 (Oquab et al., 2023) region features, which carry less closed-set semantic bias than EfficientNet-B5 (Tan & Le, 2019) features, with explicit semantic label embeddings. This decoupling of visual appearance and geometry from label semantics is better suited to purified scene graph matching, yielding ensemble Rsum gains of 14.4 and 20.0 on Flickr30K and MS-COCO 5K, respectively.

- We formulate scene graph purification as a retrieval-oriented input construction problem and build a GPT-5-generated CG-VSG dataset that provides caption-aligned supervision.
- We adapt an LVLM using SFT to learn caption-aligned scene graph generation and GRPO to further improve the triplet-content fidelity, cross-modal alignment, and structural validity through tailored rewards.
- We propose an RPCS strategy to sample the final PSG as the visual input to the retrieval model, filtering redundant or noisy graph candidates and preserving relation-aware visual semantics that are most beneficial for retrieval.
- PSGR achieves the highest Rsum among the region-based, patch-based, and scene-graph-based baselines compared in our main experiments on Flickr30K Young et al. (2014) and

MS-COCO Lin et al. (2014). The ablation studies in Section 4.3 and additional experiments in Appendix A support the effectiveness of the proposed modules.

2 Background and Related Work

Scene-graph-based image-text retrieval represents images and captions as structured entities and relations, allowing retrieval models to compare visual scene graphs (VSGs) with textual scene graphs (TSGs) beyond isolated region-word alignment. Early and recent methods, such as SGM (Wang et al., 2020b), LGSGM (Nguyen et al., 2021), CSCC (Zeng et al., 2021), SGSIN (Pei et al., 2023), and SEMScene (Liu et al., 2024), mainly improve the matching network over given scene graphs through graph encoders, cross-level consistency, or multi-level semantic matching. These designs demonstrate the utility of scene graphs for retrieval, but they still largely inherit the semantic limitations of the input VSGs.

Other image-text retrieval methods rely on visual regions or patch tokens. Region-based methods, including VSE++ (Faghri et al., 2018), SCAN (Lee et al., 2018), VSRN (Li et al., 2019), SGRAF (Diao et al., 2021), NAAF (Zhang et al., 2022), CHAN (Pan et al., 2023), HREM (Fu et al., 2023), and CORA (Pham et al., 2024), learn fine-grained region-word correspondence or global visual-semantic embeddings. CMSF (Zong et al., 2026) introduces spike-based region-word matching, while UIID (Zhang et al., 2026) distills intra-modal and inter-modal knowledge into compact retrieval encoders. Patch-based methods further exploit transformer patch representations, as in LAPS (Fu et al., 2024), AVSE (Liu et al., 2025b), DSMD (Liang et al., 2025), and D2S-VSE (Liu et al., 2025a). While these methods provide strong dense alignment, their visual inputs are either relation-agnostic regions or unstructured patches, making it difficult to align caption triplets with visual relations. Appendix C provides a detailed comparison.

3 Method

3.1 Caption-Guided LVLM Refinement

As shown in the upper part of Fig 2, PSGR first adapts an LVLM to transform generic detector-produced VSGs into retrieval-oriented Refined Scene Graphs (RSGs). Given an image, its five captions, and a generic VSG, GPT-5 first constructs a Caption Scene Graph (CSG) and then generates caption-guided VSG supervision grounded in the image. This process yields 29K samples from the Flickr30K training set, of which 22K are used for SFT and 7K are selected by output-length bin sampling for GRPO. SFT teaches the LVLM to generate grounded triplets in a parseable response structure, while GRPO further optimizes sampled responses using a weighted combination of three complementary rewards (Eq. 21). The content-order reward combines Hungarian-matched semantic and spatial triplet fidelity with preservation of the target triplet order (Eq. 25); the cross-modal reward averages the best semantic match in the generated RSG for each CSG triplet (Eq. 26); and the format reward combines required boundary-marker compliance with the valid-triplet ratio (Eq. 27). Together, they produce RSG triplets that retain visual grounding while better matching caption-level semantics. Detailed construction and objective definitions are provided in Appendix D.

3.2 Sampling PSGs from RSGs and Generic VSGs

To standardize the variable-size RSGs for downstream feature extraction and graph matching, PSGR constructs fixed-budget visual graph by sampling triplets and entities from the RSG and generic VSG. The RSG provides caption-guided triplet candidates, while the generic VSG serves as a fallback source when the RSG contains insufficient localized relations or entities. The output is a standardized Purified Scene Graph (PSG) with N_E triplets and N_O entities. As shown in the Image Branch of Fig 2, the core objective is to preserve salient triplets while maintaining triplet diversity and suppressing redundant candidates.

Prior to sampling, we perform bounding-box correction for all candidate boxes from both the RSG and generic VSG. Specifically, each bounding box is constrained within the image boundary, and degenerate boxes are rectified to ensure a minimum width and height of one pixel:

$$0 \le x_1 < x_2 \le W, \quad 0 \le y_1 < y_2 \le H, \tag{1}$$

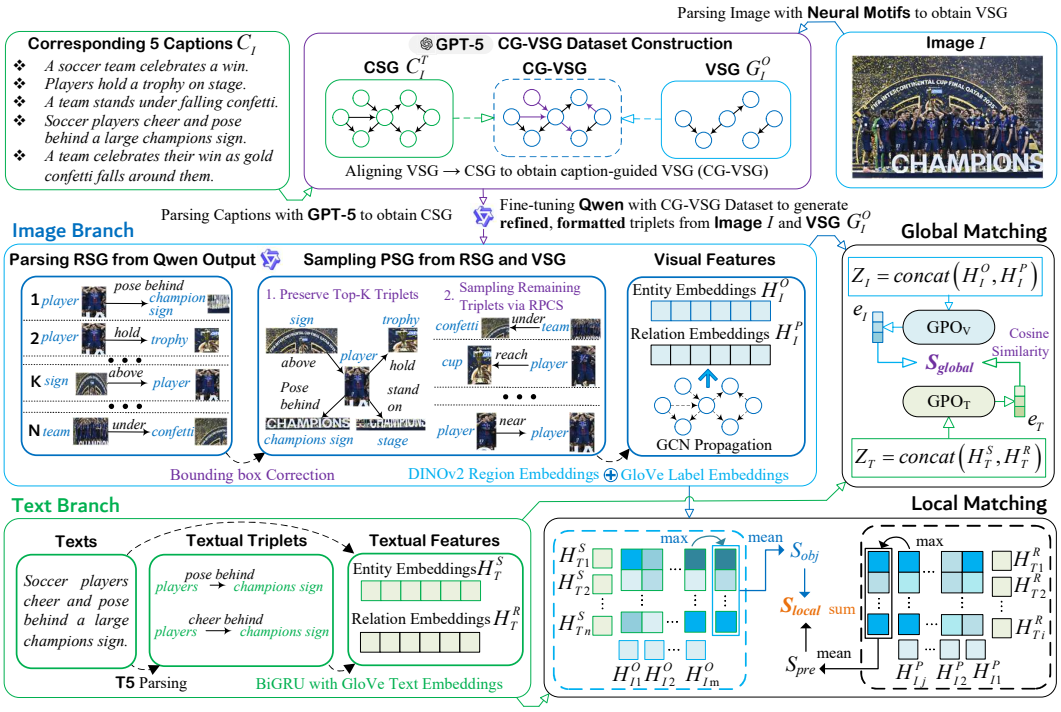

Figure 2: Overall framework of PSGR. The upper pipeline summarizes caption-guided supervision construction and LVLM adaptation, where GPT-5 generates CG-VSG targets used to fine-tune the LVLM for RSG generation via SFT and GRPO. The lower pipeline shows PSG sampling and cross-modal matching: RPCS samples PSGs from RSG and generic VSG candidates, and PSGR combines local matching with global GPO-based graph matching for visual and textual features.

where W and H denote the image width and height. This correction ensures that candidate triplets and entities from both graph sources have valid spatial dimensions before PSG selection.

Let $\hat{E}_I$ and E_I^0 denote the ordered valid triplets of the RSG and generic VSG. PSGR constructs a sequence E_I^P of N_E triplet slots, prioritizing RSG triplets and using generic triplets as fallback:

$$E_I^P = \begin{cases} \hat{E}_I, & |\hat{E}_I| = N_E, \\ \mathrm{Pad}(\hat{E}_I \,\|\, \mathrm{Sample}(E_I^0, N_E - |\hat{E}_I|)), & 0 \le |\hat{E}_I| < N_E, \\ \mathrm{RPCS}(\hat{E}_I, N_E), & |\hat{E}_I| > N_E. \end{cases} \tag{2}$$

Here, $\|$ denotes concatenation. Pad cyclically repeats selected triplets if necessary, or inserts unknown placeholders when none are available. Hence the fixed budget counts slots, not necessarily distinct relations. A missing RSG record uses the first N_E generic triplets before padding. Appendix E gives the boundary handling.

For an under-complete RSG, let $n_f = N_E - |\hat{E}_I|$. If $|E_I^0| > n_f$, fallback sampling first preserves high-ranked generic triplets as anchors:

$$\mathcal{A}_0^f = \mathrm{TopK}(E_I^0, K_f), \quad K_f = \min\{K_0, n_f, \max(1, \lceil n_f/\rho_f \rceil)\}, \tag{3}$$

where K_f is the anchor count. The remaining $n_f - K_f$ slots are sampled without replacement from all unselected generic triplets, using the same diversity-aware weights as RPCS below:

$$\pi_j^{(r)} = \frac{w_j^{(r)}}{\sum_{e_k \in E_I^0 \setminus \mathcal{A}_{r-1}^f} w_k^{(r)}}, \quad e_j \in E_I^0 \setminus \mathcal{A}_{r-1}^f. \tag{4}$$

Weights are computed relative to the selected fallback triplets, independently of the retained RSG. Selected fallback triplets are restored to their original order and appended to the RSG. If $|E_I^0| \le n_f$, all valid generic triplets are retained before padding.

For over-complete RSGs ($|\hat{E}_I| > N_E$), RPCS treats the RSG triplets as the candidate set $C_I^R = \{e_1, e_2, \ldots, e_{|\hat{E}_I|}\}$, where the triplets are ordered by their original generation rank. RPCS first preserves the top-ranked triplets as anchors:

$$\mathcal{A}_0 = \{e_1, e_2, \ldots, e_K\}, \quad K = \min(K_0, N_E). \tag{5}$$

The remaining $N_E - K$ triplets are sampled from $C_I^R \setminus \mathcal{A}_0$. At each sampling step r, for each remaining candidate $e_i \in C_I^R \setminus \mathcal{A}_{r-1}$, we compute:

$$w_i^{(r)} = \exp(-i/\tau) \frac{1 + \lambda_n n_i^{(r)} + \lambda_p p_i^{(r)} + \lambda_t t_i^{(r)}}{1 + \lambda_{sp} d_i^{(r)} + \lambda_{p'} q_i^{(r)}}, \tag{6}$$

and normalize the weights into a sampling probability:

$$P_i^{(r)} = \frac{w_i^{(r)}}{\sum_{e_j \in C_I^R \setminus \mathcal{A}_{r-1}} w_j^{(r)}}. \tag{7}$$

The selected triplet is then added to the current selected set:

$$\mathcal{A}_r = \mathcal{A}_{r-1} \cup \left\{ \text{Sample}\left(C_I^R \setminus \mathcal{A}_{r-1}, P^{(r)} \right) \right\}. \tag{8}$$

After $N_E - K$ sampling steps, RPCS outputs the selected triplets sorted by their generation order:

$$\text{RPCS}(\hat{E}_I, N_E) = \text{Sort}_{\text{rank}}(\mathcal{A}_{N_E - K}). \tag{9}$$

Here, K_0 is the maximum number of directly preserved anchors, and τ controls how strongly RPCS trusts the original generation rank. In the numerator, $p_i^{(r)}$ and $t_i^{(r)}$ indicate whether candidate e_i introduces a new relation label or a new coarse relation type with respect to the currently selected set $\mathcal{A}_{r-1}$. In the denominator, $d_i^{(r)}$ counts selected triplets with the same subject-relation pair, and $q_i^{(r)}$ counts selected triplets with the same relation label, thereby penalizing redundant relation patterns.

The node-coverage term $n_i^{(r)}$ counts the candidate's subject and object endpoints not yet covered by selected triplets. It encourages new entity coverage alongside relation diversity; selected triplets retain their original generation order.

Entity sampling is then performed from the RSG and generic VSG entity pools based on the selected triplets. PSGR first adds all entities used by E_I^P and assigns new IDs according to their first-appearance order. It then fills the remaining slots with unused RSG entities and, if necessary, generic VSG entities until the PSG contains N_O entities. If more than N_O entities are available, all entities involved in the selected triplets are retained, and the remaining entities are truncated. Finally, all triplet subjects and objects are remapped to the new entity IDs, ensuring that every PSG triplet is consistent with the final entity list.

3.3 Local and Global Matching for PSGs and TSGs

3.3.1 Feature Extraction

For each PSG, DINOv2 (Oquab et al., 2023) independently encodes entity-box crops and subject–object union-box crops for relations. These crop-level features are fused with entity and relation label embeddings, and a GCN then propagates graph context:

$$(\mathbf{H}_I^o, \mathbf{H}_I^p) = \text{GCN}(\mathbf{O}_I, \mathbf{P}_I, \mathbf{A}_I), \tag{10}$$

where $\mathbf{A}_I$ is the subject-object connection structure, and $\mathbf{H}_I^o$ and $\mathbf{H}_I^p$ are the final entity and relation embeddings. For text, caption words and TSG triplets parsed by FACTUAL-T5 (Li et al., 2023c) are encoded with BiGRU (Cho et al., 2014) and GloVe (Pennington et al., 2014), producing entity states $\mathbf{H}_T^s$ and relation states $\mathbf{H}_T^r$.

3.3.2 FINE-GRAINED LOCAL MATCHING

The local branch computes text-to-image cross attention between a textual sequence $\mathbf{Q} = \{\mathbf{q}_j\}_{j=1}^{n}$ and a visual sequence $\mathbf{K} = \{\mathbf{k}_i\}_{i=1}^{m}$:

$$\begin{aligned} \mathbf{c}_j &= \sum_{i=1}^{m} a_{ji}\mathbf{k}_i, \quad a_{ji} = \frac{\exp(\kappa\sigma(\mathbf{q}_j, \mathbf{k}_i))}{\sum_{\ell=1}^{m}\exp(\kappa\sigma(\mathbf{q}_j, \mathbf{k}_\ell))}, \\ \mathrm{XAttn}(\mathbf{Q}, \mathbf{K}) &= \frac{1}{n}\sum_{j=1}^{n}\sigma(\mathbf{q}_j, \mathbf{c}_j). \end{aligned} \tag{11}$$

Here, κ is the attention temperature and $\sigma(\cdot,\cdot)$ is cosine similarity. Each textual element attends to all visual elements to form $\mathbf{c}_j$. We compute entity-level and relation-level correspondences separately:

$$\begin{aligned} S_{\mathrm{ent}} &= \mathrm{XAttn}(\mathbf{H}_T^s, \mathbf{H}_I^o), \\ S_{\mathrm{rel}} &= \mathrm{XAttn}(\mathbf{H}_T^r, \mathbf{H}_I^p), \\ S_{\mathrm{local}} &= S_{\mathrm{ent}} + S_{\mathrm{rel}}. \end{aligned} \tag{12}$$

Here, S_{ent} captures word-to-entity region correspondence, while S_{rel} compares TSG relations with PSG relation regions; their sum retains both object and relational evidence.

3.3.3 GENERALIZED GLOBAL MATCHING

The global branch measures holistic graph consistency by aggregating image-side and text-side graph sequences with GPO (Chen et al., 2021):

$$\begin{aligned} \mathbf{Z}_I &= \{\mathbf{z}_{I,n}\}_{n=1}^{N_I}, \quad \mathbf{Z}_T = \{\mathbf{z}_{T,m}\}_{m=1}^{N_T}, \\ \mathbf{v}_I &= f_{\mathrm{GPO}}(\mathbf{Z}_I), \quad \mathbf{v}_T = f_{\mathrm{GPO}}(\mathbf{Z}_T). \end{aligned} \tag{13}$$

Image inputs concatenate PSG entities and relations, while text inputs combine caption states and TSG relations. GPO pools each feature dimension c as $[f_{\mathrm{GPO}}(\mathbf{Z})]_c = \sum_{n=1}^{N}\theta_n z_{n,c}^{\downarrow}$, where $z_{n,c}^{\downarrow}$ is the n-th largest valid value in that dimension. Weights derive from positional encodings of sorted ranks, not node identities, making pooling permutation-invariant over valid embeddings. Appendix E specifies weight generation and padding. The global score is $S_{\mathrm{global}} = \sigma(\mathbf{v}_I, \mathbf{v}_T)$.

During inference, we ensemble the local and global branches by min-max normalization:

$$S = \mathcal{N}(S_{\mathrm{local}}) + \mathcal{N}(S_{\mathrm{global}}), \quad \mathcal{N}(S) = \frac{S - \min(S)}{\max(S) - \min(S) + \epsilon}. \tag{14}$$

Here, ϵ is a small constant for numerical stability.

4 EXPERIMENTS

4.1 DATASETS AND EVALUATION METRICS

We conduct experiments on Flickr30K (Young et al., 2014) and MS-COCO (Lin et al., 2014), two standard image-text retrieval datasets where each image has five captions. Flickr30K uses 29K/1K/1K images for training/validation/testing, while MS-COCO uses the standard 113,287/5,000/5,000 split. Following prior work (Chen et al., 2021), we report MS-COCO results under both the 5-fold 1K and 5K test settings.

We report Recall@K (R@K, $K \in \{1, 5, 10\}$) for both Image→Text and Text→Image retrieval. Rsum is the sum of the six displayed recall scores and serves as the overall metric. Additional implementation details and the LVLM prompt and output format are provided in Appendix E.

Training and data usage protocol. The CG-VSG supervision used to adapt the LVLM consists of 29,000 samples from the Flickr30K training set: 22,000 for SFT and 7,000 for GRPO. After adaptation, the LVLM is frozen and used offline to generate RSG inputs for both Flickr30K and MS-COCO. The local and global retrieval matchers are then trained independently on the training split of each retrieval benchmark. Consequently, the MS-COCO results use external Flickr30K supervision at the graph-refinement stage, and the MS-COCO matchers use only the MS-COCO training set. The cross-dataset experiment in Appendix A.7 isolates transfer of the retrieval matchers and is not an end-to-end zero-shot evaluation.

Table 1: Quantitative results on Flickr30K and MS-COCO 5-fold 1K. **Bold** and underlined values denote the best and second-best results, respectively. "–" indicates an unreported result.

Method	Flickr30K							MS-COCO 5-fold 1K						
	Image→Text			Text→Image			Rsum	Image→Text			Text→Image			Rsum
	R@1	R@5	R@10	R@1	R@5	R@10		R@1	R@5	R@10	R@1	R@5	R@10	
Region-based Matching Methods														
CODER (Wang et al., 2022)	83.2	96.5	98.0	63.1	87.1	93.0	520.9	82.1	96.6	98.8	65.5	91.5	96.2	530.7
MV-VSE (Li et al., 2022b)	82.1	95.8	97.9	63.1	86.7	92.3	517.9	80.4	96.6	99.0	64.9	91.2	96.0	528.1
HREM (Fu et al., 2023)	84.0	96.1	98.6	64.4	88.0	93.1	524.2	**82.9**	96.9	99.0	67.1	92.0	96.6	534.5
RVSE++ (Li et al., 2025a)	83.6	96.5	98.6	64.3	88.2	93.0	524.2	81.6	96.6	98.8	66.6	92.1	96.6	532.3
CHAN (Pan et al., 2023)	80.6	96.1	97.8	63.9	87.5	92.6	518.5	81.4	96.9	98.9	66.5	92.1	96.7	532.5
SDE (Kim et al., 2023)	80.9	94.7	97.6	59.4	85.6	91.1	509.3	–	–	–	–	–	–	–
CORA (Pham et al., 2024)	83.4	95.9	98.6	64.1	88.1	93.1	523.2	82.8	97.3	99.0	67.3	92.4	96.9	535.7
CMSF (Zong et al., 2026)	80.7	95.0	97.6	61.3	85.9	91.3	511.8	78.9	96.2	98.7	63.6	91.0	96.3	524.7
UIID (VSRN) (Zhang et al., 2026)	67.9	89.2	94.2	56.0	86.4	92.7	486.4	72.5	94.2	97.8	64.3	91.9	95.6	516.3
Patch-based Matching Methods														
LAPS (Fu et al., 2024)	74.0	93.4	97.4	62.5	87.3	92.7	507.3	78.7	95.5	98.3	66.2	91.3	96.2	526.2
AVSE (Liu et al., 2025b)	76.0	94.6	97.5	62.7	88.4	93.1	512.3	79.8	95.6	98.3	67.0	91.5	96.3	528.5
DSMD (Liang et al., 2025)	82.0	95.5	97.7	68.4	90.8	94.4	528.8	–	–	–	–	–	–	–
D2S-VSE (Liu et al., 2025a)	82.8	96.1	98.3	68.5	91.3	94.9	531.9	80.1	97.0	99.2	68.1	92.5	96.7	533.6
Scene Graph Matching Methods														
SGM (Wang et al., 2020b)	71.8	91.7	95.5	53.5	79.6	86.5	478.6	–	–	–	–	–	–	–
LGSGM (Nguyen et al., 2021)	71.0	91.9	96.1	57.4	84.1	90.2	490.7	–	–	–	–	–	–	–
CSCC (Zeng et al., 2021)	72.7	93.4	96.5	61.2	86.7	91.5	502.0	78.8	96.1	99.0	66.6	92.5	96.4	529.4
SGSIN (Pei et al., 2023)	73.1	93.6	96.8	53.9	80.1	87.2	484.7	76.7	96.5	99.1	61.7	89.6	95.3	518.9
SEMScene (Liu et al., 2024)	77.8	95.1	97.7	63.2	87.6	92.6	514.0	76.2	96.9	98.9	64.7	91.8	96.4	524.9
PSGR (ours)	**90.3**	**99.0**	**99.5**	**78.1**	**94.7**	**97.3**	**558.9**	82.7	**97.5**	**99.3**	**71.0**	**93.3**	**97.2**	**541.0**

4.2 Main Results

Table 1 compares PSGR with representative region-based, patch-based, and scene-graph-based methods. On Flickr30K, PSGR achieves the best results across all six recall metrics. For Image→Text retrieval, it obtains 90.3/99.0/99.5 at R@1/R@5/R@10, exceeding the strongest competing values by 6.3/2.5/0.9 points. The gains are more pronounced for Text→Image retrieval: PSGR reaches 78.1/94.7/97.3 and improves over the previous best results by 9.6/3.4/2.4 points. Consequently, its Rsum of 558.9 surpasses the strongest baseline, D2S-VSE (Liu et al., 2025a), by 27.0 points. Among scene-graph-based methods specifically, PSGR exceeds the best baseline Rsum of 514.0 from SEMScene (Liu et al., 2024) by 44.9 points.

On MS-COCO 5-fold 1K, PSGR achieves the best result on five of the six recall metrics and obtains 541.0 Rsum, 5.3 points above CORA (Pham et al., 2024). Its Image→Text R@1 of 82.7 is 0.2 points below HREM (Fu et al., 2023), whereas its R@5/R@10 values of 97.5/99.3 improve on the previous best by 0.2/0.1 points. For Text→Image retrieval, PSGR obtains 71.0/93.3/97.2, yielding gains of 2.9/0.8/0.3 points over the strongest competing results. It also improves upon the strongest scene-graph baseline, CSCC (Zeng et al., 2021), by 11.6 Rsum points.

The MS-COCO 5K results in Table 4 of Appendix A.1 exhibit the same pattern. PSGR attains 63.9/88.3/94.4 for Image→Text and 50.7/78.0/86.7 for Text→Image, with particularly clear gains of 2.7/2.1/1.5 points on the three Text→Image metrics. Its Rsum reaches 462.0, outperforming HREM by 11.0 points. These MS-COCO comparisons should nevertheless be interpreted under the disclosed protocol: the frozen RSG generator was adapted with Flickr30K supervision rather than under a strictly in-dataset setting. The controlled ablations below therefore provide more direct evidence for the contribution of graph purification.

Table 2: Ablation study on LVLM fine-tuning strategies. Best results are bold. I2T and T2I respectively denotes the Image→Text and Text→Image retrievals.

Strategy	Branch	Flickr30K			MS-COCO			
		I2T R@1	T2I R@1	Rsum	5K I2T R@1	5K T2I R@1	1K Rsum	5K Rsum
SFT only	Local	82.5	72.5	540.8	56.6	47.3	532.6	442.4
	Global	85.7	73.4	544.2	60.0	46.0	533.7	445.3
	Ensemble	88.4	76.9	554.7	62.3	49.8	539.4	457.7
SFT+GRPO w/o r_{cross}	Local	81.8	71.7	541.0	57.4	47.2	533.0	443.8
	Global	84.4	72.1	541.7	59.6	45.7	531.7	443.9
	Ensemble	87.7	76.0	553.3	62.4	49.9	540.2	458.1
Full PSGR	Local	85.9	73.9	547.9	58.1	47.4	533.2	445.3
	Global	87.4	74.3	548.7	61.3	46.9	533.6	448.9
	Ensemble	**90.3**	**78.1**	**558.9**	**63.9**	**50.7**	**541.0**	**462.0**

Table 3: Ablation study on RPCS with the DINOv2 backbone and retrieval matcher fixed across all variants. Best results are bold.

Strategy	Flickr30K			MS-COCO			
	I2T R@1	T2I R@1	Rsum	5K I2T R@1	5K T2I R@1	1K Rsum	5K Rsum
VSG Top-rank	84.2	73.1	545.9	61.2	49.0	539.0	455.4
RSG Top-rank	89.3	77.8	557.4	63.2	50.5	540.3	460.3
w/o diversity	89.1	76.8	556.2	62.8	50.4	540.0	459.4
w/o redundancy	89.5	77.3	557.0	63.4	50.5	540.4	460.8
Full RPCS	**90.3**	**78.1**	**558.9**	**63.9**	**50.7**	**541.0**	**462.0**

4.3 Ablation Studies

4.3.1 Fine-tuning Strategies

We compare SFT only, GRPO without the cross-modal reward, and the full objective. SFT teaches the grounded graph format. GRPO without r_{cross} optimizes the content-order reward (Eq. 25), combining semantic and spatial fidelity with target-order preservation, and the format reward (Eq. 27), scoring boundary markers and triplet validity. The full objective adds the cross-modal reward (Eq. 26), which averages the best generated-triplet match for each caption triplet to encourage caption-level semantic alignment.

Table 2 shows that omitting this caption-level signal yields mixed results: compared with SFT only, ensemble Rsum decreases by 1.4 on Flickr30K but increases by 0.8 and 0.4 on MS-COCO 5-fold 1K and 5K. Optimizing grounded content, order, and format alone therefore does not consistently improve retrieval. The full objective improves ensemble Rsum over SFT only by 4.2, 1.6, and 4.3 on the three benchmarks, respectively, and over GRPO without r_{cross} by 5.6, 0.8, and 3.9. These results support caption-level alignment as an effective signal for retrieval-oriented graph refinement.

4.3.2 Sampling Strategies

We further analyze RPCS by comparing generic VSG top-rank sampling, RSG top-rank sampling, and two variants that remove either relation diversity or redundancy suppression. All variants use the same DINOv2 backbone, matching architecture, and training settings, so their differences isolate graph refinement and sampling rather than visual feature extraction. VSG Top-rank uses detector-generated relations, whereas RSG Top-rank tests LVLM refinement without RPCS. The two removal variants isolate relation diversity and redundancy suppression.

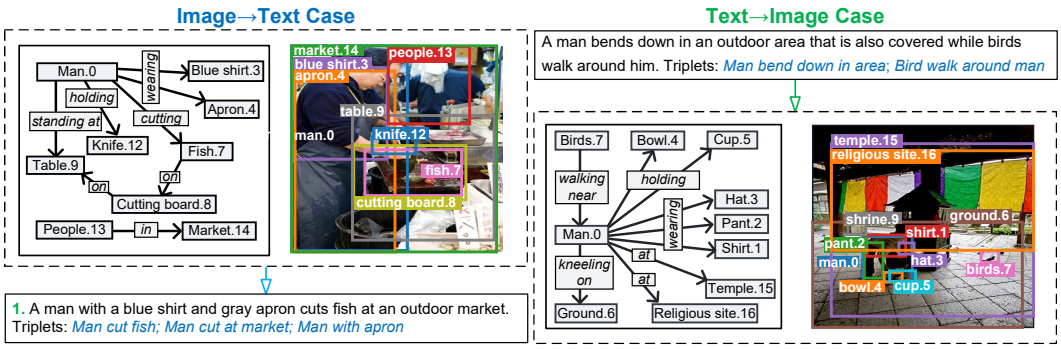

Figure 3: Qualitative retrieval examples of PSGR. PSG highlights caption-aligned entities and relations that support both Image→Text and Text→Image retrieval.

Table 3 separates the two purification steps under a matched backbone. Replacing generic VSG top-rank sampling with RSG top-rank sampling improves Rsum by 11.5 on Flickr30K and 4.9 on MS-COCO 5K, while RPCS adds a further 1.5 and 1.7 points, respectively. The total gain from generic VSG top-rank sampling to full RPCS is therefore 13.0 on Flickr30K and 6.6 on MS-COCO 5K. Removing relation diversity or redundancy suppression also reduces performance, supporting their roles in preventing repetitive candidates.

Table 5 further examines the visual backbone. Replacing EfficientNet-B5 with DINOv2 in PSGR improves ensemble Rsum by 14.4 on Flickr30K and 20.0 on MS-COCO 5K. With the same DINOv2 backbone and matching architecture, LVLM refinement and RPCS together improve Rsum by 13.0 and 6.6 over generic VSG top-rank sampling, respectively.

Additional ablations, parameter sensitivity analyses, efficiency measurements, and dataset statistics are presented in Appendix A, with qualitative examples in Appendix B.

4.4 PSGR Bidirectional Retrieval Cases

Fig. 3 shows that PSGR retrieves bidirectional matches by preserving compact, caption-aligned graph evidence. In the Image→Text case, the PSG emphasizes discriminative entities and relations that make the retrieved captions share consistent triplet semantics. In the Text→Image case, the query TSG is matched to the target image through key entities, relations, and scene context. For example, the Image→Text case preserves visual evidence around the man, apron, knife, fish, cutting board, and market, so the retrieved captions focus on consistent actions such as cutting fish and preparing food. The Text→Image case similarly keeps the man, birds, ground, and religious-site context needed by the query, allowing PSGR to distinguish the correct image from visually related candidates. This behavior is consistent with the quantitative gains in Text→Image retrieval, where the query caption benefits more directly from a compact visual graph that preserves the relations it asks for. These examples show that PSGR supports retrieval by concentrating on caption-aligned graph evidence while suppressing dense but weakly relevant visual semantics.

5 Conclusion

In this paper, we presented PSGR for retrieval-oriented scene graph purification. With tailored rewards, PSGR adapts an LVLM toward caption-aligned graph refinement and then proposes RPCS to select informative candidates before local and global graph matching. Experiments on Flickr30K and MS-COCO show strong retrieval performance under the disclosed use of a Flickr30K-adapted frozen refiner. Backbone ablations show the benefit of DINOv2 features, while LVLM refinement and RPCS provide additional gains when the backbone and matcher are fixed. These results support graph purification as a complementary direction for structured image-text retrieval, while the cross-dataset analysis should not be interpreted as end-to-end zero-shot generalization.

REFERENCES

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pp. 6077–6086, 2018.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Ming-Hsuan Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Jianwei Cao, Yanli Hu, Zhen Tan, and Xiang Zhao. Cross-modal multi-task learning for multimedia event extraction. In *Proc. AAAI Conf. Artif. Intell.*, pp. 11454–11462, 2025.

Jiacheng Chen, Hexiang Hu, Hao Wu, Yuning Jiang, and Changhu Wang. Learning the best pooling strategy for visual semantic embedding. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pp. 15789–15798, 2021.

Zuyao Chen, Jinlin Wu, Zhen Lei, Zhaoxiang Zhang, and Chang Wen Chen. Expanding scene graph boundaries: Fully open-vocabulary scene graph generation via visual-concept alignment and retention. In *European conference on computer vision*, pp. 108–124. Springer, 2024.

Zuyao Chen, Jinlin Wu, Zhen Lei, Marc Pollefeys, and Chang Wen Chen. Compile scene graphs with reinforcement learning. *arXiv preprint arXiv:2504.13617*, 2025.

Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Proc. Conf. Empirical Methods Natural Language Process.*, pp. 1724–1734, 2014.

Yuren Cong, Michael Ying Yang, and Bodo Rosenhahn. Reltr: Relation transformer for scene graph generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):11169–11183, 2023.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, 2019. doi: 10.18653/v1/N19-1423.

Haiwen Diao, Ying Zhang, Lin Ma, and Huchuan Lu. Similarity reasoning and filtration for image-text matching. In *Proc. AAAI Conf. Artif. Intell.*, pp. 1218–1226, 2021.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proc. Int. Conf. Learn. Represent.*, 2021.

Fartash Faghri, David J. Fleet, Jamie Ryan Kiros, and Sanja Fidler. VSE++: improving visual-semantic embeddings with hard negatives. In *Proc. Brit. Mach. Vis. Conf.*, pp. 12, 2018.

Jiaye Feng, Qixiang Yin, Yuankun Liu, Tong Mo, and Weiping Li. Sgg-r 3: From next-token prediction to end-to-end unbiased scene graph generation. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 19811–19830, 2026.

Zheren Fu, Zhendong Mao, Yan Song, and Yongdong Zhang. Learning semantic relationship among instances for image-text matching. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pp. 15159–15168, 2023.

Zheren Fu, Lei Zhang, Hou Xia, and Zhendong Mao. Linguistic-aware patch slimming framework for fine-grained cross-modal alignment. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pp. 26297–26306, 2024.

Byeong Su Kim, Jieun Kim, Deokwoo Lee, and Beakcheol Jang. Visual question answering: A survey of methods, datasets, evaluation, and challenges. *ACM Comput. Surv.*, 57(10):249:1–249:35, 2025.

Dongwon Kim, Namyup Kim, and Suha Kwak. Improving cross-modal retrieval with set of diverse embeddings. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pp. 23422–23431, 2023.

Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

Jiayi Kuang, Ying Shen, Jingyou Xie, Haohao Luo, Zhe Xu, Ronghao Li, Yinghui Li, Xianfeng Cheng, Xika Lin, and Yu Han. Natural language understanding and inference with MLLM in visual question answering: A survey. *ACM Comput. Surv.*, 57(8):190:1–190:36, 2025.

Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *Proc. Eur. Conf. Comput. Vis.*, pp. 212–228, 2018.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 12888–12900, 2022a.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 19730–19742, 2023a.

Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Fu. Visual semantic reasoning for image-text matching. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pp. 4653–4661, 2019.

Lin Li, Wei Chen, Jiahui Li, Kwang-Ting Cheng, and Long Chen. Relation-r1: Progressively cognitive chain-of-thought guided reinforcement learning for unified relation comprehension. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 6306–6314, 2026.

Rongjie Li, Songyang Zhang, and Xuming He. Sgtr+: End-to-end scene graph generation with transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(4):2191–2205, 2023b.

Rongjie Li, Songyang Zhang, Dahua Lin, Kai Chen, and Xuming He. From pixels to graphs: Open-vocabulary scene graph generation with vision-language models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28076–28086. IEEE, 2024.

Yongzhi Li, Duo Zhang, and Yadong Mu. Visual-semantic matching by exploring high-order attention and distraction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Zheng Li, Caili Guo, Zerun Feng, Jenq-Neng Hwang, and Xijun Xue. Multi-view visual semantic embedding. In *Proc. Int. Joint Conf. Artif. Intell.*, pp. 1130–1136, 2022b.

Zheng Li, Caili Guo, Xin Wang, Zerun Feng, and Zhongtian Du. Selectively hard negative mining for alleviating gradient vanishing in image-text matching. *IEEE Trans. Circuits Syst. Video Technol.*, 35(2):1921–1935, 2025a.

Zheng Li, Caili Guo, Xin Wang, Hao Zhang, and Lin Hu. Multi-view visual semantic embedding for cross-modal image-text retrieval. *Pattern Recognit.*, 159:111088, 2025b.

Zhuang Li, Yuyang Chai, Terry Yue Zhuo, Lizhen Qu, Gholamreza Haffari, Fei Li, Donghong Ji, and Quan Hung Tran. FACTUAL: A benchmark for faithful and consistent textual scene graph parsing. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 6377–6390, Toronto, Canada, July 2023c. Association for Computational Linguistics.

Zhengyang Liang, Meiyu Liang, Wei Huang, Yawen Li, Wu Liu, Yingxia Shao, and Kangkang Lu. Dynamic self-adaptive multiscale distillation from pre-trained multimodal large model for efficient cross-modal retrieval. In *Proc. ACM Int. Conf. Multimedia*, pp. 2851–2859, 2025.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In *Proc. Eur. Conf. Comput. Vis.*, pp. 740–755, 2014.

Chunxiao Liu, Zhendong Mao, Tianzhu Zhang, Hongtao Xie, Bin Wang, and Yongdong Zhang. Graph structured network for image-text matching. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pp. 10918–10927, 2020.

Yang Liu, Wentao Feng, Zhuoyao Liu, Shudong Huang, and Jiancheng Lv. Aligning information capacity between vision and language via dense-to-sparse feature distillation for image-text matching. In *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, pp. 21679–21688, 2025a.

Yang Liu, Mengyuan Liu, Shudong Huang, and Jiancheng Lv. Asymmetric visual semantic embedding framework for efficient vision-language alignment. In *Proc. AAAI Conf. Artif. Intell.*, pp. 5676–5684, 2025b.

Yuankun Liu, Xiang Yuan, Haochen Li, Zhijie Tan, Jinsong Huang, Jingjie Xiao, Weiping Li, and Tong Mo. Semscene: Semantic-consistency enhanced multi-level scene graph matching for image-text retrieval. *ACM Trans. Multim. Comput. Commun. Appl.*, 20(8):261:1–261:28, 2024.

Siqu Long, Soyeon Caren Han, Xiaojun Wan, and Josiah Poon. Gradual: Graph-based dual-modal representation for image-text matching. In *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, pp. 2463–2472, 2022.

Manh-Duy Nguyen, Binh T Nguyen, and Cathal Gurrin. A deep local and global scene-graph matching for image-text retrieval. *arXiv preprint arXiv:2106.02400*, 2021.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Zhengxin Pan, Fangyu Wu, and Bailing Zhang. Fine-grained image-text matching by cross-modal hard aligning network. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pp. 19275–19284, 2023.

Jiaming Pei, Kaiyang Zhong, Zhi Yu, Lukun Wang, and Kuruva Lakshmanna. Scene graph semantic inference for image and text matching. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 22 (5):1–23, 2023.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Proc. Conf. Empirical Methods Natural Language Process.*, pp. 1532–1543, 2014.

Khoi Pham, Chuong Huynh, Ser-Nam Lim, and Abhinav Shrivastava. Composing object relations and attributes for image-text matching. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pp. 14354–14363, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 8748–8763, 2021.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pp. 6105–6114. PMLR, 2019.

Kaihua Tang, Hanwang Zhang, Baoyuan Wu, Wenhan Luo, and Wei Liu. Learning to compose dynamic tree structures for visual contexts. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6612–6621. IEEE, 2019.

Yao Teng and Limin Wang. Structured sparse r-cnn for direct scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19437–19446, 2022.

Haoran Wang, Ying Zhang, Zhong Ji, Yanwei Pang, and Lin Ma. Consensus-aware visual-semantic embedding for image-text matching. In *Proc. Eur. Conf. Comput. Vis.*, pp. 18–34, 2020a. doi: 10.1007/978-3-030-58586-0_2.

Haoran Wang, Dongliang He, Wenhao Wu, Boyang Xia, Min Yang, Fu Li, Yunlong Yu, Zhong Ji, Errui Ding, and Jingdong Wang. CODER: coupled diversity-sensitive momentum contrastive learning for image-text retrieval. In *Proc. Eur. Conf. Comput. Vis.*, pp. 700–716, 2022.

Sijin Wang, Ruiping Wang, Ziwei Yao, Shiguang Shan, and Xilin Chen. Cross-modal scene graph matching for relationship-aware image-text retrieval. In *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, pp. 1497–1506, 2020b.

Hao Wu, Jiayuan Mao, Yufeng Zhang, Yuning Jiang, Lei Li, Weiwei Sun, and Wei-Ying Ma. Unified visual-semantic embeddings: Bridging vision and language with structured meaning representations. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pp. 6609–6618, 2019.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Trans. Assoc. Comput. Linguistics*, 2:67–78, 2014.

Xiang Yuan, Xinrong Chen, Haochen Li, Hang Yang, Guanyu Wang, Weiping Li, and Tong Mo. Stepwise schema-guided prompting framework with parameter efficient instruction tuning for multimedia event extraction. In *Proc. IEEE Int. Conf. Multimedia Expo*, pp. 1–6, 2025.

Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. Neural motifs: Scene graph parsing with global context. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pp. 5831–5840, 2018.

Pengpeng Zeng, Lianli Gao, Xinyu Lyu, Shuaiqi Jing, and Jingkuan Song. Conceptual and syntactical cross-modal alignment with cross-level consistency for image-text matching. In *Proc. ACM Int. Conf. Multimedia*, pp. 2205–2213, 2021.

Yan Zeng, Xinsong Zhang, and Hang Li. Multi-grained vision language pre-training: Aligning texts with visual concepts. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 25994–26009, 2022.

Kun Zhang, Zhendong Mao, Quan Wang, and Yongdong Zhang. Negative-aware attention framework for image-text matching. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pp. 15640–15649, 2022.

Qi Zhang, Zhen Lei, Zhaoxiang Zhang, and Stan Z. Li. Context-aware attention network for image-text retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Song Zhang, Nan Zheng, Liang Zhang, Xuemei Jiang, and Yifeng Jin. UIID: Unified intra-modal and inter-modal distillation for image-text retrieval. *Neurocomputing*, 683:133504, 2026. doi: 10.1016/j.neucom.2026.133504.

Yuze Zhao, Jintao Huang, Jinghan Hu, Xingjun Wang, Yunlin Mao, Daoze Zhang, Zeyinzi Jiang, Zhikai Wu, Baole Ai, Ang Wang, Wenmeng Zhou, and Yingda Chen. SWIFT: A scalable lightweight infrastructure for fine-tuning. In *Proc. AAAI Conf. Artif. Intell.*, pp. 29733–29735, 2025.

Xintao Zong, Xian Zhong, Wenxuan Liu, Jianhao Ding, Zhaofei Yu, and Tiejun Huang. Brain-inspired multimodal spiking neural network for image-text retrieval. *arXiv preprint arXiv:2603.26787*, 2026.

Table 4: Quantitative results on MS-COCO under the 5K test setting. **Bold** and underlined values denote the best and second-best results, respectively.

Method	Image→Text			Text→Image			Rsum
	R@1	R@5	R@10	R@1	R@5	R@10	
Region-based Matching Methods							
VSE∞ (Chen et al., 2021)	58.3	85.3	92.3	42.4	72.7	83.2	434.2
SCAN (Lee et al., 2018)	50.4	82.2	90.0	38.6	69.3	80.4	410.9
VSRN (Li et al., 2019)	53.0	81.1	89.4	40.5	70.6	81.1	415.7
GCN+DIST (Li et al., 2020)	51.4	81.8	89.1	40.5	73.5	84.1	420.4
CAAN (Zhang et al., 2020)	52.5	83.3	90.9	41.2	70.3	82.9	421.1
CODER (Wang et al., 2022)	62.6	86.6	93.1	42.5	73.1	83.3	441.2
MV-VSE (Li et al., 2022b)	59.1	86.3	92.5	42.5	72.8	83.1	436.3
HREM (Fu et al., 2023)	<u>64.0</u>	**88.5**	<u>93.7</u>	45.4	75.1	84.3	<u>451.0</u>
RVSE++ (Li et al., 2025a)	60.6	86.4	92.8	44.5	74.5	84.5	443.3
CHAN (Pan et al., 2023)	59.8	87.2	93.3	44.9	74.5	84.2	443.9
CORA (Pham et al., 2024)	**64.3**	87.5	93.6	45.4	74.7	84.6	450.1
Patch-based Matching Methods							
LAPS (Fu et al., 2024)	57.5	84.0	90.8	44.5	74.0	83.6	434.4
AVSE (Liu et al., 2025b)	58.8	84.3	91.0	45.1	74.3	83.9	437.4
DSMD (Liang et al., 2025)	62.1	85.9	92.0	<u>48.0</u>	75.6	84.5	448.1
D2S-VSE (Liu et al., 2025a)	60.1	85.5	92.5	46.3	<u>75.9</u>	<u>85.2</u>	445.5
Scene Graph Matching Methods							
CSCC (Zeng et al., 2021)	55.6	83.6	91.2	40.8	73.2	84.3	428.7
SEMScene (Liu et al., 2024)	54.3	81.9	89.7	41.2	71.3	82.2	420.6
PSGR (ours)	63.9	<u>88.3</u>	**94.4**	**50.7**	**78.0**	**86.7**	**462.0**

A Additional Experiments and Analyses

A.1 MS-COCO 5K Main Results

Table 4 reports the MS-COCO 5K evaluation using methods with available results. PSGR obtains 462.0 Rsum, outperforming the strongest baseline, HREM (Fu et al., 2023), by 11.0 points. For Image→Text retrieval, PSGR reaches 63.9/88.3/94.4 at R@1/R@5/R@10. Its R@1 and R@5 are only 0.4 and 0.2 points below CORA (Pham et al., 2024) and HREM, respectively, while its R@10 exceeds the previous best by 0.7 points. PSGR shows clearer advantages for Text→Image retrieval, achieving 50.7/78.0/86.7 and improving on the strongest competing results by 2.7/2.1/1.5 points. Relative to scene-graph-based methods, it surpasses CSCC (Zeng et al., 2021) by 33.3 Rsum points, indicating that the refined and sampled graph representation remains effective as the retrieval pool becomes substantially larger.

These results demonstrate a stronger bidirectional balance than methods that are competitive mainly in Image→Text retrieval. The frozen RSG generator uses Flickr30K supervision. Table 3 evaluates graph purification using the same DINOv2 backbone. Summed within each retrieval direction, PSGR obtains 246.6 for Image→Text and 215.4 for Text→Image. The former is only 0.4 points above HREM, whereas the latter is 7.3 points above the strongest competing directional sum from DSMD (Liang et al., 2025). Thus, the overall advantage on the larger 5K retrieval pool primarily comes from more reliable image ranking for textual queries rather than a uniform increase across both directions.

Table 5: Ablation study on visual feature extraction. Best results are bold.

Branch	Visual Extractor	Flickr30K			MS-COCO			
		I2T R@1	T2I R@1	Rsum	5K I2T R@1	5K T2I R@1	1K Rsum	5K Rsum
Local	EfficientNet-B5	81.3	67.6	526.6	55.6	42.8	520.9	426.3
	DINOv2 ViT-g/14	85.9	73.9	547.9	58.1	47.4	533.2	445.3
Global	EfficientNet-B5	83.5	69.8	534.1	56.6	40.8	520.6	423.7
	DINOv2 ViT-g/14	87.4	74.3	548.7	61.3	46.9	533.6	448.9
Ensemble	EfficientNet-B5	86.0	73.4	544.5	60.2	45.3	531.0	442.0
	DINOv2 ViT-g/14	**90.3**	**78.1**	**558.9**	**63.9**	**50.7**	**541.0**	**462.0**

Table 6: Ablation study on textual scene graph parsing. Best results are bold.

Branch	Parser	Flickr30K			MS-COCO			
		I2T R@1	T2I R@1	Rsum	5K I2T R@1	5K T2I R@1	1K Rsum	5K Rsum
Local	SceneGraphParser	86.2	73.4	547.1	58.6	46.9	533.3	444.1
	FACTUAL-T5	85.9	73.9	547.9	58.1	47.4	533.2	445.3
Global	SceneGraphParser	87.1	73.8	548.1	61.0	46.5	533.8	447.8
	FACTUAL-T5	87.4	74.3	548.7	61.3	46.9	533.6	448.9
Ensemble	SceneGraphParser	89.4	77.5	557.2	62.9	50.2	540.6	460.1
	FACTUAL-T5	**90.3**	**78.1**	**558.9**	**63.9**	**50.7**	**541.0**	**462.0**

A.2 Evaluation Protocol and Limitations

The CG-VSG supervision and LVLM adaptation use the Flickr30K training set. The resulting refiner is frozen and used to construct offline RSG inputs for both benchmarks. Consequently, the MS-COCO retrieval experiments include external Flickr30K supervision at the graph-refinement stage, even though the local and global matchers are trained only with MS-COCO retrieval pairs. The cross-dataset experiment in Appendix A.7 transfers MS-COCO-trained matchers to Flickr30K while retaining the Flickr30K-adapted refiner. End-to-end zero-shot transfer with source-only LVLM adaptation is not evaluated.

Table 5 compares visual backbones within full PSGR, while Table 3 compares graph refinement and sampling variants using DINOv2. Interactions between backbone choice and graph-construction variants remain untested.

A.3 Feature Extractors

We examine the sensitivity of PSGR to its input feature extractors. For the textual side, Table 6 replaces FACTUAL-T5 (Li et al., 2023c) with SceneGraphParser (Wu et al., 2019) while keeping the visual features and matching framework unchanged. The comparable MS-COCO results indicate limited sensitivity to the caption parser. For the image side, Table 5 replaces DINOv2 ViT-g/14 (Oquab et al., 2023) with EfficientNet-B5 (Tan & Le, 2019). DINOv2 improves ensemble Rsum from 544.5 to 558.9 on Flickr30K and from 442.0 to 462.0 on MS-COCO 5K. The corresponding gains are 14.4 and 20.0 Rsum. Table 3 separately evaluates graph purification using the same DINOv2 features.

A.4 RPCS Parameter Analysis

We examine RPCS sensitivity to the (a) fallback retention ratio ρ_f, (b) top-k selection threshold K_0, (c) ranking temperature τ, (d) maximum relations N_E, (e) penalty weight, and (f) reward weight.

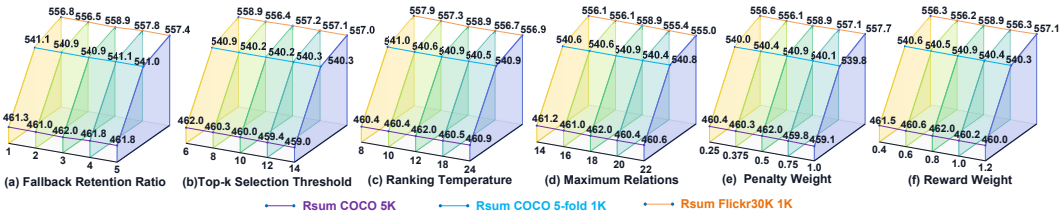

Figure 4: Parameter sensitivity analysis of RPCS on Flickr30K and MS-COCO. We report the Rsum of Flickr30K, MS-COCO 5-fold 1K, and MS-COCO 5K under different settings of the sampling-related hyperparameters.

Table 7: Evaluation for visual input with extra triplets on Flickr30K. Best results are bold.

Setting	Image→Text			Text→Image			Rsum
	R@1	R@5	R@10	R@1	R@5	R@10	
PSGR w/ extra triplets	84.0	97.6	99.0	73.7	93.6	96.6	544.5
PSGR	**90.3**	**99.0**	**99.5**	**78.1**	**94.7**	**97.3**	**558.9**

Fig. 4 shows limited variation in retrieval performance over the tested settings, supporting robustness within the evaluated parameter ranges.

The fallback retention ratio ρ_f controls the number of directly preserved generic-VSG anchors, $K_f = \min\{K_0, n_f, \max(1, \lceil n_f/\rho_f \rceil)\}$, when generic candidates outnumber the missing slots. It does not restrict the candidate pool: remaining slots are sampled from all unselected valid generic triplets.

The top-k selection threshold K_0 determines how many high-ranked RSG triplets are directly preserved as reliable anchors before the weighted sampling step. Moderate values of ρ_f and K_0 provide a balance between retaining confident relations and avoiding excessive low-quality triplets.

The ranking temperature τ appears in the rank prior $\exp(-i/\tau)$ of the sampling weight w_i. A small τ concentrates sampling on high-ranked triplets, whereas a large τ increases the probability of selecting lower-ranked candidates. Retrieval performance varies little over the tested temperature range.

Similarly, the maximum number of relations corresponds to the PSG relation budget N_E. Varying N_E tests the effect of graph size, with RPCS favoring diverse, less repetitive candidates.

For the grouped reward and penalty settings in Fig. 4, penalty weight and reward weight refer to the coefficients in the RPCS. The penalty-weight setting jointly changes the subject-relation redundancy coefficient λ_{sp} and the relation-label redundancy coefficient $\lambda_{p'}$ in the denominator, with the default pair set to $(\lambda_{sp}, \lambda_{p'}) = (0.5, 0.2)$. The reward-weight setting jointly changes the relation-label-diversity coefficient λ_p and the relation-type-diversity coefficient λ_t in the numerator, with the default pair set to $(\lambda_p, \lambda_t) = (0.4, 0.3)$. Thus, these two axes evaluate the strength of diversity rewards and redundancy penalties as grouped controls rather than isolated single-coefficient changes.

These results characterize sensitivity over the tested parameter ranges. The module ablations appear in Table 3; variability across random seeds is not measured.

A.5 PSG WITH EXTRA TRIPLETS

We evaluate whether LVLM-generated extra triplets improve retrieval when included in visual global pooling. Extra triplets undergo the same preprocessing and embedding procedure as textual triplets. Their embeddings are concatenated with the image-side feature sequence before being jointly aggregated by the visual GPO; no separate extra-triplet GPO or post-pooling MLP fusion is used. Since extra triplets lack bounding-box grounding, they are not added to the local branch. Results use the same min-max ensemble strategy.

Table 8: Ablation study on textual encoder on Flickr30K. Within-branch best results are bold.

Branch	Textual Feature	Image→Text			Text→Image			Rsum
		R@1	R@5	R@10	R@1	R@5	R@10	
Local	w/ BERT	60.8	92.6	97.3	65.6	88.9	93.6	498.8
	w/ BiGRU	**85.9**	**98.6**	**99.4**	**73.9**	**93.3**	**96.8**	**547.9**
Global	w/ BERT	**88.7**	98.0	99.2	72.8	92.3	95.6	546.6
	w/ BiGRU	87.4	**98.2**	**99.5**	**74.3**	**93.0**	**96.3**	**548.7**
Ensemble	w/ BERT	87.9	98.7	99.4	74.8	93.5	96.4	550.7
	w/ BiGRU	**90.3**	**99.0**	**99.5**	**78.1**	**94.7**	**97.3**	**558.9**

Table 9: Cross-dataset retrieval results from MS-COCO to Flickr30K. Darker cyan indicates better performance within each metric column.

Method	Image→Text			Text→Image			Rsum
	R@1	R@5	R@10	R@1	R@5	R@10	
SCAN (Lee et al., 2018)	49.8	77.8	86.0	38.4	65.0	74.4	391.4
CVSE (Wang et al., 2020a)	56.4	83.0	89.0	39.9	68.6	77.2	414.1
SGRAF (Diao et al., 2021)	65.7	87.2	93.4	48.1	73.9	81.9	450.2
HREM (Fu et al., 2023)	70.3	90.7	95.2	52.9	78.7	86.2	474.0
CHAN (Pan et al., 2023)	68.7	91.5	95.3	55.6	81.2	87.5	479.8
PSGR (ours)	76.3	94.4	96.9	64.2	87.2	92.7	511.7

Table 7 shows that this concatenation reduces ensemble Rsum by 14.4 points (558.9 to 544.5), with Image→Text and Text→Image R@1 decreasing by 6.3 and 4.4 points. A possible explanation is that ungrounded semantic embeddings alter the feature-value rankings used by visual GPO, changing the pooled representation of grounded image evidence. The results favor pooling grounded visual features alone in the tested configuration. The role of extra triplets in LVLM training and their use with alternative fusion mechanisms remain open questions.

A.6 Replacing BiGRU with BERT

Table 8 compares BiGRU (Cho et al., 2014) and BERT (Devlin et al., 2019) under the evaluated configurations. The local branch drops by 49.1 points (547.9 to 498.8), with Image→Text and Text→Image R@1 decreasing by 25.1 and 8.3 points. The global branch changes less: BERT raises Image→Text R@1 by 1.3 points but lowers Rsum from 548.7 to 546.6. The ensemble remains 8.2 points below BiGRU (550.7 vs. 558.9). Contextualized token states may alter the entity-level and relation-level correspondences needed by local matching, offering one possible explanation for this contrast. BiGRU performs better under these training settings.

A.7 Cross-Dataset Evaluation

We evaluate cross-dataset retrieval by training the local and global matchers with MS-COCO image-text pairs and directly testing them on Flickr30K without updating the matcher parameters. At test time, the Flickr30K RSG inputs have already been generated by the frozen LVLM refiner used throughout the paper, which was adapted with 22,000 SFT and 7,000 GRPO samples from the Flickr30K training set. As shown in Table 9, PSGR achieves 511.7 Rsum, exceeding the strongest baseline CHAN by 31.9 points. The improvement is particularly clear for Text→Image retrieval, where PSGR improves R@1, R@5, and R@10 over CHAN by 8.6, 6.0, and 5.2 points, respectively. For Image→Text retrieval, PSGR exceeds the strongest baseline result in each recall metric by 6.0, 2.9, and 1.6 points. The larger Text→Image gains suggest that the transferred matchers benefit

most when caption-aligned graph evidence must distinguish the target from many candidate images. These gains describe matcher transfer with a Flickr30K-adapted refiner, under the protocol specified above.

A.8 Comparison with Large-Scale Pretrained Models

Table 10 compares PSGR with large-scale pretrained models on Flickr30K 1K. CLIP learns dual-encoder representations through image-text contrastive pretraining; BLIP combines understanding and generation with caption bootstrapping; X-VLM learns multi-grained visual-textual alignments; and BLIP-2 uses a Q-Former to extract language-relevant visual features. In this comparison, CLIP is evaluated without downstream fine-tuning, BLIP and BLIP-2 transfer from COCO fine-tuning, and X-VLM uses Flickr30K fine-tuning with 4M pretraining images. The latter three employ cross-modal reranking. PSGR is trained on Flickr30K and performs retrieval through purified scene graph matching.

Table 10: Supplementary retrieval results on Flickr30K 1K under the training and inference settings described in the text. Darker cyan indicates better performance within each metric column.

Method	Image→Text			Text→Image			Rsum
	R@1	R@5	R@10	R@1	R@5	R@10	
CLIP (Radford et al., 2021)	88.0	98.7	99.4	68.7	90.6	95.2	540.6
BLIP (Li et al., 2022a)	96.7	100.0	100.0	86.7	97.3	98.7	579.4
X-VLM (Zeng et al., 2022)	96.8	99.8	100.0	86.1	97.4	98.7	578.8
BLIP-2 ViT-g (Li et al., 2023a)	97.6	100.0	100.0	89.7	98.1	98.9	584.3
PSGR (ours)	90.3	99.0	99.5	78.1	94.7	97.3	558.9

PSGR exceeds the reported zero-shot CLIP result by 18.3 Rsum, including gains of 2.3 and 9.4 points in Image→Text and Text→Image R@1, respectively. However, BLIP, X-VLM, and BLIP-2 exceed PSGR by 20.5, 19.9, and 25.4 Rsum. Their stronger results demonstrate the effectiveness of large-scale vision-language pretraining combined with cross-modal reranking. In particular, BLIP-2 uses its first-stage image encoder and Q-Former without an LLM for retrieval, followed by top-128 ITM reranking. The models differ in pretraining data, downstream supervision, and scoring procedures. Appendix C.4 discusses these approaches in relation to fine-grained retrieval.

A.9 Inference Trade-off

Fig. 5 compares retrieval accuracy with matching-stage speed relative to SGM, using precomputed graph inputs and features. The horizontal axis reports speedup over SGM, and the vertical axis reports the Rsum gain. Full PSGR improves Rsum from 478.6 to 558.9 (+80.3), but its $0.7\times$ speedup corresponds to approximately $1.43\times$ the matching latency of SGM. The standalone local and global branches reach 547.9 and 548.7 Rsum, with gains of 69.3 and 70.1 at $1.1\times$ and $1.9\times$ speed, respectively. Thus, ensembling improves accuracy at additional matching cost, whereas each standalone branch is faster than SGM under the reported timing setup. Timing covers the matching stage and excludes generic scene graph generation, LVLM refinement, and visual feature extraction. For a prebuilt image gallery, image-side preprocessing can be cached; newly added gallery images and novel image queries still require preprocessing before matching. End-to-end latency, preprocessing throughput, and cache storage are not measured.

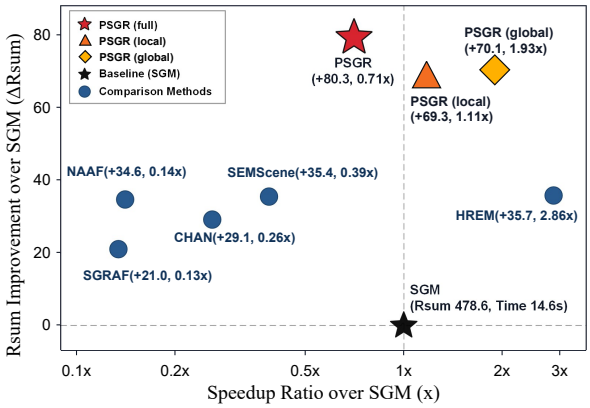

Figure 5: Matching-stage accuracy–speed trade-off; graph generation and feature extraction are excluded.

A.10 Fine-tuning Performance Across Steps

Fig. 6 illustrates the training dynamics of the LVLM during the two-stage adaptation process. In the SFT stage, the model learns to follow the required response format and produce grounded scene graph triplets from caption-guided supervision, providing a stable initialization for subsequent reinforcement learning. In the GRPO stage, the smoothed training curves show the overall optimization trend after applying the content-order, cross-modal alignment, and format rewards. The upward trend indicates that GRPO further refines the model beyond supervised imitation, encouraging outputs that are not only parseable but also more consistent with caption-level semantic relations.

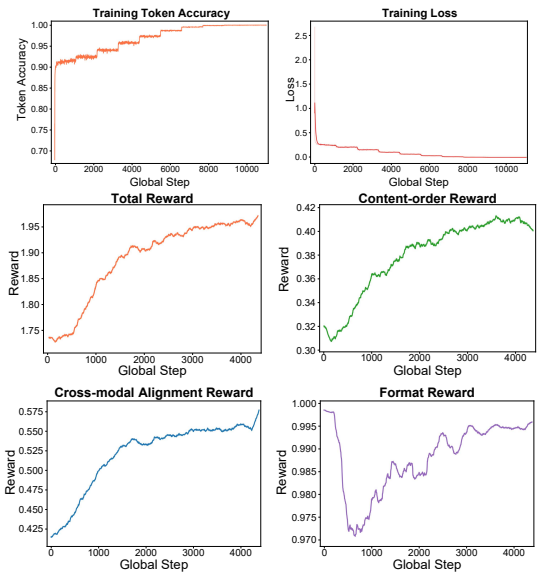

Figure 6: SFT and GRPO training curves.

A.11 Details of the CG-VSG Dataset

We further analyze the constructed CG-VSG dataset to clarify what supervision is provided to the LVLM. Each sample contains two aligned graph components. The CSG describes caption-level entities and relations extracted from the five image captions, while the RSG provides the corresponding caption-guided visual graph with grounded object labels, bounding boxes, and visual triplets.

Fig. 7(a) visualizes the triplet count distribution of the serialized RSG responses. The grounded main triplets are concentrated around a relatively stable count, while each sample contains far fewer extra triplets. Among the 29,000 RSG samples, 28,886 contain at least one extra triplet, accounting for 99.6% of the dataset. This pattern is consistent with the intended supervision design: the main part provides dense grounded visual evidence for graph construction, whereas the extra part records a small number of high-level semantic relations that are informative but not reliably localizable.

Fig. 7(b) shows that the RSG reduces relation-label and subject–relation repetition compared with the generic VSG. Fig. 7(c) further compares the semantic composition of relations, where RSGs contain a more balanced mixture of spatial, action, attribute, and semantic relations. This shift explains why the RSG is more suitable for image-text retrieval.

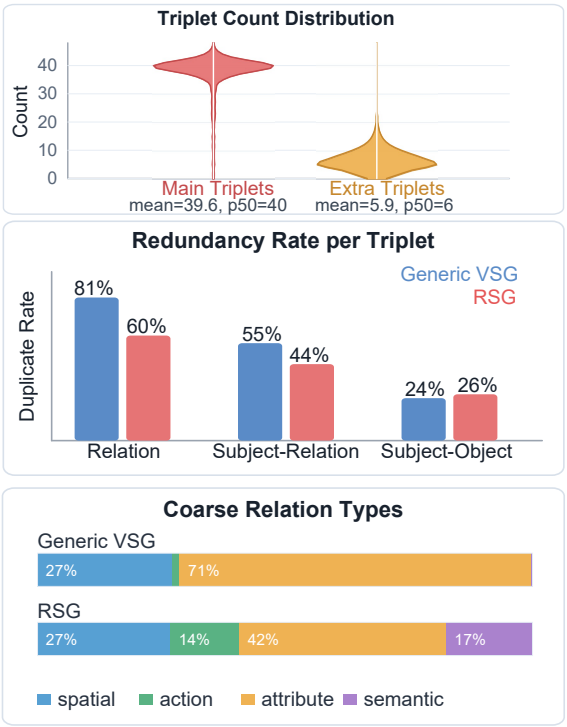

Figure 7: CG-VSG dataset statistics.

B Qualitative Analysis

B.1 Text→Image Retrieval Cases from VSG to PSG

Fig. 8 further shows how PSGR improves Text→Image retrieval through the transition from VSG to RSG and then to PSG. The generic VSG provides many generic visual entities and part-level relations, but its graph structure is weakly aligned with the query because important semantic cues are mixed with noisy or ambiguous relations. After LVLM-based refinement, the RSG introduces more query-related visual evidence by refining entity granularity, separating the key visual instances, and supplementing candidate relations that better reflect the text semantics. However, the RSG can still contain redundant body parts, weak auxiliary links, and background context. RPCS converts the RSG into a PSG by sampling the triplets that are most useful for retrieval and discarding less discriminative graph components. As a result, the final PSG keeps the entities and relations that directly support the query-image correspondence, making the Text→Image similarity less affected by noisy VSG structures and improving the ranking of the correct image.

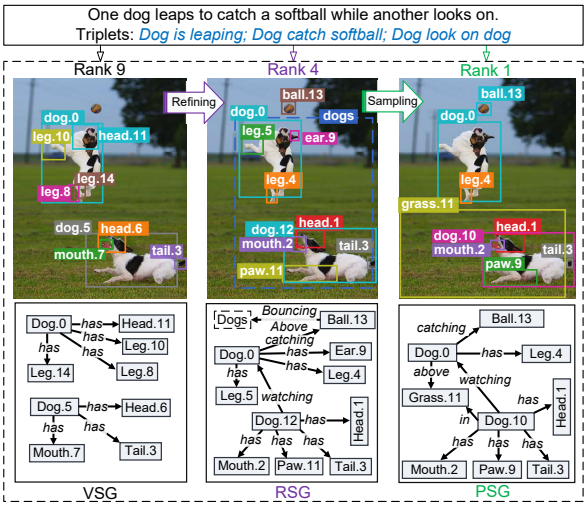

Figure 8: Text→Image case from VSG to PSG.

B.2 PSG Text→Image Retrieval Cases from SFT to Full GRPO

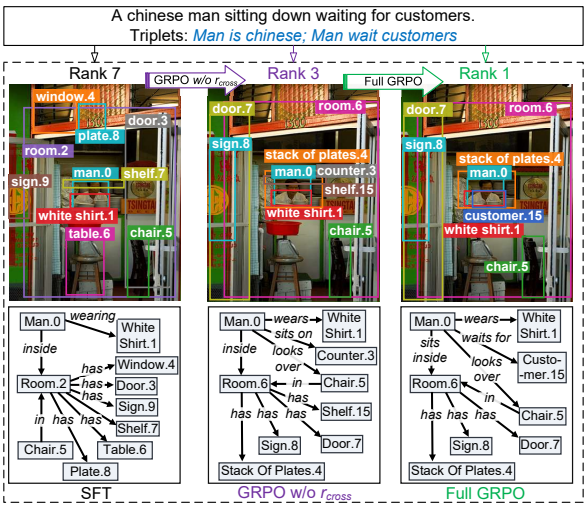

Figure 9: Retrieval case after tuning.

Fig. 9 analyzes the effect of LVLM tuning on PSG quality for Text→Image retrieval. With SFT alone, the LVLM can follow the desired output format and generate grounded scene graph triplets, but the resulting PSG still tends to describe broad scene context and common entities. Adding GRPO without the cross-modal alignment reward improves the graph by refining entity granularity and recovering more informative relations, which raises the correct image in the retrieval ranking. Full GRPO further introduces cross-modal semantic feedback, encouraging the PSG to retain relations that are directly supported by the query TSG while suppressing visually present but weakly relevant details. This progression shows that the final PSG is not merely a formatted graph produced by supervised imitation. It is shaped by reward-guided refinement to better support Text→Image matching.

C Detailed Related Work

C.1 Scene-graph-based matching methods for ITR

Unlike conventional methods that mainly align isolated visual regions with textual words, scene-graph-based methods formulate ITR as a structured cross-modal matching problem between VSGs and TSGs. SGM (Wang et al., 2020b) is the first method to represent images and captions as scene graphs and performs local-level matching for entities and relations with graph encoders. LGSGM (Nguyen et al., 2021) further integrates local-level matching with global-level alignment, employing GCNs (Kipf & Welling, 2016) to aggregate neighborhood information and enrich graph representations. CSCC (Zeng et al., 2021) explores conceptual-level and syntactical-level cross-modal alignments with a cross-level consistency constraint, encouraging fine-grained semantic correspondence and high-level structural consistency simultaneously. Recent studies further improve scene-graph-based retrieval from semantic inference and multi-level matching perspectives. SGSIN (Pei et al., 2023) constructs scene graphs within different matching modules and aggregates neighborhood information to refine entity representations, thereby improving fine-level visual-textual alignment. SEMScene (Liu et al., 2024) exploits scene graph semantics from low-level entity matching, mid-level relation matching, and high-level holistic graph matching.

Although these methods based on generic SGG demonstrate the effectiveness of scene graphs for ITR, most of them focus on designing stronger graph matching networks after the scene graphs have already been constructed. Their performance is therefore largely limited by the quality of the input VSGs. As mentioned in Section 1, SGG methods tend to produce VSGs that describe images with generic entity and relation categories, while TSGs represent caption-discriminative entities and relations. In contrast, PSGR focuses on purifying the generic VSG before retrieval, proposing RPCS to extract the purified visual graphs that are more consistent with captions.

C.2 ITR Methods Based on Other Visual Representations

C.2.1 Region-based Matching Methods

Region-based matching methods for ITR typically employ Bottom-Up and Top-Down Attention (BUTD) (Anderson et al., 2018) to extract visual regions. Building upon these region-level features, recent works learn Visual Semantic Embeddings (VSEs) to bridge the cross-modal gap. VSE++ (Faghri et al., 2018) introduces hard negative mining into the margin ranking loss and has become a widely adopted baseline. VSE∞ (Chen et al., 2021) proposes Generalized Pooling Operators

(GPO) to aggregate visual and textual features more effectively. MV-VSE (Li et al., 2022b) and MV-VSE++ (Li et al., 2025b) model multiple visual semantic views for images to capture intra-class variations. Coder (Wang et al., 2022) learns instance-level and concept-level representations with diversity-sensitive contrastive learning. CVSE (Wang et al., 2020a) incorporates consensus knowledge by constructing concept-level correlations shared across modalities, while RVSE++ (Li et al., 2025a) introduces selective hard negative mining to alleviate gradient vanishing during triplet-loss optimization.

Beyond learning VSEs, many methods explicitly reason about fine-grained region-word correspondences. SCAN (Lee et al., 2018) introduces stacked cross attention to discover latent alignments between image regions and caption words. VSRN (Li et al., 2019) performs visual semantic reasoning over region features to generate global scene-aware representations. GCN+DIST (Li et al., 2020) explores high-order attention over objects, relations, and attributes, and further reranks retrieval candidates with distraction information. GSMN (Liu et al., 2020) constructs graph-structured region representations to model sparse and dense dependencies, while CAAN (Zhang et al., 2020) introduces context-aware attention for adaptive cross-modal retrieval. GraDual (Long et al., 2022) integrates visual and textual graph information into dual-modal representations. SGRAF (Diao et al., 2021) jointly learns local-level and global-level similarities, while NAAF (Zhang et al., 2022) further considers negative effects caused by mismatched region-word pairs. CHAN (Pan et al., 2023) interprets fine-grained alignment from an information coding perspective and focuses on the most relevant region-word pairs to reduce redundant cross-attention. SDE (Kim et al., 2023) designs a similarity function for image-text embedding sets to better measure cross-modal correspondence. HREM (Fu et al., 2023) models hierarchical relations at both fragment and instance levels, and CORA (Pham et al., 2024) represents captions with scene graphs and encodes entity-attribute and entity-entity relations through graph attention.

CMSF (Zong et al., 2026) uses Faster R-CNN region features and textual embeddings as inputs to a multimodal spiking neural network. Bidirectional hard alignment and spike-fusion soft supervision improve region-word correspondence, while the fusion module is omitted during inference. Table 1 includes its BiGRU variant.

UIID (Zhang et al., 2026) unifies intra-modal relationship and inter-modal similarity distillation for efficient retrieval. We group its VSRN-teacher variant with region-based methods according to the teacher's visual representation. The deployed student instead uses ResNeXt and ALBERT to produce global image and text embeddings, without requiring the teacher at inference. Unlike PSGR, UIID transfers representation knowledge rather than explicitly purifying scene graphs.

Region-based methods improve local correspondence, but their region-level representations lack explicit relations. They do not necessarily construct explicit visual subject–predicate–object triplets for direct matching with textual scene graphs.

C.2.2 Patch-based Matching Methods

With the development of Vision Transformers (Dosovitskiy et al., 2021), patch tokens have become an alternative fine-grained visual format for ITR. LAPS (Fu et al., 2024) performs linguistically aware patch slimming to remove redundant patches and calibrate ambiguous patch-word alignment. AVSE (Liu et al., 2025b) dynamically selects multi-view image patches according to different textual inputs and computes visual-semantic similarity through meta-semantic embeddings. DSMD (Liang et al., 2025) introduces a dynamic balance strategy for multi-objective distillation, while D2S-VSE (Liu et al., 2025a) improves sparse matching by distilling dense knowledge into sparse texts.

These patch-based methods enable dense visual-language alignment, but they mainly operate on unstructured local tokens. Although this paradigm expands the alignment space for retrieval, these unstructured patches lack a one-on-one correspondence with the semantic triplets in captions. Therefore, they also fail to support the explicit modeling of entities and their relations.

In contrast to regions that lack relational context and unstructured patches with semantic misalignment, the VSG serves as a balanced and suitable representation for ITR. It reformulates ITR as a cross-modal matching process between entities and their relations. To this end, PSGR aims to purify the generic VSG, thereby narrowing the semantic gap between the VSG and the TSG to achieve more precise alignment.

C.3 Generic Scene Graph Generation Methods

Scene Graph Generation (SGG) aims to parse an image into a structured representation composed of entities and their pairwise relations. Early methods exploit visual context and statistical regularities to infer relation triplets, such as Neural Motifs (Zellers et al., 2018) and VCTree (Tang et al., 2019). Later works reformulate SGG as direct or end-to-end triplet prediction, using sparse queries, transformer architectures, or set prediction objectives (Teng & Wang, 2022; Li et al., 2023b; Cong et al., 2023). Recent advances extend SGG toward open-vocabulary and LVLM-based generation (Chen et al., 2024; Li et al., 2024), while reinforcement-learning-enhanced LVLMs improve structured relation generation and reasoning (Chen et al., 2025; Li et al., 2026; Feng et al., 2026). PSGR builds on this line of LVLM adaptation to refine scene graphs for retrieval.

The distinction of PSGR lies in the task formulation and downstream pipeline. Generic SGG and recent LVLM-based methods optimize graph generation quality or structured reasoning, whereas PSGR uses caption-guided rewards to produce retrieval-oriented candidates and applies RPCS before cross-modal matching. Captions emphasize discriminative entities and relations, while generic VSGs often retain coarse categories and generic interactions. PSGR uses caption-guided refinement and RPCS to reduce this semantic gap before graph matching.

C.4 Large-Scale Vision-Language Pretraining

CLIP (Radford et al., 2021) learns transferable image and text representations through contrastive pretraining on large-scale image-text pairs. BLIP (Li et al., 2022a) supports both vision-language understanding and generation, using caption generation and filtering to improve noisy web supervision. X-VLM (Zeng et al., 2022) learns multi-grained alignments by locating visual concepts associated with text and aligning them across granularities. These methods establish important pretraining approaches for retrieval and broader vision-language tasks; in particular, large-scale pretraining can itself support fine-grained alignment.

Our focus is fine-grained image-text retrieval through explicit scene graph purification and matching. Accordingly, region-based, scene-graph-matching, and patch-based retrieval methods form the primary comparison groups, while CLIP, BLIP, and X-VLM provide supplementary context. PSGR also uses pretrained components, while focusing on explicit graph purification. The supplementary comparison includes the training and inference settings of each model. Appendix A.8 provides the supplementary comparison table; the matched-backbone ablations remain the more direct evidence for PSGR's purification modules.

D Caption-Guided LVLM Refinement Details

D.1 Caption-Guided VSG Fine-Tuning Data Construction

Figure 10 summarizes the complete supervision-construction and LVLM-adaptation pipeline. The left part shows the three inputs: an image, its five corresponding captions, and the generic VSG extracted by Neural Motifs (Zellers et al., 2018). The middle part shows how GPT-5 first converts the captions into a Caption Scene Graph (CSG) and then combines the CSG with the image and generic VSG to construct the caption-guided VSG target. The right part summarizes the subsequent SFT and GRPO stages described in the following subsection. In the reported experiments, all CG-VSG supervision is constructed from the Flickr30K training set. Given an image I, its captions $\mathcal{C}_I = \{c_i\}_{i=1}^{N_C}$ with $N_C = 5$, and the generic visual scene graph $G_I^0 = (V_I^0, E_I^0, B_I^0)$, we leverage GPT-5 (Singh et al., 2025) to generate a caption-aligned scene graph target $\hat{G}_I = (\hat{V}_I, \hat{E}_I, \hat{B}_I)$. Rather than replicating all entities and relations in G_I^0, the objective is to refine the generic VSG into a representation that better matches caption semantics. This curated supervision is used to fine-tune one LVLM refiner, which is subsequently frozen and applied offline to both Flickr30K and MS-COCO images.

The construction prompt is structured as a two-stage task. In the first stage, GPT-5 (Singh et al., 2025) is prompted to construct a Caption Scene Graph (CSG) G_I^C solely from $\mathcal{C}_I$. This stage extracts caption-described entities and entity-relation-entity triplets that are explicitly expressed in natural language, thereby providing a caption-level semantic prior for retrieval. In the second stage, GPT-

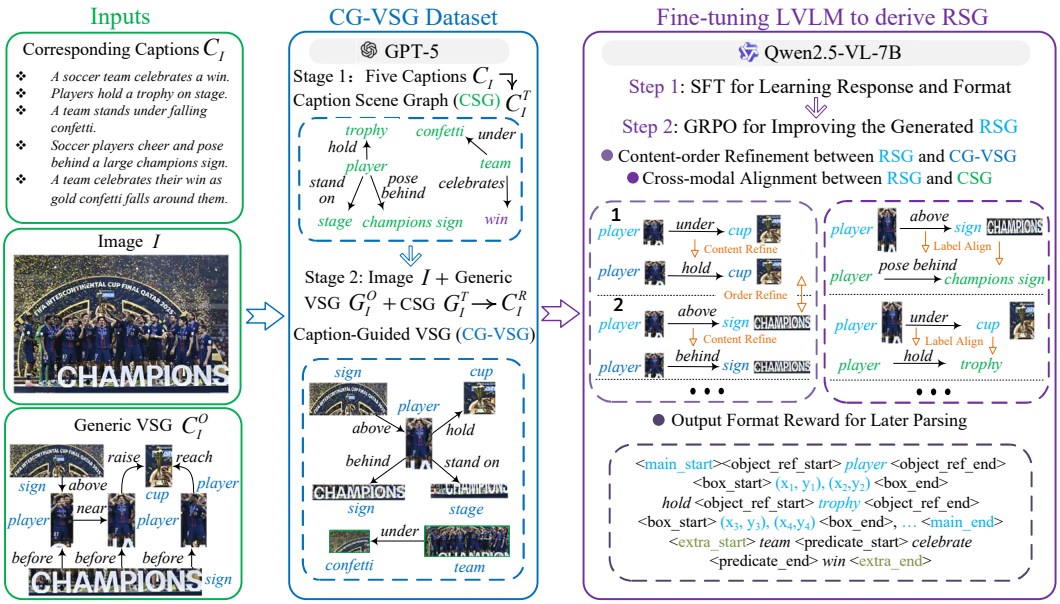

Figure 10: Overview of CG-VSG construction and LVLM adaptation for RSG generation.

5 (Singh et al., 2025) refines the generic VSG G_I^0 by jointly analyzing the image, the generic VSG, and the newly generated CSG:

$$\text{Stage 1: } \mathcal{C}_I \rightarrow G_I^C; \quad \text{Stage 2: } (I, G_I^0, G_I^C) \rightarrow \hat{G}_I. \tag{15}$$

Instead of naively rewriting the initial VSG, this stage uses G_I^C to guide targeted operations: selecting caption-relevant visual entities, removing redundant entities, and refining entities and relations for stronger visual-textual consistency. We also prompt GPT-5 to add missing caption-related entities and relations while strictly constraining such additions to visual evidence within the image. These additions are not utilized as part of the visual inputs for PSGR, but their utility and impact in our global matching branch are evaluated in the experiments.

To ensure the constructed samples are directly compatible with subsequent training pipelines, the prompt imposes strict formatting constraints. All entity labels must be expressed as natural single-word or multi-word phrases. Bounding boxes are required to be associated with the corresponding visual entities. The model is prohibited from introducing fictional elements that are not supported by either the image or the captions. Finally, the response must be a nested dictionary comprising a CSG G_I^C and caption-guided VSG $\hat{G}_I$:

$$\mathcal{P}(I, \mathcal{C}_I, G_I^0) \rightarrow \{G_I^C, \hat{G}_I\}. \tag{16}$$

Following this generation process, the parsed and normalized dictionary is serialized into structured training responses, denoted as $\mathcal{Y}$. This pipeline yields 29,000 samples from the Flickr30K (Young et al., 2014) training set, from which 22,000 samples are selected for SFT and 7,000 samples for GRPO via bin sampling based on output length. No MS-COCO sample is used to adapt the LVLM; MS-COCO enters only when the frozen refiner generates offline RSG inputs and when the retrieval branches are trained.

D.2 SFT AND GRPO FOR LVLM

The right part of Fig. 10 illustrates the two-stage adaptation of Qwen2.5-VL (Bai et al., 2025): SFT first learns the required grounded response structure, and GRPO then improves RSG content, caption alignment, ordering, and output validity through the three reward terms. The constructed fine-tuning samples are converted into the grounding-oriented dialogue format of Qwen2.5-VL. We denote one training sample as $(x, \mathcal{Y})$, where x contains the image, the instruction, and the generic VSG triplets, and $\mathcal{Y}$ is the target response for RSG generation. The response is serialized into two regions:

$$\begin{aligned} \mathcal{Y} = &\texttt{<main_start>}\, \mathcal{T}^m\, \texttt{<main_end>} \\ &\texttt{<extra_start>}\, \mathcal{T}^e\, \texttt{<extra_end>}. \end{aligned} \tag{17}$$

Here $\mathcal{T}^m$ is the set of main triplets grounded by entity references and bounding boxes, and $\mathcal{T}^e$ is the set of extra triplets that are informative but not reliably localized in the image. Each grounded triplet $\tau^m \in \mathcal{T}^m$ is serialized as

$$\begin{aligned} \tau^m &= \mathrm{Ref}(s, b^s)\ p\ \mathrm{Ref}(o, b^o), \\ \mathrm{Ref}(v, b) &= \langle \mathrm{obj} \rangle v \langle / \mathrm{obj} \rangle \langle \mathrm{box} \rangle b \langle / \mathrm{box} \rangle. \end{aligned} \tag{18}$$

Here s, p, and o denote the subject entity, relation, and object entity, respectively, while b^s and b^o represent their corresponding bounding boxes. In this way, SFT autoregressively trains the model to generate correct triplets in a parseable format and to ground entities to their respective image regions. Given the target token sequence $\mathcal{Y} = (y_1, \ldots, y_L)$ and input x, the SFT objective for optimizing the LVLM policy π_θ is:

$$\mathcal{L}_{\mathrm{SFT}} = -\sum_{t=1}^{L} \log \pi_\theta(y_t | x, y_{<t}), \tag{19}$$

where $y_{<t}$ denotes the prefix tokens.

Starting from the SFT checkpoint, GRPO (Shao et al., 2024) optimizes the LVLM by sampling a group of K candidate responses $\{\hat{\mathcal{Y}}_k\}_{k=1}^K$ for each prompt x. Rather than training a critic network, GRPO guides the policy update using a group-normalized advantage:

$$\begin{aligned} A_k &= \frac{R_k - \mu_R}{\sigma_R + \epsilon_{\mathrm{num}}}, \qquad u_{k,t} = \frac{\pi_\theta(y_{k,t}|h_{k,t})}{\pi_{\mathrm{old}}(y_{k,t}|h_{k,t})}, \\ \mathcal{J}_{\mathrm{GRPO}}(\theta) &= \mathbb{E}_{x\sim\mathcal{D},\{\hat{\mathcal{Y}}_k\}\sim\pi_{\mathrm{old}}(\cdot|x)} \left[\frac{1}{K}\sum_{k=1}^{K} \frac{1}{L_k} \sum_{t=1}^{L_k} \left(\ell_{k,t} - \beta d_{k,t} \right) \right], \\ \ell_{k,t} &= \min\left\{ u_{k,t} A_k,\ \mathrm{clip}(u_{k,t}, 1-\epsilon_{\mathrm{clip}}, 1+\epsilon_{\mathrm{clip}}) A_k \right\}. \end{aligned} \tag{20}$$

This is the standard clipped GRPO objective (Shao et al., 2024), maximized over θ. Here $h_{k,t} = (x, y_{k,<t})$, L_k is the response length, and μ_R and σ_R are the group reward mean and standard deviation. The same outcome advantage A_k is assigned to all response tokens. The numerical stabilizer ϵ_{num} is distinct from the clipping threshold ϵ_{clip}. With $v_{k,t} = \pi_{\mathrm{ref}}(y_{k,t}|h_{k,t})/\pi_\theta(y_{k,t}|h_{k,t})$, the KL penalty is $d_{k,t} = v_{k,t} - \log v_{k,t} - 1$; π_{ref} is the reference policy and β controls its weight.

In our implementation, the reward R is defined as:

$$R = \lambda_{\mathrm{co}} r_{\mathrm{co}} + \lambda_{\mathrm{cross}} r_{\mathrm{cross}} + \lambda_{\mathrm{format}} r_{\mathrm{format}}, \tag{21}$$

where r_{co}, r_{cross}, and r_{format} denote the content-order, cross-modal alignment, and format rewards. Their coefficients are all set to 1. Content-order matching uses parsed grounded main triplets; cross-modal matching uses both parsed main and extra triplets. Format checks separately score boundary markers and attempted triplet fragments.

D.2.1 Content-Order Reward

The content-order reward r_{co} aligns the triplets in the generated RSG with those of the caption-guided VSG $\hat{G}_I$ by maximizing the joint similarity of semantic labels and spatial boxes. Let $\hat{\mathcal{T}}^m$ and $\mathcal{T}^m$ denote the generated and target triplet sets, respectively. A generated triplet is represented as $\hat{t} = (\hat{s}, \hat{p}, \hat{o}, \hat{b}^s, \hat{b}^o)$, where $\hat{s}$ and $\hat{o}$ are the subject-entity and object-entity labels, $\hat{p}$ is the relation label, and $\hat{b}^s$ and $\hat{b}^o$ are their respective bounding boxes. A target triplet t is defined analogously. The pairwise matching score between a generated and a target triplet is formulated as:

$$m(\hat{t}, t) = \begin{cases} \sigma(\hat{s}, s)\sigma(\hat{p}, p)\sigma(\hat{o}, o), & \text{if } \begin{matrix} \mathrm{IoU}(\hat{b}^s, b^s) \geq \gamma_{\mathrm{iou}} \\ \mathrm{IoU}(\hat{b}^o, b^o) \geq \gamma_{\mathrm{iou}} \end{matrix}, \\ 0, & \text{otherwise}, \end{cases} \tag{22}$$

where the function $\sigma(\cdot, \cdot)$ measures the cosine similarity between two entity or relation labels, and γ_{iou} is the IoU threshold for validating head-entity and tail-entity grounding. We set $\gamma_{\mathrm{iou}} = 0.5$ in our implementation. Algorithm 1 constructs the score matrix; Hungarian matching then maximizes

Algorithm 1 Score matrix construction for Hungarian triplet matching

Require: Generated response y, target response y^*, IoU threshold γ_{iou}
Ensure: Score matrix $\mathbf{M}$ for Hungarian matching

```
Parse y and y* into generated T̂^m and target triplets T^m
Build entity box dictionaries B̂ and B* from the parsed grounded entities
Initialize M ∈ R^{|T̂^m| × |T^m|} with zeros
for each generated triplet t̂_i ∈ T̂^m and target triplet t_j ∈ T^m do
    Compute the subject-entity IoU and object-entity IoU from B̂ and B*
    if either IoU is smaller than γ_iou then
        M_ij ← 0
    else
        M_ij ← m(t̂_i, t_j)
    end if
end for
return M
```

the total score to obtain a one-to-one assignment $\mathcal{M}^*$ between generated and target triplets. The content reward is:

$$r_{\text{content}} = \frac{1}{|\mathcal{T}^m|} \sum_{(\hat{t},t)\in\mathcal{M}^*} m(\hat{t}, t). \tag{23}$$

Since both the generated and ground-truth triplets are structured as ordered sequences, we introduce an order reward to encourage the relative order of the matched generated triplets to remain consistent with that of their ground-truth counterparts. After sorting the matched pairs according to the target sequence order, let rank_p^i be the position of the generated triplet in the i-th matched pair. The order reward is:

$$r_{\text{order}} = \frac{1}{|\mathcal{M}^*| - 1} \sum_{i=1}^{|\mathcal{M}^*|-1} \mathbb{I}[\text{rank}_p^i < \text{rank}_p^{i+1}], \tag{24}$$

This expression applies when $|\mathcal{M}^*| \geq 2$. The indicator counts adjacent pairs that preserve generation order after sorting by target order. We set $r_{\text{order}} = 1$ for one matched pair and $r_{\text{order}} = 0$ for no pairs. If either generated or target main triplets are empty, the implementation returns $r_{\text{co}} = 0$ without evaluating the normalized expressions. Matched pairs are Hungarian assignments, including zero-score pairs; they are not filtered by a positive-score threshold. The final reward is:

$$r_{\text{co}} = \alpha_{\text{content}} r_{\text{content}} + \alpha_{\text{order}} r_{\text{order}}. \tag{25}$$

Here, α_{content} and α_{order} control their contributions. We set $\alpha_{\text{content}} = 0.9$ and $\alpha_{\text{order}} = 0.1$.

D.2.2 Cross-Modal Alignment Reward

The cross-modal reward r_{cross} further aligns the triplets in the generated RSG with those of the CSG G_I^C. Let $\mathcal{T}^q$ denote the CSG triplets parsed from the five training captions, and let $\hat{\mathcal{T}}^v = \hat{\mathcal{T}}_{\text{cls}}^m \cup \hat{\mathcal{T}}^e$ denote the generated RSG triplets, where grounded entity identifiers in the main triplets are converted to category names and $\hat{\mathcal{T}}^e$ denotes the extra triplets. For a CSG triplet $q = (s_q, p_q, o_q)$ and a generated triplet $\hat{v} = (\hat{s}, \hat{p}, \hat{o})$, we compute $c(q, \hat{v}) = \sigma(s_q, \hat{s}) \cdot \sigma(p_q, \hat{p}) \cdot \sigma(o_q, \hat{o})$. The cross-modal reward is:

$$r_{\text{cross}} = \frac{1}{|\mathcal{T}^q|} \sum_{q\in\mathcal{T}^q} \max\left\{0, \max_{\hat{v}\in\hat{\mathcal{T}}^v} c(q, \hat{v})\right\}. \tag{26}$$

The reward function returns zero if either triplet collection is empty. Semantic scores use spaCy `en_core_web_sm` document similarity on lowercased labels with colon-delimited instance suffixes removed, rather than retrieval-branch GloVe embeddings. Multi-word labels are passed as complete strings to spaCy, without custom phrase pooling. Both semantic terms multiply raw similarities without per-label nonnegative clipping. The zero floor above acts only after multiplication, so two negative label similarities can yield a positive product.

D.2.3 FORMAT REWARD

The format reward r_{format} ensures that the generated RSGs comply with the structural requirements of downstream ITR graph builders. It consists of a structural term and a triplet-validity term:

$$r_{\text{format}} = \alpha_{\text{structure}} r_{\text{structure}} + \alpha_{\text{triplet}} r_{\text{triplet}}. \tag{27}$$

Here, $\alpha_{\text{structure}}$ and α_{triplet} control the contributions of response structure and triplet validity, and we set both to 0.5. The structural term $r_{\text{structure}}$ assigns equal credit to the required boundary markers that define the main and extra parts as formulated in Eq. 17. The triplet-validity term r_{triplet} measures the ratio of correctly formatted triplets to the total:

$$r_{\text{triplet}} = \begin{cases} N_{\text{valid}}/N_{\text{attempt}}, & N_{\text{attempt}} > 0, \\ 0, & N_{\text{attempt}} = 0. \end{cases} \tag{28}$$

The implementation counts nonempty comma-space-separated fragments within the main and extra sections as attempted triplets and checks their respective format patterns. Main-triplet validity requires two entity references, two boxes, and an intermediate relation; extra-triplet validity checks the entity–predicate–entity pattern. Each of the four boundary markers earns one quarter of $r_{\text{structure}}$, independently of triplet validity. Thus, empty sections with all four markers receive $r_{\text{format}} = 0.5$, whereas an empty response receives zero. Each reward handler returns zero on an exception.

E IMPLEMENTATION DETAILS AND LVLM OUTPUT FORMAT

E.1 IMPLEMENTATION DETAILS

E.1.1 LVLM FINE-TUNING COST

On eight 80GB NVIDIA A100 GPUs, we fine-tune Qwen2.5-VL-7B (Bai et al., 2025) with SWIFT (Zhao et al., 2025) (version 3.8.0) in two consecutive stages. In the SFT stage, the LVLM is trained on 22,000 CG-VSG instruction samples for 10 epochs with full-parameter updates, a learning rate of 5×10^{-5}, and a batch size of 20. This stage runs for 11,000 optimization steps, taking 58,839.96 seconds (16.3 hours) and 2.45×10^{19} FLOPs. The GRPO stage uses 7,000 reward-learning samples. It is trained for 5 epochs with a learning rate of 1×10^{-6}, 8 completions per sample, and a batch size of 8 per GPU. This stage runs for 4,375 steps and takes 296,357.38 seconds (82.3 hours).

E.1.2 RETRIEVAL-BRANCH OPTIMIZATION SETTINGS

For both Flickr30K and MS-COCO, the local and global branches are trained independently with the same graph input configuration specified in Appendix E.2. The RSG inputs are precomputed by the frozen LVLM adapted on Flickr30K; only the retrieval branches are optimized on each benchmark. In the cross-dataset experiment, both branches are trained on MS-COCO and evaluated on Flickr30K without branch-level fine-tuning, while the Flickr30K RSG inputs still come from the Flickr30K-adapted refiner.

Both branches are trained for 30 epochs with a batch size of 128 and a gradient-clipping threshold of 2. The local branch is optimized with Adam, learning rate 4×10^{-4}, weight decay 0, margin 0.35, and learning-rate decay factor 0.2. The global branch is optimized with AdamW, learning rate 2×10^{-4}, weight decay 10^{-3}, margin 0.2, and learning-rate decay factor 0.1. The learning rate is decayed at epochs 15 and 25. The two branches are trained on a single NVIDIA RTX 4090 GPU.

E.2 METHOD PARAMETER SETTINGS

This subsection collects the concrete parameter values referenced by the Method section and clarifies which component each group belongs to.

PSG construction and RPCS sampling. For the PSG sampling module in Section 3.2, each PSG uses $N_E = 18$ triplet slots and $N_O = 36$ entity slots. The anchor cap is $K_0 = 6$ for both fallback and over-complete sampling, with fallback retention ratio $\rho_f = 3$ and rank temperature $\tau = 12$. The supplied graph-builder defaults include node-coverage weight $\lambda_n = 0.8$, relation-label and

coarse-type diversity weights $\lambda_p = 0.4$ and $\lambda_t = 0.3$, and redundancy penalties $\lambda_{sp} = 0.5$ and $\lambda_{p'} = 0.2$. Sampling uses an image-specific deterministic seed derived from the global seed and image identifier; fallback sampling uses a separate identifier suffix.

Empty and insufficient graph inputs. Relations with malformed endpoints, empty predicates, or unavailable endpoint boxes are removed before sampling. An existing RSG with zero valid relations uses the fallback sampler, whereas a missing RSG record directly takes the first N_E valid generic relations. If the combined selection remains short, selected relations are repeated cyclically to fill N_E slots. If none exist, all relation slots use an `unknown` self-relation. Entity slots prioritize selected relation endpoints, then unused RSG and generic entities, and cycle available entities when necessary. With no usable objects, the builder returns an all-unknown graph with boxes $[0, 0, 1, 1]$. These are structural placeholders, not inferred semantic evidence.

Coarse relation types. The RPCS implementation lowercases and strips each predicate, then applies ordered rules. Exact membership in {has, have, of, wear, wearing, holding, with, contains, covered with} gives `attribute`. Otherwise, exact membership in {in, on, at, near, behind, under, above, below, beside, next to, in front of, around, inside, outside, between, along} gives `spatial`. Otherwise, a whitespace-delimited token matching any of {ride, riding, run, running, walk, walking, sit, sitting, stand, standing, play, playing, look, looking, jump, jumping, eat, eating, drink, drinking, hold, holding, pose, posing, catch, catching, row, rowing, ski, skiing} gives `action`. All remaining predicates are `semantic`. There is no lemmatization or learned classifier; rule priority makes `holding` an attribute.

Feature extraction and graph encoding. The feature extractor uses `dinov2_vitg14` in evaluation and inference mode. Entity regions use PSG boxes; each relation region is the smallest axis-aligned box enclosing its subject and object boxes. Coordinates are clipped to the image boundary with a minimum crop width and height of one pixel. RGB crops are resized directly to 224×224 by bicubic interpolation and normalized with ImageNet mean $(0.485, 0.456, 0.406)$ and standard deviation $(0.229, 0.224, 0.225)$. Each crop is independently passed through `backbone(inputs)`; the extractor neither applies RoIAlign to a full-image patch map nor averages patches selected by a box. It stores the backbone's returned vector without additional L2 normalization. The Flickr30K and MS-COCO configurations use the same preprocessing, bfloat16 autocasting, and conversion of outputs to float32 before storage. For the feature-extraction module in Section 3.3.1, DINOv2 ViT-g/14 is used as the default visual feature extractor. It produces entity-region and relation-region features with dimension $d_v = 1536$. Entity and relation labels are initialized with 300-dimensional GloVe embeddings. The fused visual-semantic features are mapped to a 2048-dimensional GCN input space and encoded by a one-layer GCN with batch normalization, Swish activation, and a dropout rate of 0.4. The final entity and relation embeddings have dimension $d = 1024$. Caption words and TSG triplets are also initialized with 300-dimensional GloVe embeddings and encoded by a two-layer bidirectional encoder.

Global graph sequence. For the generalized global matching module in Section 3.3.3, the default PSG setting gives $N_I = N_O + N_E = 54$ image-side graph elements, with $N_O = 36$ entities and $N_E = 18$ relations. The text-side sequence length is $N_T = L + M$, where L is the caption length and M is the number of parsed TSG triplets. All image-side and text-side graph elements are represented in the shared dimension $d = 1024$ before GPO aggregation. Image batches concatenate the entity and relation blocks; the fixed entity budget ensures valid entries remain contiguous. Caption batches place relation states immediately after each caption's valid word states, followed by padding. GPO uses valid lengths to mask padding. A one-layer bidirectional GRU encodes 32-dimensional sinusoidal rank embeddings with a hidden size of 32 per direction; the two directions are averaged, linearly projected to scalar scores, and normalized across valid ranks with temperature $T_{\mathrm{GPO}} = 0.1$. Feature values are sorted in descending order independently per dimension and weighted by these rank weights. The implementation uses -10000 for masked sorting entries and zeros the padded tail before pooling. Image and text GPOs have separate learned parameters.

Reward configuration. GRPO uses all three rewards, a group size of 8, a clipping threshold of 0.2, and a KL coefficient of 0.001. Semantic reward scores use spaCy `en_core_web_sm` document similarity, as defined in Appendix D.

E.3 Retrieval Training Losses

This subsection details the training losses used by the local and global retrieval branches in Section 3.3. Given a similarity matrix S, S_{ii} denotes a matched image-caption pair, while S_{ij} and S_{ji} denote mismatched pairs in the two directions. The warm-up objective uses all in-batch negatives:

$$\begin{aligned}\ell_{\text{WARM}}(S) = &\sum_i \sum_{j\neq i} [\gamma - S_{ii} + S_{ij}]^+ \\ &+ \sum_i \sum_{j\neq i} [\gamma - S_{ii} + S_{ji}]^+ .\end{aligned} \tag{29}$$

After warm-up, hard negative mining (Faghri et al., 2018) keeps the most violating negatives:

$$\begin{aligned}\ell_{\text{MATCH}}(S) = &\sum_i \left[\gamma - S_{ii} + \max_{j\neq i} S_{ij}\right]^+ \\ &+ \sum_i \left[\gamma - S_{ii} + \max_{j\neq i} S_{ji}\right]^+ .\end{aligned} \tag{30}$$

Here, γ is the margin and $[x]^+ = \max(x, 0)$. The local branch is optimized with $\ell_{\text{MATCH}}(S_{\text{local}})$. For the global branch, we use the first epoch as a warm-up stage to stabilize the GPO-based graph-level representation, and then switch to hard negative mining:

$$\ell_{\text{global}} = \begin{cases} \ell_{\text{WARM}}(S_{\text{global}}), & \text{if 1st epoch,} \\ \ell_{\text{MATCH}}(S_{\text{global}}), & \text{otherwise.} \end{cases} \tag{31}$$

E.4 Prompt and Output of LVLM

Fig. 11 presents the template prompt and response format used for Refined Scene Graph (RSG) generation. The prompt takes an input image together with 25 triplets detected by Neural Motifs (Zellers et al., 2018), referred to as the generic VSG. Rather than directly preserving all detector outputs, the LVLM is instructed to remove uninformative or redundant triplets, merge semantically overlapping relations, and infer higher-level entities or relations only when they are supported by the visual evidence.

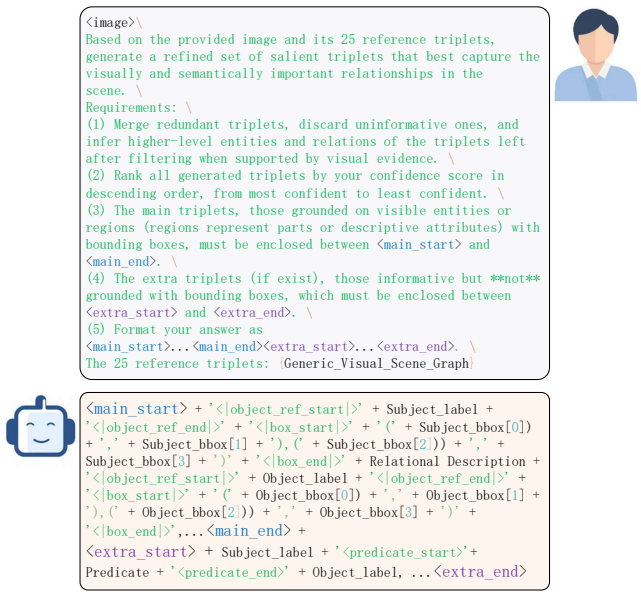

Figure 11: Template prompt and structured response format for RSG generation.

The response is divided into two complementary parts. The `main triplets`, enclosed by `<main_start>` and `<main_end>`, describe salient visual relationships whose subject and object entities, or entity-related regions, can be grounded by bounding boxes. These triplets retain explicit object references and spatial coordinates, enabling subsequent region-level graph construction. The `extra triplets`, enclosed by `<extra_start>` and `<extra_end>`, contain additional informative semantic relations that cannot be reliably grounded to specific bounding boxes.

The prompt additionally requires the generated triplets to be ranked by confidence. This order carries the LVLM's semantic preference into RPCS, which preserves high-confidence anchors while balancing relation diversity and redundancy. The explicit boundary markers allow the parser to separate grounded and ungrounded content before constructing the candidate pools.

This template converts the LVLM response into a structured and directly parseable representation. In particular, the grounded main triplets provide reliable visual anchors for constructing the PSG, while the extra triplets supplement the graph with high-level semantic information that may not correspond to a single visible region. The resulting refined scene graph therefore preserves salient visual evidence while reducing noise from the original detector-produced VSG.

F AI DISCLOSURE

In this work, we used GPT-5 (Singh et al., 2025) to construct the Caption-Guided VSG (CG-VSG) supervision dataset. Given an image, its five captions, and a detector-produced VSG, GPT-5 generated a Caption Scene Graph and a caption-guided visual scene graph target. The resulting 29,000 samples were constructed from the Flickr30K training set, with 22,000 used for SFT and 7,000 for GRPO. The generated responses were parsed, normalized, and serialized for LVLM adaptation, as detailed in Appendix D. These targets are AI-generated supervision and may contain semantic or grounding errors; they should not be treated as manually verified annotations. We take responsibility for the use of these generated artifacts and for the final content and claims of this work.