



VKG-IR: Reliability-Aware Visual Graph Embedding for Fine-Grained Image Retrieval

Nam Nguyen^{1,2} , Long Nguyen³ , and Thanh Le¹

¹ University of Sciences, Hue University, 77 Nguyen Hue St., Hue, Vietnam
namnp@hcmue.edu.vn, lmthanh@hueuni.edu.vn

² Ho Chi Minh City University of Education, 280 An Duong Vuong St., Ho Chi Minh, Vietnam

³ University of Science, VNU-HCM, 227 Nguyen Van Cu St., Ho Chi Minh, Vietnam
nhblong@fit.hcmus.edu.vn

Abstract. Global vector-based methods for image retrieval often fail to capture the fine-grained interactions between objects in complex scenes. We present VKG-IR, a reliability-aware visual graph retrieval framework that represents each image as a set of visual triplets extracted by a pretrained scene-graph generator. VKG-IR includes a reliability module that estimates a reliability score for each triplet and uses these scores to reweight relations during structural aggregation, reducing the impact of inaccurate or redundant triplets on the image representation. The model is trained with a joint objective that combines an image-level ranking loss and a relation-level reliability supervision term. At inference, a query image is encoded once and ranked against indexed embeddings using cosine similarity. Experiments on the VG-COCO dataset under an IDF-weighted relevance labeling protocol show that VKG-IR improves Recall@K and mAP over global and graph-based baselines. Ablation studies further examine the contribution of the reliability mechanism and the linear fusion of graph and global representations.

Keywords: Semantic Image Retrieval · Scene Graph · Visual Triplets · Reliability Estimation · Fine-grained Relational Reasoning

1 Introduction

Semantic image retrieval aims to match images based on high-level conceptual similarity [1], in contrast to traditional content-based image retrieval systems that rely primarily on low-level visual primitives such as color, shape, and texture [2, 3]. The advent of deep learning and large-scale vision-language pre-training has enabled the extraction of robust representations, which has led to consistent improvements in ranking performance on standard benchmarks [4].

Despite these advancements, existing architectures face challenges in compositional generalization. Prior studies [5, 6] indicate that performance often degrades when queries involve rare or out-of-distribution combinations of objects, attributes, and relationships. For instance, distinguishing a “horse handler” from a “horse rider” requires modelling role-specific relations between entities, rather than merely recognizing the set of co-occurring objects. Figure 1 illustrates this challenge: two images share the same entities

(person, horse) but exhibit distinct predicates (leading versus riding). In Fig. 1a, the fine-grained predicate “leading” does not appear among the top-10 predictions, as the generic predicate “near” is ranked above it. In contrast, in Fig. 1b, the predicate “riding” is correctly prioritized.

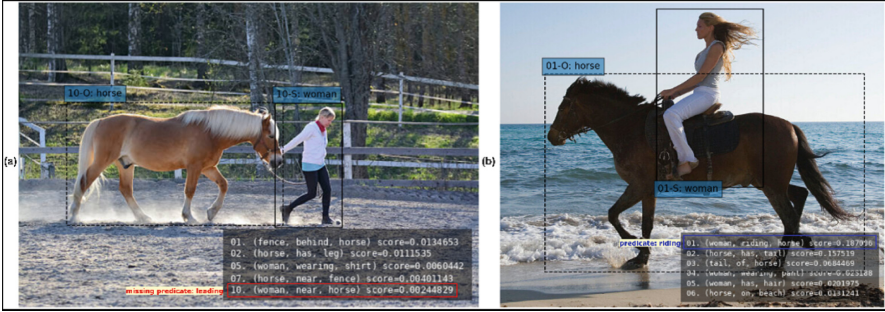


Fig. 1. Contrast in predicted predicate labels for two images that contain the same entity pair (woman, horse). In (a), the fine-grained predicate “leading” does not appear among EGTR’s top-10 predictions and is ranked below the generic predicate “near”. In (b), the predicate “riding” is ranked first.

A fundamental bottleneck stems from the reliance on global image representations in dual-encoder architectures. In conventional vision-language models, an entire image is compressed into a single embedding vector—often obtained via global pooling or a dedicated [CLS] token—to support direct matching in a shared embedding space [7, 8]. Although computationally efficient, this design creates an information bottleneck, and fine-grained interaction details can be abstracted away. Consequently, models may exhibit a bag-of-objects limitation and fail to resolve compositional ambiguities, especially when scenes share the same object set but differ in their relational structure.

To mitigate these structural limitations, recent approaches leverage scene graphs (SGs) to explicitly encode subject–predicate–object triplets for retrieval [9, 10]. A common limitation in existing methods is treating generated SGs as highly reliable and aggregating all triplets with equal weight. This assumption may not hold in practice, as automatically extracted relationships are often inaccurate or redundant. This motivates adaptive assessment of relationship reliability when synthesizing the final image-level representation.

In practice, the fidelity of input graphs varies substantially. Empirical studies [11–13] indicate that Scene Graph Generation (SGG) models are prone to producing spurious or trivial predicates that are misaligned with the visual context. In downstream applications, Maheshwari et al. [14] observed that retrieval performance depends strongly on the structural integrity of the graph. If a model integrates unreliable graphs without filtering, relationship-level errors can propagate to the global embedding, degrading ranking quality.

To address these challenges, we propose VKG-IR, a framework that integrates triplet-level reliability estimation into the retrieval pipeline through dual-level optimization. At the local level, a dedicated reliability module assigns confidence scores to extracted

triplets, attenuating the influence of spurious or redundant relationships. At the global level, these scores serve as adaptive attention weights during feature aggregation, shaping the final image representation. The entire pipeline is trained end-to-end under a dual-level objective, aligning relationship reliability refinement with ranking performance. Accordingly, VKG-IR targets role- and relation-sensitive queries, while offline graph extraction enables efficient online retrieval.

In summary, our main contributions are as follows:

- **VKG-IR framework.** We propose VKG-IR, a framework that represents images using explicit visual triplets to learn a structure-aware embedding, integrating it with global features via a simple linear combination.
- **Triplet reliability for structure synthesis.** We introduce a triplet-level reliability module that assigns reliability scores to extracted relationships. The module is trained with ground truth-aligned partial label supervision and uses these scores to modulate triplet contributions during feature synthesis.
- **Evaluation scope.** We evaluate VKG-IR on VG-COCO within the VG150 label space under the triplet-overlap protocol. Under identical settings, VKG-IR is compared against comparable baselines and improves ranking performance. Ablation studies further quantify the contribution of the reliability module.

The remainder of this paper is organized as follows. Section 2 reviews related work. Section 3 details the proposed VKG-IR framework. Section 4 presents the experimental analysis, and Sect. 5 concludes the study.

2 Related Work

2.1 Semantic Image Retrieval

Semantic image retrieval aims to bridge the “semantic gap” between low-level visual features (e.g., color, texture) and high-level concepts. Early approaches largely framed the task as keyword matching through automatic tag assignment or tag completion. However, these methods rely heavily on manual annotations and often represent semantics as discrete, isolated labels, thereby failing to capture the relational structure among objects [3].

Recent work has shifted toward deep representation learning on large-scale datasets. Convolutional neural network (CNN)-based architectures, optimized with ranking objectives (e.g., triplet loss), have improved retrieval performance [3, 15]. To capture fine-grained details, methods such as DELF [16] use attention mechanisms to focus on salient local regions. Nevertheless, these representations are not explicitly structured to model subject–predicate–object interactions. This becomes a limitation when scenes share the same set of entities but differ in spatial or semantic configuration.

Beyond standard CNNs, large-scale vision-language models (VLMs) such as CLIP [7] and BLIP [17] have shown strong zero-shot retrieval capabilities. However, both traditional CNNs and contemporary VLMs primarily rely on holistic alignment in a shared embedding space. As a result, structural information regarding entity interactions is often attenuated. Recent studies on compositional generalization [5, 6] indicate that performance often degrades in out-of-distribution scenarios. Without explicit structural priors,

these models may conflate subject and object roles or fail to correctly bind attributes to entities, resulting in a bag-of-objects bias in which semantically distinct scenes become effectively indistinguishable.

However, prevailing methods often treat SGs as static structures and regard all triplets as fully reliable labels. This simplifying assumption can be limiting in practice because generated graphs frequently contain erroneous relationships or redundant links [13]. To address this, we introduce a triplet-level reliability estimation mechanism. Specifically, we assign a reliability score to each triplet, which serves as a weighting factor during feature aggregation. This strategy avoids treating all relationships uniformly and attenuates the impact of spurious or contextually misaligned connections on the final image representation.

2.2 Visual Knowledge Graphs and Relational Reasoning

Scene graph generation models transform unstructured image data into graph representations that explicitly capture entities and their relationships. Many SGG methods leverage global context and graph-based modeling to improve predicate prediction, particularly under long-tailed distributions and unreliable or redundant predicate labels [18–20]. In this work, we treat a scene graph as a visual knowledge graph, providing a basis for leveraging relational information when constructing image-level representations. This enables semantic image retrieval to incorporate structured knowledge, moving beyond sole reliance on holistic alignment of global features.

More recently, attention has shifted toward modeling the reliability of information in knowledge graphs to refine embeddings and inference processes. Egger et al. [21] proposed a reliability measure tailored to the knowledge graph embedding space, while a recent survey [22] synthesizes approaches to knowledge reliability from a representation learning perspective. These findings suggest that triplet-level reliability modeling can provide a basis for weighting relations when synthesizing graph representations.

Although graph neural network models are effective for reasoning over well-formed structured graphs, their direct application to the visual domain remains contingent upon the quality of the input graph. Automated scene graph extraction often produces erroneous or redundant relation labels, and the resulting distribution is typically imbalanced, rendering the uniform processing of all triplets inappropriate [11–13]. Without a triplet-level reliability assessment in message passing to weight individual relations, contextually misaligned triplets can propagate errors into the image-level representation and degrade ranking performance in semantic image retrieval. This motivates incorporating triplet-level reliability scores into the weighted aggregation mechanism for image representation learning, explicitly aligned with the ranking objective.

3 Method

3.1 Problem Definition and Visual Knowledge Graph Representation

Formally, we define the semantic image retrieval over a dataset $\mathcal{I} = \{I_i\}_{i=1}^N$. Each image I_i is associated with visual knowledge graph $G_i = (V_i, E_i)$, where V_i denotes the set of detected entities and $E_i \subseteq V_i \times \mathcal{R} \times V_i$ is the set of edges capturing their relationships,

with relation labels drawn from a predicate set $\mathcal{R}$. For computation, G_i is decomposed into a set of visual triplets:

$$T_i = \{\tau_{i,m}\}_{m=1}^{M_i}, \quad \tau_{i,m} = (s_{i,m}, r_{i,m}, o_{i,m})$$

where $s_{i,m}, o_{i,m} \in V_i$ are the subject and object entities, and $r_{i,m} \in R$ is the relation (predicate) label. M_i denotes the number of triplets for image I_i . Let $T_G(I_i)$ denote the ground-truth triplet set associated with I_i .

Our objective is to learn an embedding function $F : (I_i, G_i) \mapsto g_i \in \mathbb{R}^d$ that projects the multimodal input into a d -dimensional embedding space. Retrieval ranks a candidate image I_j against a query I_q using cosine similarity $\text{sim}(g_q, g_j)$. The mapping F is optimized to satisfy the ranking constraint, where $\mathcal{P}(I_q)$ and $\mathcal{N}(I_q)$ denote the positive and negative image sets for query I_q , respectively:

$$\forall j \in \mathcal{P}(I_q), \quad \forall k \in \mathcal{N}(I_q) : \quad \text{sim}(g_q, g_j) > \text{sim}(g_q, g_k)$$

This constraint enforces higher similarity for images that share semantic structure with the query than for those with divergent structure.

3.2 VKG-IR

Figure 2 provides an overview of VKG-IR, which adopts a dual-branch architecture comprising a graph-reasoning branch and a global visual branch. In the graph branch, an input image I_i is processed to construct a visual knowledge graph $G_i = (V_i, E_i)$, from which a set of visual triplets T_i is derived. Each triplet $\tau_{i,m} \in T_i$ is processed by the Multimodal Triplet Fusion block to produce a fused embedding $z_{i,m}$. The Reliability Module R maps $z_{i,m}$ to a triplet-level reliability score $\lambda_{i,m} = R(z_{i,m})$, which is normalized across triplets via a softmax to obtain aggregation weights $\alpha_{i,m}$. These weights govern the weighted aggregation to form the structural embedding g_i^{graph} .

In parallel, the global branch encodes I_i into a holistic representation g_i^{global} , which is fused with the structural embedding g_i^{graph} to form the final representation g_i . During training, the model optimizes a joint objective: L_{ret} enforces image-level ranking constraints, while L_{rel} supervises the reliability module to learn triplet reliability scores $\lambda_{i,m}$, which induce aggregation weights $\alpha_{i,m}$ that regulate each triplet's contribution to g_i^{graph} . Structurally, the graph branch comprises three functional units: multimodal triplet encoding, reliability assessment, and weighted aggregation.

Visual Knowledge Graph and Triplet Encoding. Rather than processing the entire graph $G_i = (V_i, E_i)$ monolithically, VKG-IR operates on a decomposed set of triplets $T_i = \{\tau_{i,m}\}_{m=1}^{M_i}$. This design decomposes the graph into local relational units of the form (s, r, o) , where each triplet $\tau_{i,m}$ is matched against ground-truth triplet set $T_G(I_i)$ to derive positive/negative labels. The resulting L_{rel} signal backpropagates to the Reliability Module R , enabling it to learn triplet-level reliability scores $\lambda_{i,m}$. These scores are normalized to obtain weights $\alpha_{i,m}$, which regulate each triplet's contribution during feature aggregation and yield the graph-level representation g_i^{graph} .

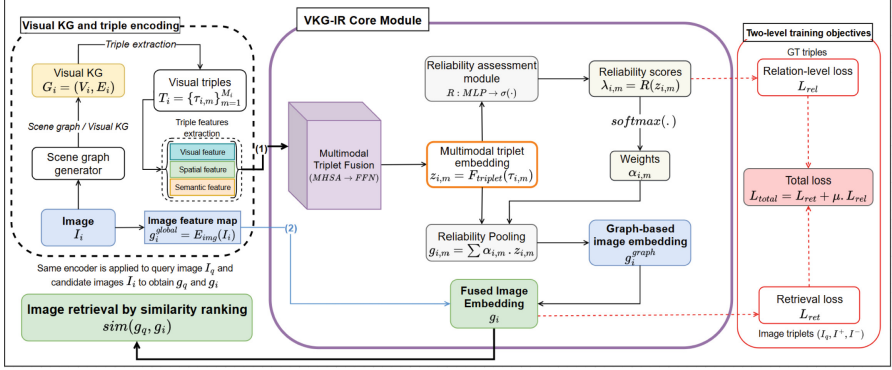


Fig. 2. Overall architecture of VKG-IR

Specifically, each triplet $\tau_{i,m}$ is characterized by three feature modalities: (i) visual features extracted from the subject and object regions; (ii) spatial features encoding geometric attributes and relative positioning; and (iii) semantic features derived from word embeddings of entity and predicate labels. The Multimodal Triplet Fusion block integrates these sources to produce a unified triplet embedding $z_{i,m}$.

Multimodal Triplet Fusion Module. The Multimodal Triplet Fusion module integrates visual, spatial, and semantic modalities using self-attention over a token sequence constructed from triplet features. For a given $\tau_{i,m}$, the three feature components are projected into input tokens. A multi-head self-attention (MHSA) layer models cross-modal interactions by attending over these tokens, and the resulting representation is passed through a feed-forward network to produce the triplet embedding:

$$z_{i,m} = F_{\text{triplet}}(\tau_{i,m}), \quad z_{i,m} \in \mathbb{R}^d$$

The vector $z_{i,m}$ serves as a unified multimodal triplet embedding of $\tau_{i,m}$, consolidating cues from the visual, spatial, and semantic modalities in a shared embedding space.

In VKG-IR, the self-attention in the triplet fusion block is designed to model intra-triplet interactions. Specifically, MHSA is applied only to the token sequence constructed from a single triplet $\tau_{i,m}$, preventing token interactions across different triplets T_i . This design yields a triplet-specific embedding $z_{i,m}$ before reliability estimation. The Reliability Module R then maps $z_{i,m}$ to a triplet-level reliability score $\lambda_{i,m}$. These scores are normalized across triplets to obtain aggregation weights $\alpha_{i,m}$, which control each $z_{i,m}$'s contribution during weighted aggregation when constructing g_i^{graph} .

Reliability Assessment Module and Weighted Synthesis Mechanism. To modulate the contribution of individual triplets when constructing the graph representation, VKG-IR incorporates a reliability assessment module R . Given a triplet embedding $z_{i,m}$, R is implemented as an Multi-Layer Perceptron (MLP) followed by a sigmoid activation and outputs a triplet-level reliability score $\lambda_{i,m}$:

$$\lambda_{i,m} = R(z_{i,m}), \quad \lambda_{i,m} \in [0, 1]$$

During training, R is optimized using the reliability loss L_{rel} with ground-truth-matched (GT) triplets as positive supervision. Because triplet annotations are incomplete, non-matching triplets are treated as unlabeled rather than hard negatives to reduce the risk of penalizing valid but unlabeled triplets. For each image I_i , the raw scores $\{\lambda_{i,m}\}_{m=1}^{M_i}$ are normalized across triplets via softmax to obtain aggregation weights $\alpha_{i,m}$, where $\alpha_{i,m} \geq 0$ and $\sum_{m=1}^{M_i} \alpha_{i,m} = 1$:

$$\alpha_{i,m} = \frac{\exp(\lambda_{i,m})}{\sum_{k=1}^{M_i} \exp(\lambda_{i,k})}.$$

Subsequently, the structural representation is obtained by weighted aggregation:

$$g_i^{\text{graph}} = \sum_{m=1}^{M_i} \alpha_{i,m} z_{i,m}.$$

Our architecture decouples indexing-time preprocessing from the retrieval phase: SGG inference and graph encoding are performed offline during indexing, so online retrieval reduces to computing a query embedding followed by nearest-neighbor search. Within the graph encoder, reliability scoring, softmax normalization, and weighted aggregation scale linearly with the number of triplets, i.e., $O(M_i)$. Triplet self-attention is also linear in M_i because it operates on a fixed-length token sequence within each triplet. We cap the number of triplets per image to a fixed budget $M_{\text{max}} = 25$ to balance structural coverage and computational cost.

After softmax normalization, g_i^{graph} becomes a convex combination of triplet embeddings $z_{i,m}$, where $\alpha_{i,m}$ controls each triplet's contribution to the structural representation. The resulting graph-based embedding g_i^{graph} is then linearly fused with the global representation g_i^{global} to form the final image embedding g_i . Note that this subsection describes the mechanism of soft-weighted pooling; the contribution of the reliability component is examined through ablation analysis in the experimental section.

Linear Fusion of Image Embeddings. The graph representation g_i^{graph} is derived from triplet embeddings and is designed to emphasize fine-grained relational information. To retain global visual context, VKG-IR includes a parallel global branch that encodes the image using an image encoder E_{img} producing: $g_i^{\text{global}} = E_{\text{img}}(I_i)$. The final embedding g_i for image I_i is obtained by linearly combining the graph-based and global components:

$$g_i = \beta g_i^{\text{graph}} + (1 - \beta) g_i^{\text{global}}, \quad \beta \in [0, 1].$$

Here, β controls the trade-off between structural information from the graph branch and global visual context from the image encoder. The resulting embedding g_i is used to compute cosine similarities for retrieval and to rank database images for a given query.

3.3 Two-Stage Training and Inference for Image Retrieval

Image-Level Ranking Loss. VKG-IR learns an embedding space in which relevant images are more similar to a query than non-relevant ones. For each query image I_q , we

form a positive set $\mathcal{P}(I_q)$ and a negative set $\mathcal{N}(I_q)$, and sample training triplets (I_q, I^+, I^-) with $I^+ \in \mathcal{P}(I_q)$ and $I^- \in \mathcal{N}(I_q)$. We optimize a triplet margin loss:

$$L_{\text{ret}} = \sum_{(q, +, -)} \max(0, \gamma_{\text{ret}} - \text{sim}(g_q, g^+) + \text{sim}(g_q, g^-)),$$

where $\gamma_{\text{ret}} > 0$ is the margin and $\text{sim}(\cdot, \cdot)$ denotes cosine similarity. This objective encourages:

$$\text{sim}(g_q, g_j) \geq \text{sim}(g_q, g_k) + \gamma_{\text{ret}}, \quad \forall j \in \mathcal{P}(I_q), \quad \forall k \in \mathcal{N}(I_q)$$

which aligns with the ranking objective in image retrieval.

Relational-Level Reliability Loss. Beyond image-level ranking, VKG-IR optimizes the reliability module R to assign higher reliability scores to triplets that match the ground truth within each image. For each triplet embedding $z_{i,m}$, R outputs a reliability score: $\lambda_{i,m} = R(z_{i,m})$, where $\lambda_{i,m} \in [0, 1]$. Based on the alignment with the ground-truth triplet set $T_G(I_i)$, we define the index set of matched triplets:

$$\mathcal{P}_i = \{m \mid \tau_{i,m} \in T_G(I_i)\}$$

and treat the remaining predictions as unlabeled candidates:

$$\mathcal{U}_i = \{m \mid \tau_{i,m} \notin T_G(I_i)\}$$

We train R using a margin-based ranking objective that encourages matched triplets to receive higher scores than unlabeled candidates within the same image:

$$L_{\text{rel}} = \sum_i \sum_{(p,u) \in \mathcal{P}_i \times \mathcal{U}_i} \max(0, \gamma_{\text{rel}} - \lambda_{i,p} + \lambda_{i,u}),$$

where $\gamma_{\text{rel}} > 0$ is the margin. This encourages $\lambda_{i,p} \geq \lambda_{i,u} + \gamma_{\text{rel}}$ for matched triplets p relative to unlabeled candidates u . Since triplet annotations are incomplete, non-matching predictions are treated as unlabeled rather than hard negatives, reducing the risk of penalizing valid but unlabeled triplets. The scores $\lambda_{i,m}$ are then normalized into aggregation weights $\alpha_{i,m}$ to synthesize g_i^{graph} , which in turn affects the final image embedding g_i used for retrieval.

Joint Objective and Inference Strategy for Image Retrieval. During training, VKG-IR optimizes an image-level ranking loss and a relational-level reliability loss via a joint objective:

$$L_{\text{total}} = L_{\text{ret}} + \mu L_{\text{rel}}, \quad \mu > 0,$$

where μ balances the ranking objective and the reliability supervision signal. Under this formulation, L_{ret} optimizes image embeddings for retrieval, while L_{rel} provides supervision for reliability estimation and backpropagates through the reliability branch to learn reliable scores $\lambda_{i,m}$.

During inference, given a query image I_q and a database $\mathcal{D} = \{I_j\}_{j=1}^{N_D}$, retrieval proceeds in two stages: indexing and querying. In the indexing stage, database embeddings $\{g_j\}$ are computed offline (including graph extraction and encoding when used) and stored or indexed for efficient search. In the querying stage, the model computes g_q via a forward pass, evaluates cosine similarities $\text{sim}(g_q, g_j)$ for $I_j \in \mathcal{D}$, and ranks images in descending order to return the retrieval results.

4 Experiments

4.1 Dataset and Evaluation Methods

Dataset. We evaluate VKG-IR on the VG-COCO dataset, formed by the set of images shared by Visual Genome [23] and MS-COCO 2014 [24]. The images are taken from Visual Genome, and the corresponding captions are obtained from MS-COCO for the same image IDs. We adopt the VG150 label space (150 object categories and 50 predicates), which is consistent with EGTR outputs and supports controlled structure-aware evaluation under the triplet-overlap protocol. We split the dataset by image ID into non-overlapping training, validation, and test sets with an 80/10/10 ratio.

Scene Graph Preprocessing. The graph branch uses a pre-trained EGTR model to extract visual triplets from each input image. For image I_i , EGTR produces a triplet set $T_i = \{\tau_{i,m}\}_{m=1}^{M_i}$, where $M_i \leq M_{\max}$ and we cap the output at 36 detected objects and $M_{\max} = 25$ relation edges (i.e., triplets) per image. Manually annotated triplets from Visual Genome are used as ground truth $T_G(I_i)$. We align predicted and ground-truth triplets using exact label matching on (s, r, o) to supervise the reliability module; this supervision is GT-aligned and should not be interpreted as exhaustive, since Visual Genome annotations are incomplete. Graph extraction is performed offline for static or periodically updated databases, where SGG inference is executed once (or rerun periodically) and amortized across many queries. In dynamic settings, newly added images undergo SGG inference and embedding, index updates before they can be retrieved.

Ranking Process. Retrieval proceeds in two phases: indexing and querying. In the indexing phase, we pre-compute and store embeddings $\{g_j\}$ in the database $\mathcal{D} = \{I_j\}_{j=1}^{N_D}$. In the querying phase, given a query image I_q , the model computes g_q via a forward pass and ranks database images by cosine similarity $\text{sim}(g_q, g_j)$ in descending order.

Relevance Criteria. We evaluate semantic relevance at the relational level to assess whether retrieval preserves role assignments in subject–relation–object triplets, targeting the bag-of-objects failure mode where images share similar entities but differ in their interactions. Importantly, relevance is computed from ground-truth triplets T_G only for evaluation, whereas the retrieval model relies on EGTR-predicted triplets at inference time.

Given a query image I_a and a candidate image I_b , we define the overlap of their ground-truth triplet sets, exact match on (s, r, o) as:

$$C(a, b) = T_G(I_a) \cap T_G(I_b).$$

Image I_b is deemed relevant if $C(a, b)$ satisfies at least one of the following conditions: (i) it contains at least one high-IDF triplet, or (ii) it contains two or more matching triplets. Triplet rarity is quantified using inverse document frequency (IDF) computed over the training split:

$$w(\tau) = \log\left(\frac{|D_{train}|}{df(\tau) + 1}\right),$$

where $df(\tau)$ is the number of training images containing triplet τ in D_{train} . We treat a triplet as high-IDF if $w(\tau) \geq \text{median}(w)$, where the median is computed over triplet weights in the training set.

This IDF weighting emphasizes informative semantic units when assigning relevance. Scene-graph relations are typically long-tailed, so common triplets can dominate overlaps [25, 26]. Prioritizing high-IDF triplets helps reduce spurious overlaps induced by frequent triplets, leading to more distinctive relevance labels [27].

For each query image I_q , we define the positive (relevant) set and the negative set as:

$$\mathcal{P}(I_q) = \{I_j \in \mathcal{D} \setminus \{I_q\} \mid \exists \tau \in C(q, j) : w(\tau) \geq \text{median}(w) \text{ or } |C(q, j)| \geq 2\}$$

$$\mathcal{N}(I_q) = \mathcal{D} \setminus (\mathcal{P}(I_q) \cup \{I_q\})$$

This protocol measures relational semantic consistency rather than coarse-grained entity overlap; improvements under this criterion should therefore be interpreted as stronger fine-grained structural matching.

Evaluation Metrics. We report Recall @ K (with $K \in \{1, 5, 10\}$) and mean average precision (mAP). Here, Recall @ K is computed in a hit-based form, i.e., a query is counted as correct if at least one relevant image appears within the top- K results. mAP measures overall ranking quality by averaging average precision over all queries, where relevance is defined by the positive set $\mathcal{P}(I_q)$.

4.2 Baseline Methods

To assess the contribution of structural modeling beyond visual representation, we benchmark VKG-IR against two baseline categories under the protocol in Sect. 4.1:

- **Global baselines:** We evaluate Global-CNN (ResNet-101 with average pooling) and CLIP zero-shot. We also report CLIP-FT CLIP-FT, which fine-tunes the encoder using L_{ret} without the graph branch; this provides a reference point for separating structural effects from simple domain adaptation.
- **Graph baselines:** We compare against Graph-Avg, which uniformly averages triplet embeddings, and Graph-GCN, which connects triplets that share entities and applies GCN layers. Graph-GCN is trained with L_{ret} only and does not use reliability supervision.

For methodological consistency, we restrict comparisons to methods that explicitly model structure using scene-graph triplets. We therefore omit token-level region alignment methods (e.g., FILIP), as their objectives are not directly comparable under the VG150 triplet-overlap relevance criterion.

4.3 Implementation Details

All experiments are implemented in PyTorch and run on a single NVIDIA L4 GPU (24 GB). Images are resized to 224×224 . The global branch produces a 256-dimensional l_2 -normalized feature g_i^{global} . Triplets are encoded by an MHSA-based multimodal fusion module to obtain unified triplet embeddings $\{z_{i,m}\}$. The reliability module R takes $z_{i,m}$ as input and predicts triplet-level reliability scores $\lambda_{i,m} \in [0, 1]$, which are normalized into aggregation weights $\alpha_{i,m}$ for constructing g_i^{graph} . We optimize the joint objective $L_{\text{total}} = L_{\text{ret}} + \mu L_{\text{rel}}$ using AdamW with gradient clipping and a cosine learning-rate schedule with warmup. During inference, candidates are ranked by cosine similarity using the fused embedding:

$$g_i = \beta g_i^{\text{graph}} + (1 - \beta) g_i^{\text{global}}$$

with $\beta = 0.7$ by default (see Sec. 4.5).

4.4 Experimental Results

As shown in Table 1, VKG-IR achieves 15.3 Recall@1 and 20.5 mAP, outperforming CLIP-FT (6.5/4.6) and Graph-GCN by 9.5 points at Recall@1 and 17.2 points in mAP, with consistent gains at Recall@5 and Recall@10. In contrast, Graph-Avg performs notably worse, suggesting that uniform aggregation of triplet embeddings is insufficient without reliability-weighted synthesis and fusion with global features.



Fig. 3. Qualitative retrieval results on the VG-COCO dataset. For each query image, we show the top-4 retrieved candidates from a baseline and from VKG-IR.

Figure 3 reports qualitative examples on VG-COCO. For each query, we compare the top-4 retrieved candidates from a baseline and from VKG-IR, highlighting differences in retrieved relations and role assignments

Table 1. Quantitative results on VG-COCO. Results are reported for VKG-IR and the compared global and graph-based baselines

Methods	Recal@K			mAP
	K = 1	K = 5	K = 10	
Global-CNN	5.0	17.0	24.0	3.5
CLIP zero-shot	7.5	23.0	32.0	6.0
CLIP-FT (w/ L_{rel})	6.5	21.8	30.5	4.6
Graph-Avg	7.1	22.7	31.7	10.1
Graph-GCN	5.8	19.6	27.5	3.3
VKG-IR	15.3	40.2	50.1	20.5

4.5 Ablation Studies

Impact of the Reliability Mechanism and L_{rel} . Table 2 evaluates the reliability module by comparing the full model with unweighted and unsupervised variants. Removing L_{rel} reduces mAP from 20.5 to 17.4, and uniform weighting yields the lowest mAP (15.5). This indicates that relational-level supervision improves the quality of aggregation weights $\alpha_{i,m}$ learned from triplets, and that reliability-weighted pooling helps downweight inaccurate or redundant triplets during aggregation.

Table 2. Ablation analysis of the reliability module and L_{rel}

	Recall@K			mAP
	K = 1	K = 5	K = 10	
VKG-IR	15.3	40.2	50.1	20.5
w/o L_{rel} ($\mu = 0$)	13.8	37.5	47.2	17.4
w/o weighting ($\lambda = 1$)	12.8	35.5	45.0	15.5

The Role of the Global Branch and the Coefficient β . We analyze the effect of the fusion coefficient β , which balances the graph representation g_i^{graph} and the global feature g_i^{global} (Table 3). The two extremes, $\beta = 0$ (global-only) and $\beta = 1$ (graph-only), perform worse than intermediate settings. The best result is obtained at $\beta = 0.7$, indicating that structural relations captured by g_i^{graph} and implicit contextual cues in g_i^{global} provide complementary information for ranking.

Table 3. Ablation analysis of the global feature branch and the fusion coefficient β

	β	Recall@K			mAP
		K = 1	K = 5	K = 10	
Only g^{global}	0.0	6.4	20.3	28.1	5.2
Linear fused	0.3	11.5	32.2	41.1	14.5
Linear fused	0.5	13.5	36.2	45.5	17.0
Linear fused (VKG-IR)	0.7	15.3	40.2	50.1	20.5
Only g^{graph}	1.0	11.8	33.5	42.6	14.8

5 Conclusion

In this paper, we introduced VKG-IR, a framework for semantic image retrieval that leverages visual knowledge graphs derived from scene graphs. VKG-IR integrates relation-level structural representations with global visual features via a simple linear fusion, and includes a reliability module to reduce the influence of triplets that are mis-aligned with the visual context. Experiments on the VG-COCO dataset within the VG150 label space, evaluated under the triplet-overlap protocol, show that VKG-IR improves ranking performance over the baselines under identical settings. Our current evaluation is limited to the VG-COCO domain. Extending the pipeline to other benchmarks will require careful ontology alignment and consistent triplet grounding, which we leave for future work. Future directions include assessing cross-dataset transfer and incorporating external knowledge to improve robustness for rare or ambiguous relations.

Disclosure of Interests. The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Thanh, L.T.V., Van, T.T., Le, T.M.: Semantic-based image retrieval using RS-tree and knowledge graph. In: Lecture Notes in Computer Science, vol. 13757, pp. 481–495. LNAI, Springer, Cham (2022). https://doi.org/10.1007/978-3-031-21743-2_38

2. Zhang, C., Liu, J.: Content based deep learning image retrieval: a survey. In: ACM International Conference Proceeding Series, pp. 158–163 (2023). <https://doi.org/10.1145/3638884.3638908>

3. Dubey, S.R.: A decade survey of content based image retrieval using deep learning. IEEE Trans. Circuits Syst. Video Technol. **32**(5), 2687–2704 (2022). <https://doi.org/10.1109/TCSVT.2021.3080920>

4. Cao, D., Zhou, H., Yang, H.: Survey of semantic-based image-to-image retrieval. In: 2024 5th International Conference on Artificial Intelligence and Computer Engineering (ICAICE), pp. 51–54. IEEE (2024). <https://doi.org/10.1109/ICAICE63571.2024.10863997>

5. Abbasi, R., Rohban, M.H., Baghshah, M.S.: Deciphering the Role of Representation Disentanglement: Investigating Compositional Generalization in CLIP Models, pp. 35–50. ECCV (2025). https://doi.org/10.1007/978-3-031-73024-5_3
6. Wang, Y., Xu, S., Zhu, X., Li, Y.: MSCI: addressing CLIP’s inherent limitations for compositional zero-shot learning. In: Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, pp. 2009–2017. International Joint Conferences on Artificial Intelligence Organization, California (2025). <https://doi.org/10.24963/ijcai.2025/224>
7. Radford, A., et al.: Learning transferable visual models from natural language supervision. In: Proceedings of the 38th International Conference on Machine Learning, vol. 139, pp. 8748–8763. PMLR (2021) [Online]. Available: <https://proceedings.mlr.press/v139/radford21a.html>
8. Miyawaki, S., Hasegawa, T., Nishida, K., Kato, T., Suzuki, J.: Scene-text aware image and text retrieval with dual-encoder. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics, pp. 422–433 (2022). <https://doi.org/10.18653/V1/2022.ACL-SRW.34>
9. Johnson, J., et al.: Image retrieval using scene graphs. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3668–3678. IEEE (2015). <https://doi.org/10.1109/CVPR.2015.7298990>
10. Schroeder, B., Tripathi, S.: Structured query-based image retrieval using scene graphs. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, pp. 680–684 (2020). <https://doi.org/10.1109/CVPRW50498.2020.00097>
11. Li, L., Chen, L., Huang, Y., Zhang, Z., Zhang, S., Xiao, J.: The devil is in the labels: noisy label correction for robust scene graph generation. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 18847–18856 (2022). <https://doi.org/10.1109/CVPR52688.2022.01830>
12. Jung, D., Kim, S., Kim, W.H., Cho, M.: Devil’s on the edges: selective quad attention for scene graph generation. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 18664–18674 (2023). <https://doi.org/10.1109/CVPR52729.2023.01790>
13. Zhou, H., Yang, Y., Luo, T., Zhang, J., Li, S.: A unified deep sparse graph attention network for scene graph generation. *Pattern Recogn.* **123**, 108367 (2022). <https://doi.org/10.1016/j.patcog.2021.108367>
14. Maheshwari, P., Chaudhry, R., Vinay, V.: Scene graph embeddings using relative similarity supervision. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, no. 3, pp. 2328–2336 (2021). <https://doi.org/10.1609/aaai.v35i3.16333>
15. Gordo, A., Almazán, J., Revaud, J., Larlus, D.: Deep Image Retrieval: Learning Global Representations for Image Search, pp. 241–257. ECCV (2016). https://doi.org/10.1007/978-3-319-46466-4_15
16. Noh, H., Araujo, A., Sim, J., Weyand, T., Han, B.: Large-scale image retrieval with attentive deep local features. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 3476–3485. IEEE (2017). <https://doi.org/10.1109/ICCV.2017.374>
17. Li, J., Li, D., Xiong, C., Hoi, S.: BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. In: Proceedings of the 39th International Conference on Machine Learning, vol. 162, pp. 12888–12900. PMLR (2022)
18. Zellers, R., Yatskar, M., Thomson, S., Choi, Y.: Neural motifs: scene graph parsing with global context. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5831–5840. IEEE (2018). <https://doi.org/10.1109/CVPR.2018.00611>
19. Suhail, M., et al.: Energy-based learning for scene graph generation. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 13931–13940 (2021). <https://doi.org/10.1109/CVPR46437.2021.01372>

20. Im, J., Nam, J., Park, N., Lee, H., Park, S.: EGTR: extracting graph from transformer for scene graph generation. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 24229–24238 (2024). <https://doi.org/10.1109/CVPR52733.2024.02287>
21. Egger, M.K., et al.: ReliK: a reliability measure for knowledge graph embeddings. In: WWW 2024—Proceedings of the ACM Web Conference, pp. 2009–2019 (2024). https://doi.org/10.1145/3589334.3645430/SUPPL_FILE/RFP0660.MP4
22. Yang, Y., Chen, J., Xiang, Y.: A review on the reliability of knowledge graph: from a knowledge representation learning perspective. *World Wide Web* **28**(1), 4 (2024). <https://doi.org/10.1007/S11280-024-01316-W>
23. Krishna, R., et al.: Visual genome: connecting language and vision using crowdsourced dense image annotations. *Int. J. Comput. Vis.* **123**(1), 32–73 (2017). <https://doi.org/10.1007/S11263-016-0981-7/TABLES/9>
24. Lin, T.Y., et al.: Microsoft COCO: Common objects in context. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 8693, No. Part 5, pp. 740–755. LNCS (2014). https://doi.org/10.1007/978-3-319-10602-1_48
25. He, T., Gao, L., Song, J., Cai, J., Li, Y.-F.: Learning from the Scene and Borrowing from the Rich: Tackling the Long Tail in Scene Graph Generation [Online]. Available: <https://github.com/htlsn/issg> (2020). Accessed 13 Dec 2025
26. Nguyen, T.S., Yang, H., Fernando, B.: Effective scene graph generation by statistical relation distillation. In: *Proceedings 2025 IEEE Winter Conference on Applications of Computer Vision, WACV 2025*, pp. 8420–8430 (2025). <https://doi.org/10.1109/WACV61041.2025.00816>
27. Li, H., et al.: Scene graph generation: a comprehensive survey. *Neurocomputing*. **566**, 127052 (2024). <https://doi.org/10.1016/j.neucom.2023.127052>