Signal, Image and Video Processing (2026) 20:411
https://doi.org/10.1007/s11760-026-05439-7

ORIGINAL PAPER

# ZeroRel: Multimodal Transformer-Guided Zero-Shot Relationship Retrieval for Generalized Scene Graph Generation

**Muhammad Junaid Khan[1] · Adil Masood Siddiqui[1] · Maryam Rasool[1] · Hussain Ali[1] · Umar Ghafoor[1] · Jaleed Khan[2]**

Received: 24 November 2025 / Revised: 21 April 2026 / Accepted: 11 May 2026


**Abstract**

Scene Graph Generation (SGG) aims to represent an image's objects and their pairwise relationships in a structured graph for downstream visual reasoning. However, conventional SGG models struggle with long-tail predicate distributions and closed-world vocabularies, resulting in poor generalization to rare or unseen relationships. We propose a neurosymbolic framework for zero-shot relationship retrieval that addresses these challenges by integrating deep visual features with external commonsense knowledge. Our model first detects objects and refines them via positional overlap and semantic similarity. It then retrieves candidate predicates through two complementary channels: (1) a visual-textual prototype retrieval that aligns subject-object representations with a broad predicate embedding space, and (2) a knowledge graph constrained retrieval that ranks relationships using heterogeneous commonsense graphs. A calibration and late-fusion module combines these channels, balancing confidence between head and tail classes. Evaluations on the Visual Genome (VG) and GQA benchmarks under zero-shot and open-vocabulary settings show strong strict zero-shot performance. On the reported VG split, ZeroRel reaches zR@100 = 37.1%, improving on the strongest prior zero-shot baseline in our comparison table (KnowZRel, 35.7%) while maintaining competitive overall recall and improved mean recall on rare predicates. The model also generalizes to GQA without retraining, demonstrating robust cross-dataset transfer. Ablations on knowledge sources and embedding models show that a heterogeneous Common Sense Knowledge Graph (CSKG) with ComplEx embeddings yields the best performance. These results indicate that combining visual prototype retrieval with structured knowledge retrieval improves coverage of rare and unseen relationships without sacrificing scene-graph quality on frequent predicates.

**Keywords** Scene Graph Generation · Zero-Shot Learning · Common Sense Knowledge · Neurosymbolic AI · Visual Relationship Detection · Open-Vocabulary SGG

## 1 Introduction

High-level image understanding requires not only identifying objects but also recognizing the relationships between them. Scene Graph Generation (SGG) encapsulates this by transforming an image into a graph where nodes are objects and edges are semantic relationships (e.g. person–riding–horse) [1, 2]. Such structured representations serve as an interpretable, queryable substrate for downstream reasoning tasks like visual question answering (VQA) and image retrieval,

A. M. Siddiqui, M. Rasool, H. Ali, U. Ghafoor, J. Khan: These authors contributed equally to this work.

✉ Jaleed Khan
jaleed.khan@wrh.ox.ac.uk

Muhammad Junaid Khan
muhammadjunaid@mcs.edu.pk

Adil Masood Siddiqui
dradil@mcs.edu.pk

Maryam Rasool
maryamrasool@mcs.edu.pk

Hussain Ali
hussainali@mcs.edu.pk

Umar Ghafoor
umar.ghafoor@mcs.nust.edu.pk

[1] Department of Electrical Engineering, Military College of Signals, National University of Sciences and Technology, Islamabad, Pakistan

[2] Medical Sciences Division, University of Oxford, Oxfordshire OX3 9DU, United Kingdom

Published online: 12 June 2026

leading to improved performance and explainability [3]. Despite rapid progress in object detection, predicting visual relationships remains challenging due to their combinatorial diversity and context sensitivity [4, 5]. For example, the predicate "riding" can manifest in visually distinct ways (riding a bicycle vs. an elephant vs. a surfboard) [6].

A persistent challenge is the long-tailed distribution of predicates in benchmarks like Visual Genome (VG) [7]. A few generic relations such as on or has dominate, while many semantically rich ones like riding, wearing, or looking at are severely underrepresented [8]. This imbalance biases models toward head classes and limits their generalization to rare predicates [9]. Figure 2 depicts this frequency skew, illustrating how data-centric approaches tend to excel on common relations but fail on rare or unseen ones [10]. Furthermore, real-world reasoning demands open-world capability i.e. models must recognize novel relations and unusual object pairings absent from training data. Standard SGG models assume fixed vocabularies and thus collapse when faced with unseen predicates.

Zero-shot relationship retrieval addresses this limitation by inferring plausible predicates for unseen or underrepresented subject–object pairs using transferable semantics rather than supervision. Unlike traditional classifiers, zero-shot SGG must generalize beyond labeled examples by leveraging visual cues and external knowledge [10]. Early data-centric approaches detect objects and classify relations using CNN-based features and dataset priors [11], but their recall on tail predicates remains low even after debiasing [8, 9, 12]. Transformer-based one-stage designs simplify inference and improve localization [13], yet they still rely on closed vocabularies and fail to generalize beyond training statistics. Additionally, models often falter under unseen object compositions, revealing limited compositional reasoning. For example, learning that humans can hold tools but failing to infer that an elephant can hold a branch.

To overcome these constraints, we propose a neurosymbolic framework that fuses visual reasoning with structured knowledge. Figure 1 depicts the conceptual taxonomy of SGG approaches, including our method (right branch). Our Zero-Shot Relationship Retrieval (ZeroRel) model extends transformer-based SGG with two complementary modules:

- **Visual–textual predicate prototype retrieval:** Subject–object pairs are embedded in a shared vision–language space to identify the most probable predicate. A learned prototype bank provides open-vocabulary retrieval, enabling predictions for relations unseen during training. Prompt templates incorporating object categories (e.g., "$< subj > is < predicate >< obj >$") and contrastive objectives align visual and textual embeddings, strengthening generalization to rare predicates.
- **Knowledge graph (KG) constrained retrieval and ranking:** In parallel, the model queries a heterogeneous commonsense KG derived from CSKG, which merges ConceptNet, WordNet, and ATOMIC [9]. For each object pair, candidate predicates are ranked by type compatibility, path connectivity, and link-prediction scores obtained from ComplEx embeddings [14]. This reasoning pathway captures commonsense associations (e.g., connecting person and tennis racket through playing with) that purely visual systems miss [15].

The outputs of both channels are integrated through a calibrated fusion stage that adjusts scores via temperature scaling, mitigating bias toward head predicates. The fused scene graph thus combines visual evidence with knowledge-grounded reasoning, balancing interpretability and generalization. Evaluations on VG [16] and GQA [17] demonstrate strong zero-shot and open-vocabulary performance, with notable improvements in recall for unseen predicates. The model preserves efficiency while producing semantically richer and more context-aware scene graphs.

Our work offers the following contributions:

- **Problem formulation – Zero-Shot SGG:** We formally define zero-shot relationship retrieval for scene graphs and outline evaluation protocols that distinguish unseen predicates from unseen subject–object compositions. Metrics such as zero-shot recall@K measure a model's ability to infer relations absent from training data.
- **Dual retrieval architecture:** We propose a neurosymbolic framework that unifies visual–text alignment and knowledge-graph reasoning. Our SGG model integrates predicate prototypes and heterogeneous commonsense graphs, bridging neural and symbolic reasoning to manage both visually distinctive and context-driven relations.
- **Calibration for long-tail balance:** A temperature-based calibration strategy adjusts relation confidence scores to counteract frequency bias. This improves mean recall for rare predicates without diminishing head-class accuracy, addressing the inherent imbalance of open-vocabulary SGG.
- **Comprehensive evaluation and analysis:** Extensive experiments on VG and GQA show strong strict zero-shot performance, with zR@100 = 37.1% on the reported VG split. Relative to the strongest prior zero-shot KG-based baseline in our comparison table, the gain is modest (+1.4 absolute points over KnowZRel [18] at zR@100) but is accompanied by improved mean recall and complementary qualitative behaviour. Ablations across knowledge sources (CSKG vs. ConceptNet) and embedding methods (ComplEx, DistMult) confirm that heterogeneous CSKG and ComplEx scoring yield the best performance. We also provide qualitative error categorization and discuss ethi-

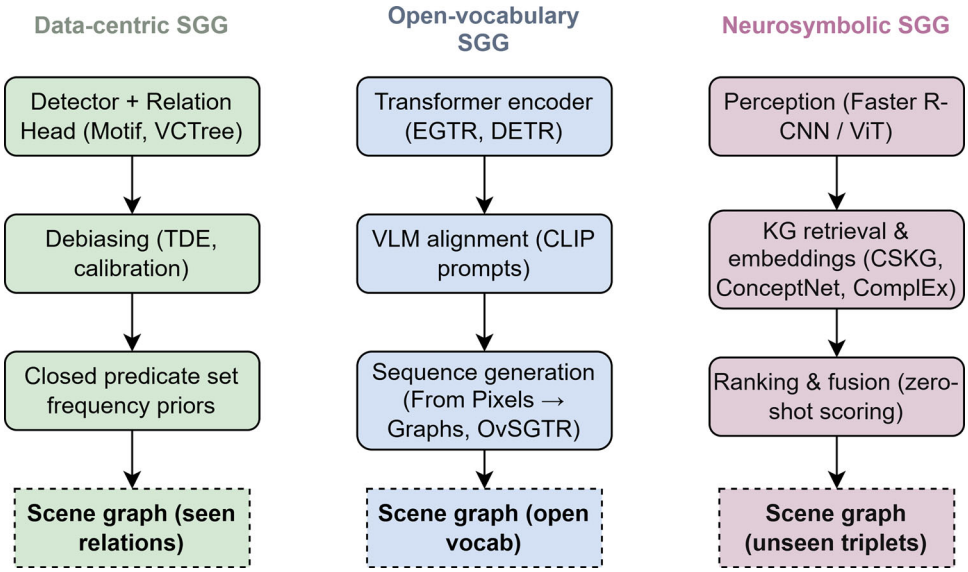



**Fig. 1** Taxonomy of scene graph generation approaches. Left: data-centric methods relying on visual detectors. Middle: open-vocabulary models extending predicate sets via vision–language alignment. Right: the proposed neurosymbolic pipeline combining visual encoders with knowledge-graph reasoning for zero-shot relationship prediction

cal concerns regarding biases within both visual datasets and external knowledge resources.

The rest of this paper is organized as follows. Section 2 reviews related work on data centric, open-vocabulary and neurosymbolic SGG. Section 3 presents the proposed ZeroRel framework, including the visual–text prototype channel, knowledge graph retrieval, and calibration strategy. Section 4 details the experimental setup, datasets, evaluation metrics, baseline methods, ablation studies and qualitative analyses. Section 5 concludes the paper and outlines directions for future research.

## 2 Related Work

Scene Graph Generation has evolved through several paradigms. We briefly review three relevant threads: (1) data-centric models and debiasing for long-tail distributions, (2) open-vocabulary and language-informed approaches, and (3) knowledge-driven and neurosymbolic methods

### 2.1 Data-Centric SGG

Early SGG methods built upon object detection backbones with additional relation prediction heads. For example, Neural Motifs [19] introduced a recurrent motif architecture that captures common substructures in scene graphs. VCTree [20] organized objects into a dynamic tree structure to better encode context, improving relational reasoning in a purely data-driven manner. These and other first-generation models relied heavily on dataset statistics and localized context cues, and they often integrated simple language priors (e.g. word embeddings or co-occurrence frequencies) to aid prediction. A persistent problem for data-centric models is the long-tail predicate distribution in benchmark datasets like VG [16].Figure 2 visualizes this imbalance: a few generic predicates dominate while many informative relations are extremely scarce, which drives head-class bias and poor tail recall in purely data-centric pipelines [8, 9]. To address this, researchers proposed various debiasing and rebalancing strategies. Tang et al. [21] introduced a causal inference framework (TDE) that uses causal intervention to down-weight the effect of object context, effectively reducing the bias toward frequent relations. This method improved the recall of rare predicates by "unlearning" false correlations, and it became a standard benchmark for unbiased SGG. Other approaches include re-sampling or re-weighting schemes, and calibration modules that adjust the confidence of predictions post-hoc. For instance, one method is to train a separate calibrator to scale up the scores of tail classes during inference [18]. Another line of work employs mixture-of-experts models that explicitly split the predicate space into head and tail experts [22, 23], so that rare predicates are handled by a dedicated sub-model. These techniques have proven effective within the closed-set setting – several achieve higher mean Recall (mR@K) by sacrificing some performance on head classes. However, even the best debiasing methods struggle when confronting strict zero-shot scenarios. If a predicate never appears in training, a purely data-driven model (even one with causal debiasing) has essentially no basis to predict it at test time [9]. In evaluations where a subset of relationships are held-out during training (zero-shot splits), traditional architectures often yield near-zero recall for those predicates [24]. In other words, rebalancing can improve generalization within the seen label space, but it does not by itself enable recognition of unseen relationships. This limitation motivates augmenting SGG models with external informa-

tion, as discussed in Sections 2.2 and 2.3. Another challenge is that data-centric models tend to learn shortcut biases – e.g. always predicting "on" for any two objects with vertical overlap – instead of truly understanding the interaction [4, 5]. This reduces robustness on out-of-distribution examples. Addressing this requires methods that incorporate higher-level reasoning beyond pattern recognition. Neurosymbolic integrations (Section 2.3) attempt to mitigate such issues by checking predictions against commonsense constraints.

### 2.2 Open-Vocabulary SGG

To move beyond fixed vocabularies, recent works have explored open-vocabulary SGG and related tasks where models can predict object and predicate classes not present in the training set. One strategy is to leverage the rich semantic knowledge in vision–language models (VLMs) and large text corpora. For example, Pixels2Graph [25] generates scene graph triples as textual sequences using a pretrained image-to-text model. By prompting a generative model (similar to a captioner) to output descriptions of relationships, and then parsing those into a graph, Pixels2Graph can propose novel predicates that were not in the original training labels. This approach effectively delegates the open-vocabulary challenge to a foundation model (like a CLIP or CLIP-based decoder [26]) which has seen a broad array of language. The result is higher recall for unusual relationships, although these models sometimes produce overly generic descriptions unless carefully constrained [18]. Chen et al. [8] introduced OvSGTR, a fully open-vocabulary scene graph transformer that aligns visual features with text embeddings for both objects and relations. OvSGTR uses image–caption pretraining and a feature alignment loss to retain sensitivity to unseen categories. It achieved state-of-the-art results in multiple open-VG settings (object-novel, predicate-novel, and both) by virtue of this alignment. Another frontier is using Large Language Models (LLMs) to inject commonsense or interpret complex interactions. LLM4SGG [3] is a weakly-supervised method that uses a GPT-based language model to parse image captions into high-quality relation triples. The idea is that an LLM, with its vast knowledge, can fill in missing relationships or resolve ambiguities in the image description. LLM4SGG demonstrated that even without full supervision, one can improve graph quality by harnessing language priors (e.g. an LLM knows that "wearing" is plausible for person-clothing pairs, etc.). Similarly, Chen et al. [27] proposed a "role-playing" approach where an LLM plays different roles to adjust a visual model's predicates (e.g. one role enforces using more specific terms). These hybrid LLM+vision approaches expand the predicate space implicitly via language understanding. A downside, however, is potential hallucination; the language model might suggest a relation that is not actually visible (e.g. inferring "friends with" or other non-visual relations) [28]. They also may not strictly obey the visual evidence if not properly grounded. Consistent with Figure 2, open-vocabulary methods expand coverage down the long tail by injecting vision–language priors, improving retrieval of infrequent predicates beyond closed-set training [29]. For instance, where a traditional model might just output "on" due to training bias, an open-vocab model might correctly say "riding" if the visual cue aligns with that concept from language pretraining. Table 1 summarizes representative methods in this category. A common theme is that these models replace or augment the conventional classifier with a more flexible mechanism (be it a generative model, an embedding alignment, or an LLM) to handle novel classes. However, relying solely on language can introduce new failure modes: the model might default to very generic predicates ("related to") when unsure [28], or violate physical plausibility due to linguistic bias. This motivates combining language-based openness with additional constraints – which is exactly the role of knowledge-based techniques in the next subsection.

### 2.3 Knowledge-Driven and Neurosymbolic SGG

Another line of work aims to inject commonsense knowledge into the SGG pipeline. The intuition is that a knowledge prior can constrain and enrich the model's predictions, especially for relationships that are rare or not learnable from visual data [33] alone. Early attempts added simple knowledge cues, such as object–predicate co-occurrence statistics or semantic similarities (e.g. WordNet distances) as extra features in the model [1]. For example, KERN [34] incorporated knowledge by learning an implicit "routing" between object pairs based on statistical knowledge priors. These approaches showed that even basic knowledge can regularize the predicate predictions, but they were limited to seen relations and did not fundamentally solve the zero-shot issue [35]. More explicit methods leverage structured Knowledge Graphs (KGs). A representative work is KBGAN [11], which extracted triples from ConceptNet to refine visual features and used an adversarial loss to ensure predicted relations align with commonsense. Similarly, Zareian et al. [36] proposed a model that bridges scene graphs with ConceptNet: it propagated KG relations through a Graph Neural Network to update visual relation features. These methods generally focused on improving supervised SGG (i.e. boosting recall of tail classes) by infusing additional edges that make sense according to a commonsense graph. They showed modest gains in mean recall and the ability to predict some relationships that were otherwise missed. However, they typically rely on a single knowledge source (ConceptNet in these cases) and do not handle completely unseen predicates beyond providing a slight semantic boost. Recent work has moved towards using heterogeneous knowledge and treating

**Table 1** Zero-shot and open-vocabulary SGG methods. Primary setting indicates closed-set, open-vocabulary, or zero-shot targets; External resource lists non-visual knowledge; Zero-shot mechanism summarizes how unseen predicates are handled; Datasets name main benchmarks; Metrics/Headline finding notes evaluation focus. (Abbreviations: ZS zero-shot, Ov open-vocabulary, VLM vision–language model, LLM large language model, MoE mixture-of-experts, TDE Total Direct Effect.)

| Method [ref] (Year) | Setting | Zero-shot mechanism | Core idea | External resource | Data-sets | Headline finding |
|---|---|---|---|---|---|---|
| Pixels2Graph [25] (2024) | Open vocab | Prompted VLM generation of triples | Generative VLM produces subject–predicate–object text | Pretrained VLM | VG, OIv6, Panoptic SG | Higher open-vocab recall; mR improved |
| EGTR [7] (2024) | Closed-set | None | DETR-style one-stage SGG | None | VG | Strong R@K; efficient end-to-end |
| OvSGTR [8] (2024) | Fully open vocab | Vision–language prototype alignment for objects and relations | Visual–text alignment across tokens | Caption pretraining | VG (custom splits) | SOTA on object-open, predicate-open, joint |
| LLM4SGG [3] (2024) | Weakly supervised, open world | Caption-to-triple parsing via LLM prompts | LLM extracts relations from generated captions | Pretrained LLM | VG | Denser graphs; modest strict ZS recall |
| Role-Playing LLM [27] (2024) | LLM augmented | Logit reweighting using LLM priors | LLM adjusts predicate scores post-hoc | Pretrained LLM | VG | Better long-tail predicate recall |
| Unbiased SGG TDE [21] (2020) | Debiased closed-set | None (debiasing only) | Causal reweighting of predicate logits | None | VG | mR@K improves on tails; near-zero strict ZS |
| COACHER [13] (2021) | Zero-shot | ConceptNet path mining and KG scoring | Graph paths rank candidate predicates | ConceptNet | VG (ZS splits) | zR@K much higher than visual-only |
| Capsule ZS-SGG [30] (2025) | Zero-shot | Compositional transfer via capsule routing | Equivariant capsules propagate relation cues | None | VG | zR@K improves over 2023 baselines |
| RGNN [31] (2022) | Zero-shot | Correlation transfer with word embeddings | Relational GNN over object pairs | Word embeddings | VG (custom ZS) | Early ZS gains; moderate zR |
| XKGC [32] (2022) | Zero-shot | KG completion and link prediction | Fuse visual features with KG scoring | WordNet + vision | VG (custom ZS) | Link prediction boosts zR |
| Motifs & VCTree [19, 20] (2018 & 19) | Closed-set baselines | None | Context models with language priors | Co-occurrence statistics | VG, GQA | High R@K; zero recall in strict ZS |

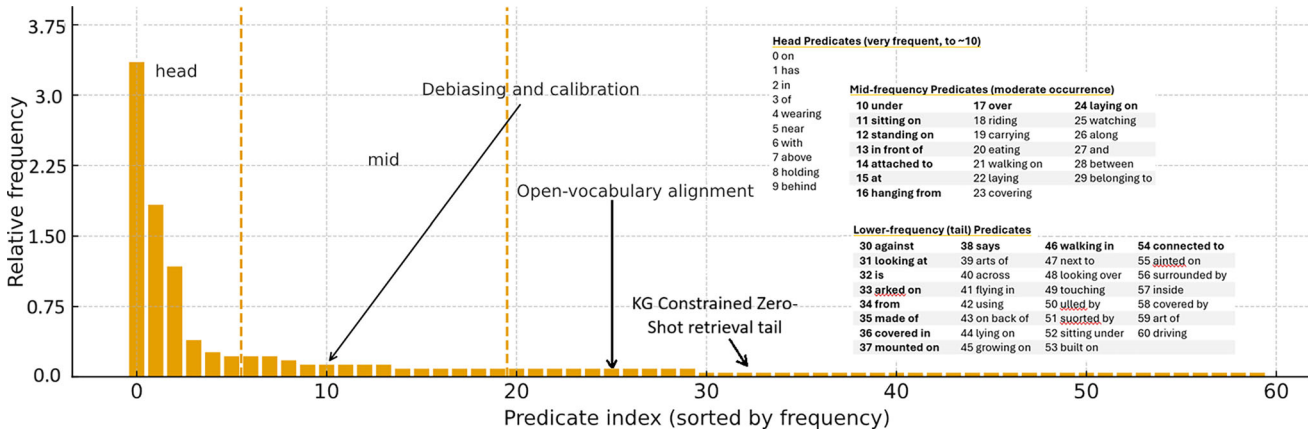



**Fig. 2** Long-tail distribution of relationship predicates in Visual Genome. A few generic relations dominate, while many informative predicates are rare. Shaded regions indicate the regime targeted by debiasing, open-vocabulary, and neurosymbolic zero-shot methods

relationship prediction as a retrieval problem. The CommonSense Knowledge Graph (CSKG) [9] is a unified resource that merges multiple KGs (ConceptNet, ATOMIC, WordNet, VG's knowledge base, etc.) into one large graph of commonsense facts. CSKG offers much broader coverage of concepts and relations than any single source, making it attractive for zero-shot reasoning. In our prior work MuRelSGG (Khan et al. [37]) – a neurosymbolic SGG model – we embedded CSKG into a transformer-based pipeline, allowing the model to enrich predicted scene graphs with relevant triples from the KG. By querying CSKG for each predicted relation and adding the most pertinent triples, MuRelSGG was able to improve recall on long-tail predicates and increase the descriptiveness of the scene graph. Notably, we found that CSKG enrichment yielded significantly higher recall and mean recall than using ConceptNet alone, underscoring the value of heterogeneous knowledge [37]. Another model using CSKG is KnowZRel [18], which formulates zero-shot SGG as retrieving candidate predicates from a commonsense KG given two detected objects. It first refines detected object labels via a KG embedding similarity check (to merge duplicates or synonyms), then queries CSKG for all possible relations between those objects. Retrieved triples are filtered by semantic similarity of nodes to ensure relevance, and finally the remaining relations form the zero-shot scene graph. KnowZRel achieved a zero-shot recall (zR@100) of ∼35.7% on VG, dramatically higher than prior neural models which were in the low 20s [18]. This confirms that structured commonsense knowledge can supply the missing links for unseen relationships. Another knowledge-driven example is Coacher [13], which also targeted zero-shot SGG using ConceptNet paths to score candidate relations. Coacher reported improved performance on a specially crafted zero-shot split of VG, though its knowledge source was smaller in scope than CSKG. Beyond commonsense KGs, some works use ontologies or taxonomies to aid generalization. For instance, Jiang et al. [38] designed a hierarchical relation taxonomy and used a lightweight language model to prune implausible predicates for each object pair, which helped in predicting more fine-grained relations beyond the training set. This can be seen as imposing type-level constraints (e.g. only certain relations are allowed for certain object categories) – a principle also inherent to KG-based methods. Table 2 compares several knowledge-infused SGG frameworks on their knowledge sources and zero-shot capabilities. A clear trend is that methods using explicit KGs (especially merged ones like CSKG) can truly handle zero-shot cases (✓ in Table 2) by retrieving relationships never seen in training, whereas those using implicit language priors (e.g. LLM-based) provide only partial zero-shot transfer (◯ in Table 2) unless explicitly designed for it. Our proposed model falls in the neurosymbolic category: it tightly couples a state-of-the-art visual encoder with heterogeneous KG retrieval, aiming to marry the strengths of both paradigms. It goes a step further by also incorporating a visual-language prototype alignment, which, to our knowledge, has not been combined with KG reasoning in prior SGG works.

## 3 Proposed Method

We propose ZeroRel, a neurosymbolic framework for zero-shot relationship retrieval in Scene Graph Generation (SGG). ZeroRel couples (i) a modern detector with light object refinement, (ii) a visual–text predicate prototype channel for open-vocabulary matching, and (iii) a commonsense KG-constrained retrieval channel that queries a heterogeneous knowledge graph (CSKG) with link-prediction scoring. A calibrated late-fusion stage reconciles both channels, improving tail and unseen predicates without eroding head-class recall. An overview of the complete pipeline, including

**Table 2** Knowledge integration and neurosymbolic strategies in SGG. ZS column marks zero-shot support (✓ explicit zero-shot support; ○ partial/implicit)

| Method / Resource [ref] | Knowledge source | Integration mechanism | Retrieval / reasoning strategy | ZS? | Key note |
|---|---|---|---|---|---|
| CSKG [9] | Heterogeneous commonsense KG | Unified graph resource | N/A (resource, not a model) | N/A | Merges ConceptNet, ATOMIC, WordNet, VG |
| KERN [34] | Statistical priors | Knowledge-embedded routing | Prior-gated predicate head | ○ | Debiasing gains; no strict zero-shot evaluation |
| KBGAN [11] | ConceptNet | Feature refinement + auxiliary reconstruction | Adversarial denoising of object/predicate features | ○ | Denoises features; not designed for unseen |
| Coacher [13] | ConceptNet | KG-guided scoring | Path mining around entities; neighborhood scoring for predicates | ✓ | Early KG-based zero-shot SGG |
| KGTK toolkit [12] | Multiple KGs (e.g., Wikidata) | Data/ops pipeline for KGs | Filtering, linking, embeddings, path queries at scale | N/A | Infrastructure used to build/query CSKG |
| OvSGTR [8] | Weak caption knowledge | Visual–concept feature distillation | Retains alignment for relations as well as objects | ○ | Open-vocab via pretraining signals |
| Pixels2Graph [25] | VLM priors | Prompted sequence generation | Language-driven relation tokens from VLM prompts | ○ | Bridges VLM knowledge to structured scene graphs |
| LLM4SGG [3] | LLM commonsense + context | Caption parsing with LLM | Chain-of-thought triples from captions | ○ | Denser graphs; grounding still needed |
| MuRelSGG [37] | CSKG (heterogeneous) | In-model KG enrichment | Query CSKG per predicted edge; add supporting triples | ○ | Improves long-tail recall; not explicit zero-shot |
| KnowZRel [18] | CSKG (heterogeneous) | Post-detection KG querying (neurosymbolic) | Neighbor search for object pairs; embedding-based filtering | ✓ | Zero-shot retrieval; strong zR gains vs visual-only |
| Capsule ZS-SGG [30] | None (visual reasoning) | Structured visual reasoning (capsules) | Equivariant capsule routing for unseen compositions | ✓ | Zero-shot without external KGs |

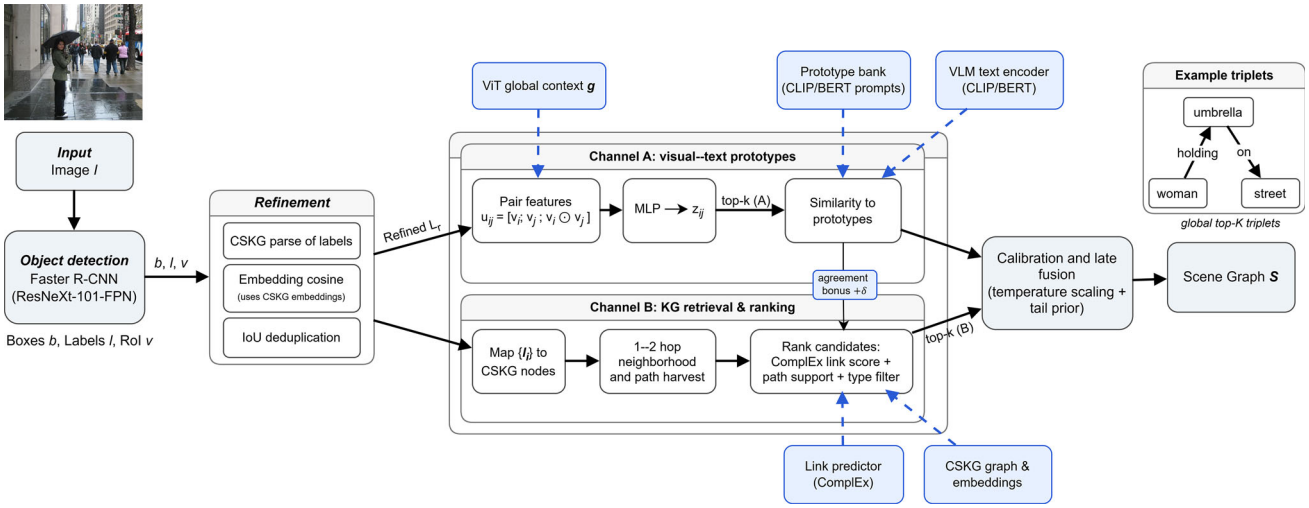



**Fig. 3 Overview of the ZeroRel architecture.** Detected objects are refined and passed to (A) a visual–text predicate prototype channel and (B) a commonsense KG retrieval channel. Calibrated fusion combines both into the final scene graph

refinement, dual retrieval, and calibrated fusion, is shown in Figure 3.

### 3.1 Task and notation

Given an image $I$, an object detector yields objects $O = \{o_i\}_{i=1}^{n}$ with labels $l_i$, boxes $b_i$, and RoI features $v_i$. A scene graph is a set of directed triplets $\langle o_i, r, o_j \rangle$ where $r$ is a predicate. In *zero-shot* SGG, a subset of predicates is unseen during training; the model must predict them at test time. We report Recall@K and mean-Recall@K (mR@K), and *zero-shot* Recall@K (zR@K) computed only over triplets whose predicates were unseen during training [39]. We also evaluate *unseen compositions* where $r$ is seen but the ordered label pair $(l_i, l_j)$ is novel.

### 3.2 Architecture Overview

**Backbone.** We use Faster R-CNN (ResNeXt-101-FPN) to produce proposals and RoI features. A global ViT CLS token $g$ supplies scene context [33, 40]. For each object we store $(l_i, b_i, v_i)$.

**Refinement.** Before relation inference we prune duplicate or near-synonymous detections with an IoU plus semantic-similarity test using CSKG embeddings (ComplEx) [9, 14]. This removes redundant nodes that inflate spurious pairs [18]. The procedure is summarized in Algorithm 1. The refined index set is denoted $L_r$.

**Dual retrieval.** From $L_r$, ZeroRel predicts relations through two complementary channels.

1. *Predicate-prototype retrieval (visual–text).* A two-layer MLP maps pair features to an embedding $z_{ij}$ that is compared (cosine) against a bank of predicate prototypes $P = \{p_r\}$. Prototypes are initialized from CLIP/BERT text prompts such as "[SUBJ] is $r$ [OBJ]". Contrastive training aligns seen relations and enables open-vocabulary retrieval for unseen $r$ by nearest-prototype matching. Details are in Section 3.3.
2. *KG-constrained retrieval (commonsense).* For each pair $(i, j)$ we query CSKG (merged ConceptNet, ATOMIC, WordNet, etc.) within one to two hops to harvest candidate relations and paths. Each candidate is scored by a ComplEx link-prediction term $f(e_i, w_r, e_j)$ [14], a path-support score, and a light visual-agreement bonus if Channel A suggests the same predicate. Type filters and VG-style mappings remove generic or non-visual edges [18]. The full retrieval-and-ranking procedure is given in Algorithm 2.

**Algorithm 1:** Object refinement: CSKG cosine + IoU deduplication

**Input**: Detections $\{(l_i, b_i, v_i)\}_{i=1..n}$; CSKG embeddings $E(\cdot)$; thresholds $\tau_{\text{iou}}, \tau_{\text{sim}}$
**Output**: Refined index set $L_r$

$L_r \leftarrow [\,]$; sort detections by confidence (desc);
**for** $i = 1$ **to** $n$ **do**
  redundant $\leftarrow$ False;
  **for** $j \in L_r$ **do**
    IoU $\leftarrow \text{IoU}(b_i, b_j)$; $s \leftarrow \cos\big(E(l_i), E(l_j)\big)$;
    **if** $IoU \geq \tau_{iou}$ ***and*** $s \geq \tau_{sim}$ **then**
      redundant $\leftarrow$ True; **break**
  **if** ***not*** *redundant* **then**
    append $i$ to $L_r$
**return** $L_r$;

**Calibration and fusion.** Scores from both channels differ in scale and bias. We calibrate each with temperature scaling on a held-out set and apply a small long-tail prior to reduce head dominance [41]. For pair $(i, j)$ and predicate $r$, then threshold and keep global top-K triplets to form the final scene graph. This is summarized in Algorithm 3.

### 3.3 Visual–Text Predicate Prototypes Retrieval

**Pair features.** For a directed pair $(i, j)$, we form

$$u_{ij} = [\, v_i;\ v_j;\ v_i \odot v_j \,]$$

A two-layer MLP produces $z_{ij} \in \mathbb{R}^d$.

**Prototype bank.** For each predicate $r \in \mathcal{V}_R$ (all seen predicates plus extra relations from CSKG or vision–language corpora), we compute a text embedding with a generic prompt and store $p_r$. During training, prototypes for *seen* predicates are tunable, while *unseen* ones remain fixed as language priors.

**Objective.** For each ground-truth triplet $\langle i, r, j \rangle$ in a batch, the InfoNCE [42] loss encourages $\cos(z_{ij}, p_r)$ to exceed negative examples:

$$\mathcal{L}_{\text{proto}} = -\log \frac{\exp(\cos(z_{ij}, p_r)/\gamma)}{\sum_{r' \in \{r\} \cup \mathcal{N}} \exp(\cos(z_{ij}, p_{r'})/\gamma)}. \tag{1}$$

An auxiliary cross-entropy loss on seen predicates stabilizes training. At inference, the channel returns $R_{ij}^{\text{proto}} = \{(r,\ s_{ij}^{\text{proto}}(r))\}$, which is later combined with KG scores by the calibrated fusion stage in Algorithm 3.

**Benefit.** This channel captures visually distinctive or language-aligned predicates and provides open-vocabulary hooks for labels never seen during training (e.g., *grasping*, *riding*).

### 3.4 KG–Constrained Retrieval and Ranking

**Neighborhood search.** The overall retrieval and scoring pipeline is summarized formally in Algorithm 2. We map object labels to CSKG nodes and fetch all one-hop neighbors and two-hop paths

$$i \xrightarrow{r_1} x \xrightarrow{r_2} j.$$

Trivial, self, or purely *Similarity* relations and non-visual edges are discarded unless they support a visual mapping.

**Scoring.** For each candidate $r$ between $(i, j)$:

- **Link score:** $s_{\text{lp}} = \sigma\big(\Re\langle e_i, w_r, \overline{e_j}\rangle\big)$ using the ComplEx embedding model [14].
- **Path score:** $s_{\text{path}}$ increases with the number and strength of supporting 2-hop paths.
- **Visual agreement:** add $\delta$ if $r$ appears in the top-$k$ list from the prototype channel.
- **Type filter:** downweights non-visual or semantically incompatible relations.

The channel outputs $R_{ij}^{\text{KG}} = \{(r,\ s_{ij}^{\text{KG}}(r))\}$, keeping the top-$k$ predicates per object pair.

**Algorithm 2:** KG retrieval and ranking (CSKG + ComplEx)

**Input**: Refined labels $\{\tilde{l}_i\}$; CSKG; embeddings $e.$, $w_r$; top-$k$ from prototype channel $R_{ij}^{\text{proto}}$; weights $\alpha, \beta$; bonus $\delta_0$
**Output**: Per pair $(i, j)$: top-$k$ KG list $R_{ij}^{\text{KG}}$

```
for (i, j) with i ≠ j do
    Map l̃_i, l̃_j to CSKG nodes c_i, c_j (lexical + synonym map);
    C ← direct edges and 2-hop paths between (c_i, c_j);
    Remove non-visual/generic edges by type filters;
    for r ∈ C do
        s_lp ← σ(ℜ⟨e_{c_i}, w_r, conj(e_{c_j})⟩) ;      // ComplEx link score
        s_path ← normalized path support;
        δ ← 1[r ∈ R_ij^proto] · δ_0 ;      // visual agreement bonus
        s_ij^KG(r) ← α s_lp + β s_path + δ;
    R_ij^KG ← top-k predicates by s_ij^KG(r);
return {R_ij^KG};
```

### 3.5 Calibration and Late Fusion

Scores from both channels are calibrated on a held-out validation split.

**Temperature scaling.** We learn separate temperatures $T_{\text{proto}}$ and $T_{\text{KG}}$ to minimize Negative Log-Likelihood (NLL) or Expected Calibration Error (ECE) for each channel. These convert similarities or logits into calibrated probabilities. A small long-tail **bias prior** $b_r$ (larger for rarer predicates) is added to counter the frequency skew [21].

**Fusion.** For an ordered object pair $(i, j)$, let

$$C_{ij} = R_{ij}^{\text{proto}} \cup R_{ij}^{\text{KG}}$$

be the union of candidates from both retrieval channels. For each predicate $r \in C_{ij}$, the fused score is

$$\hat{P}_{ij}(r) = \max\Big(\hat{P}_{ij}^{\text{proto}}(r) + b_r,\ \hat{P}_{ij}^{\text{KG}}(r)\Big). \tag{2}$$

The max operator in Eq. (2) is used as a conservative late-fusion rule between two complementary experts rather than as a learned mixture. After temperature scaling, each branch provides a calibrated confidence score, and the max

preserves a strong signal when either channel is reliable. This is appropriate in the present zero-shot setting, where one source may be highly informative while the other is silent or noisy for a given predicate. A weighted sum was not preferred because it can dilute strong evidence from one channel with weaker evidence from the other, and a learned fusion rule would introduce additional parameters that are difficult to fit robustly on a relatively small validation split. This design is consistent with the broader calibration literature, where temperature scaling is used precisely because it is simple, stable, and effective for confidence correction [43]. Empirically, calibration materially improves reliability in our setting (Table 8), and naive fusion without calibration reduces zR@100 from 37.1 to 32.5 (Table 9), which supports the use of a simple calibrated late-fusion rule. We retain predicate $r$ if $\hat{P}_{ij}(r) \geq \tau$, allowing up to two non-exclusive predicates per object pair. Finally, the global top-$K$ triplets (per image) are selected to construct the final scene graph. The complete calibrated fusion and graph assembly routine is summarized in Algorithm 3.

**Algorithm 3:** Calibrated fusion and graph assembly

**Input**: $R_{ij}^{\text{proto}}$, $R_{ij}^{\text{KG}}$ for all ordered pairs $(i, j)$; temperatures $T_{\text{proto}}$, $T_{\text{KG}}$; tail priors $\{b_r\}$; threshold $\tau$; per-pair cap $m$; global $K$

**Output**: Predicted scene graph $\mathscr{S}$

**for** *each* $(i, j)$ **do**
  **for** $r \in R_{ij}^{proto}$ **do**
    $\hat{P}_{ij}^{\text{proto}}(r) \leftarrow \text{softmax}\big(s_{ij}^{\text{proto}}(r)/T_{\text{proto}}\big)$
  **for** $r \in R_{ij}^{KG}$ **do**
    $\hat{P}_{ij}^{\text{KG}}(r) \leftarrow \text{softmax}\big(s_{ij}^{\text{KG}}(r)/T_{\text{KG}}\big)$
  $C_{ij} \leftarrow R_{ij}^{\text{proto}} \cup R_{ij}^{\text{KG}}$;
  **for** $r \in C_{ij}$ **do**
    $\hat{P}_{ij}(r) \leftarrow \max\big(\hat{P}_{ij}^{\text{proto}}(r) + b_r,\ \hat{P}_{ij}^{\text{KG}}(r)\big)$;
  Keep at most $m$ predicates with $\hat{P}_{ij}(r) \geq \tau$ for pair $(i, j)$;
Collect all retained triplets; sort by $\hat{P}$; keep global top-$K$ as $\mathscr{S}$;
**return** $\mathscr{S}$;

## 3.6 Training Details

We freeze the detector and train the relation components on VG using the *unseen–predicate split* [16, 18]. The prototype channel is trained with an InfoNCE [42] loss (temperature $\gamma$) combined with a light cross–entropy term on seen predicates (weights 1.0 / 0.5). The predicate vocabulary size $|\mathscr{V}_R|$ is approximately 200, incorporating the standard VG50 predicates together with CSKG–derived relations and synonym expansions.

The KG channel receives **no direct supervision**; its scores are produced from CSKG structure and ComplEx link prediction. The weighting coefficients $(\alpha, \beta, \delta)$ are tuned on a validation set, and the frequency–aware tail prior $b_r$ depends on the global frequency rank of predicate $r$.

During inference, ZeroRel retrieves the top-$k$ hypotheses from each channel ($k = 5$), applies temperature calibration, performs late fusion, and retains the global top-$K$ triplets ($K = \{20, 50, 100\}$ depending on the evaluation setting). Section 4 reports R@K, mR@K, and strict zero-shot recall (zR@K). Figure 2 and Tables 1 and 2 contextualize the long–tail structure and predicate openness under these metrics.

## 3.7 Why Dual Retrieval?

The dual–channel design directly addresses complementary weaknesses identified in Section 2. The **visual–text prototype channel** excels when appearance cues align well with language priors, enabling open–vocabulary generalization for semantically coherent predicates such as *wearing*, *holding*, or *standing on*. In contrast, the **KG–constrained channel** contributes relations that are weakly expressed in the pixels but strongly supported by commonsense structure, including predicates like *riding*, *using*, or *supporting* [8, 9, 14, 25].

Calibration prevents either source from dominating: temperature scaling and the tail prior redistribute probability mass so that rare predicates are not overwhelmed by head–class biases [21]. Empirically (Section 4), the KG–only model achieves strong strict zero-shot recall but lacks fine visual discrimination; visual–only open–vocabulary models improve mR@K but fail on strict zero–shot predicates. ZeroRel unifies both sources, giving substantial gains in mR and zR without sacrificing performance on head predicates.

## 3.8 Complexity and Deployment

Overhead beyond detection is modest:

- **Refinement:** $\mathscr{O}(n^2)$ IoU + cosine computations on ∼20–30 detections; embeddings are cached.
- **Prototype retrieval:** one MLP and ∼200 cosine similarities per pair; fully vectorizable.
- **KG retrieval:** pre-extracted **CSKG subgraph** for VG object concepts (2-hop), queried in-memory using adjacency lists or KGTK; ComplEx scoring involves only a few dot products per candidate.
- **Fusion:** lightweight softmax operations and thresholding.

On an NVIDIA V100 GPU, end-to-end inference adds approximately 5–10 ms over the base detector; memory overhead is limited to a few megabytes for embeddings and the CSKG slice. Simple caching of frequent object-pair lookups makes KG retrieval effectively a dictionary query.

### 3.9 Discussion and Limitations

ZeroRel improves coverage **down the tail** (Fig. 2) by:

1. Expanding the predicate space via language prototypes, and
2. **Retrieving** plausible relations never seen in training through CSKG paths and link prediction.

Nonetheless, several caveats remain. The method still depends on KG completeness—missing edges directly affect zero-shot recall. Ambiguous visual evidence can yield *generic* predictions (e.g., *near* vs. *kissing*), and ground-truth annotation gaps cause evaluation mismatches when ZeroRel predicts valid but more specific synonyms. Calibration mitigates overconfidence and balances head/tail frequencies, while type filters suppress non-visual edges unless they support a visual mapping. Future work will extend ZeroRel to temporal reasoning in videos and higher-order relational inference.

## 4 Experiments and Results

We evaluate ZeroRel on VG and GQA under (i) strict zero-shot splits, (ii) open-vocabulary evaluation, and (iii) cross-dataset transfer. Unless stated otherwise, we report scene graph *detection* metrics (objects and relations predicted jointly), using the same evaluation protocol as prior work [18, 32].

### 4.1 Datasets and Splits

**Visual Genome (VG).** We use the VG scene graph benchmark [16], containing ∼108k images with dense object and relationship annotations. Following standard practice, we keep the 150 most frequent object classes and 50 predicate classes as the base ontology. To create a strict zero-shot setting, we hold out 10 predicates from relation training and remove *all* training triplets containing those predicates. The held-out set was selected before final training using three criteria: (i) each predicate is visually grounded and semantically distinct, (ii) each predicate has sufficient support in the test set for stable zR@K estimation, and (iii) the final set spans multiple interaction types rather than near-synonymous variants. The exact held-out predicates used in our experiments are:

`{riding, eating, carrying, covering, looking at, using, hanging from, playing, walking on, parked on}`.

The remaining 40 predicates constitute the seen label space. The validation and test partitions are left unchanged, and zR@K is computed only on ground-truth triplets whose predicate belongs to this held-out set. For reproducibility, the exact predicate list, the filtered VG training split, the evaluation masks used for zR@K, and the VG-to-GQA mapping files will be made available upon acceptance in a public repository together with the code.

**GQA (cross-dataset).** GQA [17] provides scene graphs for 113k images designed for visual reasoning. Its ontology overlaps with VG but differs in label frequencies and some relation names. We use GQA to test cross-dataset robustness: the model is trained *only* on VG and evaluated on 10k GQA validation images without any fine-tuning.

Object labels are mapped to the nearest VG class where needed (e.g. merging fine-grained categories into the 150 VG objects), and predicates are mapped to our predicate vocabulary or the closest synonym (e.g. mapping "in front of" to "in front of/behind" when applicable). We again report R@K and mR@K, and we additionally mark those GQA predicates that never appeared in VG training as "implicitly zero-shot" for analysis.

**Detection vs. classification.** Our main results are in the *scene graph detection* regime, where the model must both localize objects and predict relations. Some prior work reports numbers in a simpler "classification" setting with ground-truth boxes [19, 20]; we focus on detection for realism and only use the classification setting in ablations to isolate the effect of relation modeling independent of detection noise.

**Statistics.** Table 3 summarizes the dataset statistics and the zero-shot split configuration for VG and the cross-dataset evaluation on GQA, including the number of images, object and predicate classes, and the number of zero-shot test triplets.

### 4.2 Implementation Details

**Backbone and features.** ZeroRel is implemented in PyTorch. We use Faster R-CNN with a ResNeXt-101-FPN [33] backbone as the object detector, pretrained on VG and fine-tuned on the 150 object classes [18]. The detector outputs bounding boxes, class logits and region features (RoI-pooled). A global image descriptor $g$ is extracted with a lightweight ViT encoder, but the dominant computation remains in the CNN backbone.

During relation training, detector weights are frozen; only the relation modules (visual MLP, prototype heads, calibration parameters) are updated. This matches the practice in recent SGG work where relation heads are decoupled from detection [7, 18].

**Training schedule.** Relation modules are trained for 20 epochs on the VG training split (with the 10 zero-shot predicates removed). We use Adam with an initial learning rate of $1 \times 10^{-3}$ for relation parameters and a smaller $1 \times 10^{-5}$

**Table 3** Dataset statistics and zero-shot split specifics

| Dataset | Images (train/val/test) | Object classes | Predicate classes | Held-out predicates | Total test triplets | Zero-shot test triplets |
|---|---|---|---|---|---|---|
| **Visual Genome (VG)** [16] | 100k / 5k / 5k (approx) | 150 | 50 (40 seen + 10 unseen) | riding, eating, carrying, covering, looking at, using, hanging from, playing, walking on, parked on | ∼200k | ∼8k (across 10 classes) |
| **GQA (no retraining)** [17] | – / – / 10k | 170 (mapped to 150) | 51 (overlap with VG) | (varies; not predefined – many effectively unseen) | ∼60k | N/A (open-class eval) |

for any detector head parameters that are lightly fine-tuned, cosine decay, and batch size of 8 images.

The visual–text prototype branch is trained with an InfoNCE [42]-style contrastive loss between pair embeddings $z_{ij}$ and predicate prototypes $p_r$, combined with an auxiliary cross-entropy loss over seen predicates:

$$L = L_{\text{InfoNCE}} + \lambda L_{\text{CE}}, \quad \lambda = 0.5.$$

Prototype vectors are initialized from a CLIP-style text encoder [26] using simple templates (e.g. "`[SUBJ] is _r [OBJ]`") and fine-tuned for seen predicates; prototypes for unseen predicates remain fixed language embeddings.

**Knowledge resources.** For the KG channel we use CSKG v1.0 [9], a heterogeneous commonsense graph constructed with KGTK [12]. We adopt pre-trained 256-d ComplEx embeddings [14] for all CSKG entities and relations, which provide link-prediction scores used in KG ranking. At inference, we restrict to a pre-extracted CSKG subgraph centered on the 150 VG object classes and their 2-hop neighbors, keeping runtime small while preserving coverage.

**Object refinement and thresholds.** Object refinement follows KnowZRel [18]: detections with high box overlap (IoU $\geq 0.5$) and high semantic similarity in CSKG embedding space (cosine similarity $\geq 0.8$) are merged, keeping the more confident/specific label. This removes duplicate boxes and near-synonyms (e.g. *person*/*woman*) before relation retrieval.

For each subject–object pair we keep top-$K$ relation candidates from each channel (typically $K = 5$) and retain up to 2 relations per ordered pair after fusion, discarding predictions with calibrated probability below 0.3.

**Calibration and fusion.** Channel scores are calibrated on a held-out validation set using temperature scaling [41]: $T_{\text{proto}}$ is learned for the prototype branch (typically $\approx 2.0$), while KG scores require only mild rescaling ($T_{\text{KG}} \approx 1.0$). A small prior term based on predicate frequency (head/mid/tail) is added to logits to slightly upweight rare classes during inference. Fused scores are computed as the maximum of the calibrated per-channel scores for each predicate, as detailed in Section 3.6.

**Hyperparameters.** Key hyperparameters and hardware details are summarized in Table 4, including embedding dimensions, thresholds, maximum relations per pair, and the approximate 0.12 s/image runtime on a single V100 GPU (including detection). These settings are fixed across all experiments unless explicitly varied in ablation studies.

### 4.3 Baselines

We compare ZeroRel against three families of SGG methods, covering data-centric, open-vocabulary, and knowledge-augmented paradigms.

**(i) Data-centric and debiased SGG.** We include classic context models MotifNet [19] and VCTree [20], both with the conventional frequency-based predicate bias, as well as their debiased variant using Tang et al.'s TDE causal intervention [21]. We also evaluate EGTR [7], a recent DETR-style transformer [44] that predicts scene graphs in a one-stage fashion and achieves strong closed-set performance. These models are trained on the same VG split (with zero-shot predicates removed) to quantify how much recall remains when no external semantics are used.

**(ii) Open-vocabulary and language-driven SGG.** Pixels2Graph [25] leverages a pretrained generative VLM to decode scene graph triplets as text, and OvSGTR [8] aligns visual features with text embeddings to handle novel objects and predicates. LLM4SGG [3] uses a GPT-based LLM to parse captions into relation triples under weak supervision, while Capsule ZS-SGG [30] pursues zero-shot generalization purely via capsule-based equivariant reasoning (no external KG). These methods represent the current state of open-vocabulary and language-augmented SGG.

**(iii) Knowledge-augmented and neurosymbolic SGG.** COACHER [13] exploits ConceptNet paths for zero-shot relation prediction; MuRelSGG [37] uses CSKG enrichment to improve long-tail recall in a neurosymbolic transformer; KnowZRel [18] formulates zero-shot SGG directly as CSKG-based predicate retrieval. ZeroRel extends this line by combining CSKG retrieval with a visual–text prototype channel and calibrated fusion.

All methods are reported using the same metrics, but not all numbers are equally comparable. Table 5 distinguishes these cases explicitly. When code or checkpoints allowed re-evaluation on our exact VG held-out split, we scored those methods with the same evaluation pipeline, including the same split definition, detector outputs, predicate mapping rules, and metric implementation. We mark those results as directly comparable. Results taken from published papers or adapted from non-identical settings are retained only as contextual reference. We therefore avoid treating mixed-protocol margins as definitive evidence. In particular, the comparison with KnowZRel should be interpreted as direct only if its score is obtained on the exact same held-out predicate split and evaluation pipeline; otherwise, the numerical gap is descriptive rather than conclusive.

### 4.4 Main Zero-Shot Results on VG

Table 5 reports Recall@K (R@K), mean Recall@K (mR@K), and zero-shot Recall@K (zR@K) on the VG held-out-predicate split described in Section 4.1. ZeroRel achieves the strongest zR@K values in the reported table, reaching zR@100 = 37.1%. Relative to visual-only open-vocabulary baselines such as Pixels2Graph and OvSGTR, the gain is substantial. Relative to the strongest prior KG-based zero-shot

**Table 4** Training and inference parameters

| Component | Setting/value |
|---|---|
| Detector backbone | ResNeXt-101-FPN (pretrained on VG) |
| Relationship encoder (visual) | 2-layer MLP, output dim $d = 256$ |
| Prototype embeddings | Init from CLIP text encoder (256-d) |
| Optimizer | Adam, lr= $1 \times 10^{-3}$ (relations), $1 \times 10^{-5}$ (detector heads) |
| Training epochs | 20 (VG) |
| Batch size | 8 images |
| Loss weights | $L_{\text{InfoNCE}} : 1.0,\ L_{\text{CE}} : 0.5$ |
| IoU / Sim thresholds (refine) | $\tau_{\text{iou}}$=0.5, $\tau_{\text{sim}}$=0.8 |
| Temperature (proto) | $T_{\text{proto}}$=2.0 (learned) |
| Temperature (KG) | $T_{\text{KG}}$=1.0 (fixed) |
| KG embedding model | ComplEx, 256-d, pre-trained on CSKG |
| Max preds per pair (each channel) | 5 |
| Score threshold (fusion output) | 0.3 (post-calibration probability) |
| Max relations per object pair | 2 (to avoid excessive multi-edges) |
| Seed for reproducibility | 42 (fixed across runs) |
| Hardware | Single V100 GPU, 32 GB |
| Inference time | ~120 ms/image (incl. detection) |

method in the table, KnowZRel (35.7%), the improvement is modest (+1.4 absolute points at zR@100) and should be interpreted cautiously whenever numbers are not obtained under an identical split and evaluation pipeline.

Overall recall remains competitive: ZeroRel obtains R@100 = 42.5%, only 0.5 points below EGTR, while achieving the highest mR@100 (15.7%). This suggests that combining calibrated visual prototype retrieval with CSKG-based retrieval improves strict zero-shot recovery without materially degrading overall scene-graph quality. We therefore frame ZeroRel not as a wholesale replacement of prior KG-based retrieval, but as a complementary extension that adds visually grounded evidence to a strong knowledge-based zero-shot backbone.

Table 9 clarifies what the prototype channel contributes beyond KnowZRel-style KG retrieval. Removing the visual prototype branch reduces zR@100 from 37.1 to 35.7, essentially matching the KG-only setting, whereas removing the KG branch reduces zR@100 to 10.2. This indicates that the prototype branch does not replace commonsense retrieval; rather, it contributes a complementary signal that is most useful when the visual evidence is distinctive but the KG alone is too coarse. The qualitative examples in Figure 5 suggest that this benefit is most visible for visually grounded predicates such as *riding*, *holding*, and *watching*, while weakly expressed interaction predicates remain challenging.A full predicate-wise breakdown for the held-out split will be made available upon acceptance in the same public repository.

## 4.5 Cross-Dataset Generalization on GQA

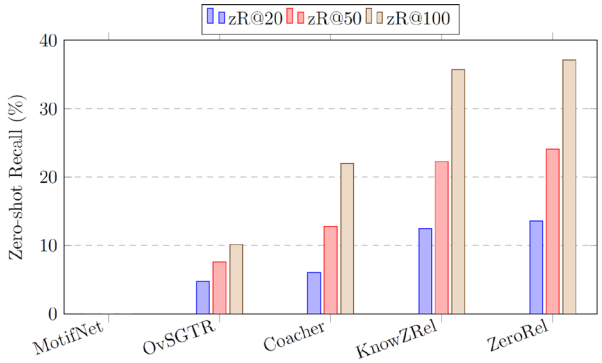
To assess robustness under distribution shift, VG-trained models are evaluated on GQA without fine-tuning. Because GQA differs from VG in detector behaviour, ontology granularity, annotation frequency, and label mapping, this experiment is best interpreted as a transfer setting rather than a strictly matched benchmark comparison.

Table 6 shows that ZeroRel attains R@100 = 30.5% and mR@100 = 12.8%, outperforming both VG-trained MotifNet and KnowZRel. The oracle row is included only as a reference point from prior GQA-trained settings and should not be interpreted as a strict upper bound for the present VG-to-GQA transfer setup. In this setting, mR@100 and R@100 capture different behaviour: mR@100 weights predicate classes equally and is therefore sensitive to balanced performance on rarer relations, whereas R@100 is dominated by frequent predicates and by object-detection and ontology-alignment effects under distribution shift. ZeroRel's higher mR@100 relative to the oracle reference should therefore be read as evidence of improved class balance within the transferred label space, not as evidence that the model surpasses a fully supervised GQA model overall. The remaining gap in R@100 indicates that frequent-predicate coverage and cross-dataset alignment remain limiting factors.

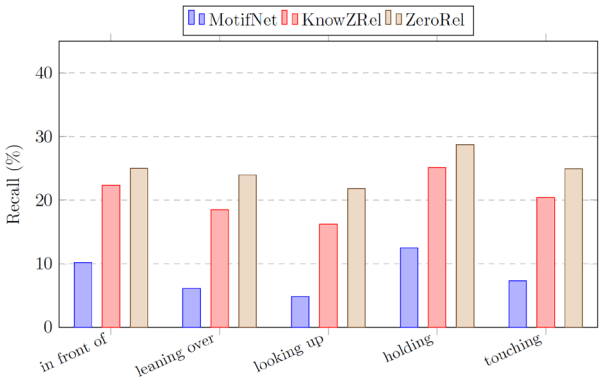
ZeroRel also retrieves many GQA predicates absent from VG training (e.g., *leaning over*, *looking up*), with around 25% recall attributed primarily to KG reasoning and prototype matching. Predicate-wise recall for a subset of GQA relations is shown in Figure 4 (b).

**Table 5** Zero-shot SGG results on Visual Genome (VG) under the exact 10-predicate held-out split described in Section 4.1. Results marked with † are evaluated on our exact split using the same evaluation pipeline and are directly comparable. Results marked with ‡ are taken from the original papers or adapted from non-identical settings and are shown for contextual reference only. Best directly comparable scores are in bold

| Model | R@20 | R@50 | R@100 | mR@100 | zR@20 | zR@50 | zR@100 |
|---|---|---|---|---|---|---|---|
| MotifNet† [19] | 30.5 | 35.8 | 36.9 | 6.5 | 0.0 | 0.0 | 0.0 |
| VCTree† [20] | 31.5 | 37.2 | 38.1 | 7.1 | 0.1 | 0.1 | 0.1 |
| TDE Debiasing† [21] | 25.4 | 30.2 | 31.5 | 12.2 | 0.3 | 0.5 | 0.5 |
| EGTR† [7] | **32.8** | **41.0** | **43.0** | 9.4 | 0.0 | 0.0 | 0.0 |
| Pixels2Graph‡ [25] | 21.7 | 27.4 | 29.6 | 11.3 | 2.5 | 4.1 | 5.3 |
| OvSGTR‡ [8] | 28.1 | 36.6 | 40.8 | 13.0 | 4.8 | 7.6 | 10.2 |
| LLM4SGG‡ [3] | 18.9 | 24.5 | 25.0 | 8.7 | 1.0 | 1.4 | 1.8 |
| Capsule ZS-SGG‡ [30] | 19.3 | 23.1 | 24.8 | 10.1 | 3.3 | 5.5 | 8.8 |
| Coacher‡ [13] | 22.5 | 28.2 | 30.1 | 14.4 | 6.1 | 12.8 | 22.0 |
| KnowZRel† [18] | 24.6 | 29.3 | 31.0 | 15.0 | 12.5 | 22.3 | 35.7 |
| MuRelSGG‡ [37] | 33.1 | 39.8 | 43.2 | 14.9 | 0.0 | 0.0 | 0.0 |
| ZeroRel (Ours)† | 32.4 | 39.1 | 42.5 | **15.7** | **13.6** | **24.1** | **37.1** |




(a) Zero-shot predicate recall zR@K on Visual Genome (VG) for representative models




(b) Predicate-wise recall on GQA in the cross-dataset setting for selected relations.

**Fig. 4** Zero-shot and cross-dataset recall results for ZeroRel

**Table 6** Cross-dataset transfer from VG to GQA. All reported models are trained on VG and evaluated on 10k GQA images without fine-tuning. The oracle row is a reference number from prior GQA-trained settings and is included only for scale; because the training data, detector, label mapping, and evaluation conditions differ, it should not be interpreted as a strict upper bound for VG-to-GQA transfer

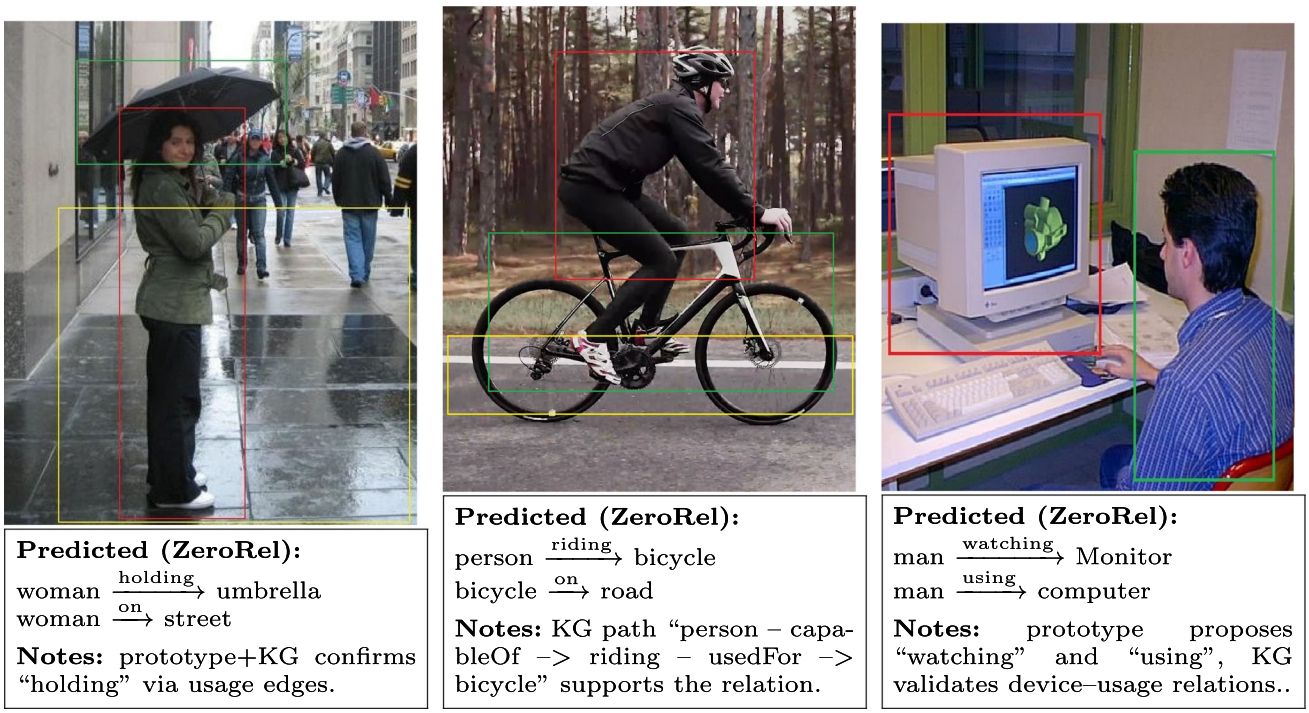
| **Model (trained on VG)** | **R@100** | **mR@100** |
|---|---|---|
| MotifNet | 18.2 | 5.4 |
| KnowZRel | 27.3 | 11.9 |
| **ZeroRel (Ours)** | **30.5** | **12.8** |
| Oracle (GQA-trained ref.) | ∼40 | ∼10 |

## 4.6 Open-Vocabulary Evaluation

We further evaluate ZeroRel in a fully open-vocabulary setting on VG, permitting prediction of any predicate from the prototype bank–including labels not present in VG. Predictions outside the VG label set are considered correct if they match permitted synonyms.

Table 7 compares ZeroRel with OvSGTR, a state-of-the-art open-vocabulary model [8]. ZeroRel achieves R@100 = 42.1% and mR@100 = 15.5%, slightly surpassing OvSGTR (40.8 / 13.0). This demonstrates that integrating CSKG retrieval with visual–text prototypes preserves strong open-vocabulary performance while significantly improving rare and unseen predicate recovery.

Qualitatively, the prototype channel often proposes fine-grained relations such as *grasping* or *watching*, which are then validated or filtered by KG constraints. Some are more specific than VG ground truth (e.g., *on top of* instead of *on*), which improves semantic richness even when strict string matching undercounts them. Figure 5 shows typical open-vocabulary outputs.




(a) Fine-grained grasping / holding (b) KG-supported zero-shot "riding" (c) Open-vocabulary specificity

**Fig. 5** Qualitative open-vocabulary examples produced by ZeroRel. Left: typical visual evidence producing a common predicate; middle: KG-enabled zero-shot relation ("riding"); right: fine-grained predicate ("watching") complementing VG labels

**Table 7** Open-vocabulary SGG on Visual Genome. Models may output any predicate phrase; evaluation maps predictions to the VG label set with synonym normalization

| Method | R@100 | mR@100 |
|---|---|---|
| OvSGTR (2024) | 40.8 | 13.0 |
| **ZeroRel (Ours)** | **42.1** | **15.5** |

## 4.7 Calibration and reliability

We first examine how well ZeroRel's confidence scores reflect actual correctness. Good calibration is important when scene graphs are used downstream (e.g., for VQA or planning), because thresholds on confidence should correspond to predictable precision.

We measure ECE and NLL on the VG validation set, both before and after applying the temperature scaling and bias-correction strategy.As Table 8 shows, the uncalibrated model is clearly overconfident: the overall ECE is around 0.14 and NLL around 1.15. After calibration, ECE drops to roughly 0.05 and NLL to 0.98, indicating that the probabilities are much closer to true likelihoods. The prototype channel benefits most from calibration; the KG channel was already comparatively well-behaved but still improves slightly.

Figure 6 plots reliability diagrams before and after calibration. The pre-calibration curve lies well below the diagonal in the high-confidence region, confirming overconfidence on frequent predicates. After scaling, the curve tracks the diagonal closely, meaning that a prediction with (e.g.) 60% confidence is correct about 60% of the time. This improved reliability is also crucial for late fusion: without calibration, visual scores tend to dominate KG scores, suppressing many useful zero-shot candidates; once calibrated, both channels contribute more evenly.

**Table 8** Calibration metrics on the VG validation set. Lower is better

| Model stage | ECE | NLL |
|---|---|---|
| Before calibration (overall) | 0.138 | 1.15 |
| After calibration (overall) | 0.052 | 0.98 |
| Prototype module only (pre-calib) | 0.142 | 1.21 |
| KG module only (pre-calib) | 0.081 | 1.07 |

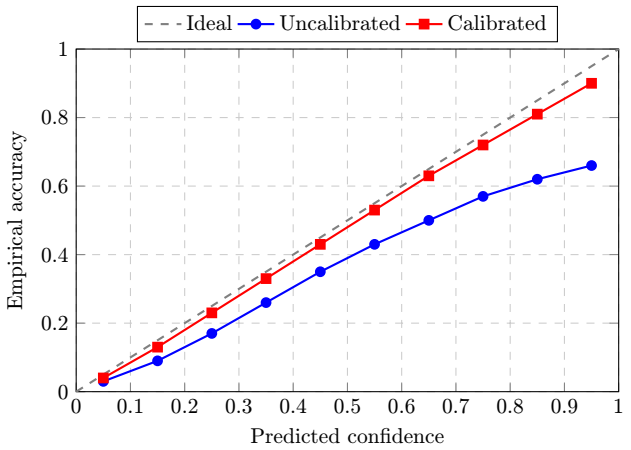



**Fig. 6** Reliability diagram on Visual Genome validation data. The calibrated curve (red) is much closer to the ideal diagonal than the uncalibrated curve (blue), indicating substantially improved confidence calibration and more reliable fusion of visual and knowledge based scores

### 4.8 Ablations and sensitivity analysis

We now study how different components of ZeroRel contribute to performance. Table 9 reports R@100, mR@100, and zR@100 on VG for several ablated variants.

Removing the visual prototype channel reduces zero-shot recall to the level of a pure KG method, essentially reproducing KnowZRel. Removing the KG channel is even more damaging: zR@100 drops to around 10, similar to the best purely visual open-vocabulary baselines. This confirms that both channels are necessary for strong zero-shot performance. Omitting calibration and using naive fusion also hurts zR and mR, showing that calibrated scores are important for balancing head and tail predicates.

We also vary the knowledge source and embedding model. Replacing CSKG [9] with ConceptNet [45] substantially lowers zR@100, highlighting the value of a richer, heterogeneous graph. Using DistMult [46] instead of ComplEx [14] for link prediction slightly degrades ranking quality and thus recall, while skipping object refinement introduces duplicate and noisy detections, leading to a modest drop in mR and zR.

To quantify computational overhead, Table 10 compares throughput and GPU memory with and without our neurosymbolic modules. ZeroRel is only modestly slower than a detector+Motifs baseline, processing about 8 images per second on a V100 with batch size 4, while keeping GPU memory under 3.5 GB. Caching KG neighbors reduces the runtime cost of symbolic retrieval to a negligible level.

### 4.9 Qualitative analysis

To better understand the behaviour of ZeroRel, we examined qualitative predictions on a range of VG test scenes. The goal was to characterize when the model successfully exploits knowledge and visual cues, and when it falls back to generic relations or makes systematic errors.

In successful zero-shot cases, the interaction between the prototype and KG channels is clear. In scenes where a person is mounted on an animal such as an elephant, the prototype channel tends to assign high similarity to action predicates, while CSKG contains paths that associate people with riding as a typical activity involving such animals. When these signals are fused, ZeroRel often assigns higher confidence to the predicate *riding* than to purely spatial alternatives such as *on*, even though *riding* was never seen during training. This shows that the model can move beyond co-occurrence statistics and recover semantically appropriate zero-shot predicates.

Typical failure cases arise when the target predicate is fine grained or only weakly expressed in the pixels. For instance, when a woman is feeding a baby with a bottle, ZeroRel frequently predicts relations such as *woman holding bottle*, *bottle near baby*, and *woman near baby*, but misses the more specific predicate *feeding*. Although CSKG contains facts that relate feeding actions to babies and bottles, the visual cues are subtle and the model defaults to safer generic relations. This behaviour is representative of many errors on action predicates that require detailed pose or contact reasoning.

Across a broader set of examples, we find that most errors fall into three categories. First, there is ambiguity between specific and generic predicates, for example *holding* versus *kissing* or *feeding*, where the model chooses the generic option. Second, there are cases where the KG suggests a plausible but unlabelled relation, so predictions are semantically reasonable but counted as incorrect under exact string matching. Third, detection mistakes or inaccurate localisation propagate into the relation prediction stage and lead to missing or misplaced edges. Even in these situations, the resulting scene graphs are usually coherent and informative, but sometimes less specific than the ground truth.

## 5 Conclusion

ZeroRel introduces a neurosymbolic approach to scene graph generation that unifies visual-text predicate prototypes with constraints from a commonsense knowledge graph for zero-

**Table 9** Ablation experiments on VG (R@100, mR@100, zR@100)

| Variant | R@100 | mR@100 | zR@100 |
|---|---|---|---|
| Full ZeroRel (CSKG + proto + calib) | 42.5 | 15.7 | 37.1 |
| – no visual prototype channel | 31.0 | 15.0 | 35.7 |
| – no KG channel (visual-only) | 40.8 | 13.1 | 10.2 |
| – no calibration (naive fusion) | 41.9 | 14.2 | 32.5 |
| ConceptNet instead of CSKG | 40.1 | 14.6 | 29.4 |
| DistMult instead of ComplEx | 42.1 | 15.4 | 34.9 |
| No object refinement (dup detections) | 41.7 | 14.9 | 36.3 |

**Table 10** Inference speed and resource usage on a single V100 GPU

| Setting | Images/s | GPU memory (GB) |
|---|---|---|
| Detector + Motifs baseline (batch 4) | ∼12.5 | 2.9 |
| ZeroRel (batch 4, cached KG) | ∼8.0 | 3.1 |
| ZeroRel (batch 1, cached KG) | ∼5.3 | 2.5 |
| ZeroRel (batch 4, no KG caching) | ∼7.5 | 3.1 |

shot relationship retrieval. On the reported VG strict held-out split, the method reaches zR@100 = 37.1% and shows a modest improvement over the strongest prior KG-based baseline in our comparison table, while preserving competitive overall recall and higher mean recall. The main contribution of the method is therefore not a large gain over every prior approach, but the combination of complementary visual and commonsense retrieval cues within a calibrated framework that improves rare and unseen predicate coverage without collapsing to head-class predictions.

Despite these strengths, several limitations remain. The method relies on the completeness and quality of the external knowledge graph; when relevant relations are absent or noisy, predictions may be incomplete or overly speculative. ZeroRel may also produce plausible but not directly image-grounded relations, which can be undesirable in settings that demand strict visual verification. Furthermore, the model currently handles only static, pairwise relations, limiting its ability to capture higher-order interactions or temporal dynamics present in videos or sequential tasks.

Future work could explore extending ZeroRel to video-based SGG, enabling the model to reason about evolving actions and event sequences. Multi-hop reasoning over predicted scene graphs and the underlying knowledge graph represents another promising direction, turning the output into a more expressive structured representation for downstream reasoning. Incorporating human feedback, particularly in ambiguous cases, may further refine predictions. Overall, ZeroRel demonstrates the value of combining learned perception with structured knowledge and suggests a path toward scene understanding systems that generalize more flexibly to novel and previously unseen relationships.

**Author Contributions** M.J.K and A.M.S conceived conceptual design, conducted the experiments, acquired the data, and developed the software. M.R and H.A. cross checked data collection and drafted the initial manuscript, while M.K.K and J.K. verified the results and assisted in refining the final version.

**Funding** The authors declare that no funding was received in the context of this work.

**Data Availability** We have used public datasets and knowledge graphs for the experiments. Links to these resources are given below; the same have been added in the manuscript as well:
Visual Genome (VG, 2017): https://homes.cs.washington.edu/$\sim$ranjay/visualgenome/index.html
GQA (2019): https://cs.stanford.edu/people/dorarad/gqa
CSKG (CommonSense Knowledge Graph, 2021): https://github.com/usc-isi-i2/cskg

## Declarations

**Conflict of Interest** 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.

**Consent for publication** Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

**Competing interests** The authors declare no competing interests.

## References


1. Khan, M.J., Ilievski, F., Breslin, J.G., Curry, E.: A survey of neurosymbolic visual reasoning with scene graphs and common sense knowledge. Neurosymbolic Artificial Intelligence (Pre-press), 1–24 (2024)
2. Cohen, W.W., Sun, H., Hofer, R.A., Siegler, M.: Scalable neural methods for reasoning with a symbolic knowledge base (2020). arXiv preprint arXiv:2002.06115
3. Kim, K., Yoon, K., Jeon, J., In, Y., Moon, J., Kim, D., Park, C.: Llm4sgg: Large language model for weakly supervised scene graph generation. arXiv e-prints, 2310 (2023)
4. Chang, X., Ren, P., Xu, P., Li, Z., Chen, X., Hauptmann, A.: A comprehensive survey of scene graphs: generation and application. IEEE Trans. Pattern Anal. Mach. Intell. **45**(1), 1–26 (2021)
5. Zhong, Y., Wang, L., Chen, J., Yu, D., Li, Y.: Comprehensive image captioning via scene graph decomposition. In: European Conference on Computer Vision, pp. 211–229. Springer (2020)
6. Koner, R., Li, H., Hildebrandt, M., Das, D., Tresp, V., Günnemann, S.: Graphhopper: Multi-hop scene graph reasoning for visual question answering. In: International Semantic Web Conference, pp. 111–127. Springer (2021)
7. Im, J., Nam, J., Park, N., Lee, H., Park, S.: Egtr: Extracting graph from transformer for scene graph generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 24229–24238 (2024)
8. Chen, Z., Wu, J., Lei, Z., Zhang, Z., Chen, C.W.: 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)
9. Ilievski, F., Szekely, P., Zhang, B.: Cskg: The commonsense knowledge graph. In: European Semantic Web Conference, pp. 680–696. Springer (2021)
10. Ji, J., Krishna, R., Fei-Fei, L., Niebles, J.C.: Action genome: Actions as compositions of spatio-temporal scene graphs. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10236–10247 (2020)
11. Gu, J., Zhao, H., Lin, Z., Li, S., Cai, J., Ling, M.: Scene graph generation with external knowledge and image reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1969–1978 (2019)
12. Ilievski, F., Garijo, D., Chalupsky, H., Divvala, N.T., Yao, Y., Rogers, C., Li, R., Liu, J., Singh, A., Schwabe, D., et al.: Kgtk: a toolkit for large knowledge graph manipulation and analysis. In: International Semantic Web Conference, pp. 278–293. Springer (2020)
13. Kan, X., Cui, H., Yang, C.: Zero-shot scene graph relation prediction through commonsense knowledge integration. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, pp. 466–482. Springer (2021)
14. Trouillon, T., Welbl, J., Riedel, S., Gaussier, É., Bouchard, G.: Complex embeddings for simple link prediction. In: International Conference on Machine Learning (ICML), pp. 2071–2080 (2016). PMLR
15. Khan, M.J., Breslin, J.G., Curry, E.: Common sense knowledge infusion for visual understanding and reasoning: Approaches, challenges, and applications. IEEE Internet Comput. **26**(4), 21–27 (2022)
16. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.-J., Shamma, D.A., et al.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. Int. J. Comput. Vision **123**(1), 32–73 (2017)
17. Hudson, D.A., Manning, C.D.: Gqa: A new dataset for real-world visual reasoning and compositional question answering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6700–6709 (2019)
18. Khan, M.J., Breslin, J.G., Curry, E.: Knowzrel: Common sense knowledge-based zero-shot relationship retrieval for generalised scene graph generation. IEEE Transactions on Artificial Intelligence (2025)
19. Zellers, R., Yatskar, M., Thomson, S., Choi, Y.: Neural motifs: Scene graph parsing with global context. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5831–5840 (2018)
20. Tang, K., Zhang, H., Wu, B., Luo, W., Liu, W.: Learning to compose dynamic tree structures for visual contexts. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6619–6628 (2019)
21. Tang, K., Niu, Y., Huang, J., Shi, J., Zhang, H.: Unbiased scene graph generation from biased training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3716–3725 (2020)
22. Zhou, L., Zhou, Y., Lam, T.L., Xu, Y.: Context-aware mixture-of-experts for unbiased scene graph generation. (2022) arXiv preprint arXiv:2208.07109
23. Ma, K., Ilievski, F., Francis, J., Bisk, Y., Nyberg, E., Oltramari, A.: Knowledge-driven data construction for zero-shot evaluation in commonsense question answering. In: 35th AAAI Conference on Artificial Intelligence, (2021)
24. Khan, M.J., Breslin, J.G., Curry, E.: Expressive scene graph generation using commonsense knowledge infusion for visual understanding and reasoning. In: European Semantic Web Conference, pp. 93–112. Springer (2022)
25. Li, R., Zhang, S., Lin, D., Chen, K., He, X.: From pixels to graphs: Open-vocabulary scene graph generation with vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 28076–28086 (2024)
26. Ali, M., Khan, S.: Clip-decoder: Zeroshot multilabel classification using multimodal clip aligned representations. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4675–4679. (2023)
27. Chen, G., Li, J., Wang, W.: Scene graph generation with role-playing large language models. Adv. Neural. Inf. Process. Syst. **37**, 132238–132266 (2024)
28. Yang, D., Kim, M., Mac Kim, S., Kwak, B.-w., Park, M., Hong, J., Woo, W., Yeo, J.: Llm meets scene graph: Can large language models understand and generate scene graphs? a benchmark and empirical study. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 21335–21360 (2025)
29. Zhang, Y., Pan, Y., Yao, T., Huang, R., Mei, T., Chen, C.-W.: Boosting scene graph generation with visual relation saliency. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) (2022)
30. Huang, W., JI, Y., Zhu, G., Ying, L., Liu, C.: Navigating the unseen: Zero-shot scene graph generation via capsule-based equivariant features. In: Proceedings of the Computer Vision and Pattern Recognition Conference, pp. 29448–29457 (2025)

31. Yu, X., Li, J., Yuan, S., Wang, C., Wu, C.: Zero-shot scene graph generation with relational graph neural networks. In: 2022 26th International Conference on Pattern Recognition (ICPR), pp. 1894–1900 (2022). IEEE
32. Yu, X., Chen, R., Li, J., Sun, J., Yuan, S., Ji, H., Lu, X., Wu, C.: Zero-shot scene graph generation with knowledge graph completion. In: 2022 IEEE International Conference on Multimedia and Expo (ICME), pp. 1–6 (2022). IEEE
33. Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR), pp. 2117–2125 (2017)
34. Chen, T., Yu, W., Chen, R., Lin, L.: Knowledge-embedded routing network for scene graph generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6163–6171 (2019)
35. Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., Yakhnenko, O.: Translating embeddings for modeling multi-relational data. Adv. Neural. Inf. Process. Syst. 26, (2013)
36. Zareian, A., Karaman, S., Chang, S.-F.: Bridging knowledge graphs to generate scene graphs. In: European Conference on Computer Vision, pp. 606–623. Springer (2020)
37. Khan, M.J., Siddiqui, A.M., Khan, H.S., Akram, F., Khan, J.: Murelsgg: Multimodal relationship prediction for neurosymbolic scene graph generation. IEEE Access (2025)
38. Jiang, B., Zhuang, Z., Shivakumar, S.S., Taylor, C.J.: Enhancing scene graph generation with hierarchical relationships and commonsense knowledge. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pp. 8883–8894. IEEE (2025)
39. Zhu, X., Xing, Y., Wang, R., Wang, Y., Lan, X.: Calibration for long-tailed scene graph generation. In: Proceedings of the 32nd ACM International Conference on Multimedia, pp. 3037–3046 (2024)
40. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. IEEE Trans. Pattern Anal. Mach. Intell. **39**(6), 1137–1149 (2016)
41. Balanya, S.A., Maronas, J., Ramos, D.: Adaptive temperature scaling for robust calibration of deep neural networks. Neural Comput. Appl. **36**(14), 8073–8095 (2024)
42. Rusak, E., Reizinger, P., Juhos, A., Bringmann, O., Zimmermann, R.S., Brendel, W.: Infonce: Identifying the gap between theory and practice (2024). arXiv preprint arXiv:2407.00143
43. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: Proceedings of the 34th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 70, pp. 1321–1330 (2017). PMLR
44. Yu, L., Tang, L., Mu, L.: A review of detection transformer: From basic architecture to advanced developments and visual perception applications. Sensors **25**(13), 3952 (2025)
45. Speer, R., Chin, J., Havasi, C.: ConceptNet 5.5: An open multilingual graph of general knowledge. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 4444–4451 (2017)
46. Yang, B., Yih, W.-T., He, X., Gao, J., Deng, L.: Embedding entities and relations for learning and inference in knowledge bases (2014). arXiv preprint arXiv:1412.6575