DUAL-CUE RELATION MODELING WITH SYNTHETIC CUE AUGMENTATION FOR UNBIASED SCENE GRAPH GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Scene Graph Generation (SGG) represents visual scenes as structured subject-predicate-object triplets, yet relation prediction remains challenging due to semantic ambiguity, limited generalization to unseen triplet compositions, and long-tailed predicate distributions. To address these challenges of relational evidence in SGG, we propose a framework of Dual-Cue Relation Modeling (DCRM), which models visual-spatial and semantic-context evidence in separate relation branches. To enhance semantic modeling and compositional generalization, we further incorporate pretrained encoders into this dual-cue framework, leveraging its cross-modal image representations and encoded predicate descriptions to provide semantic guidance and fine-grained predicate prototypes. An adaptive router dynamically coordinates the predictions of the two branches, enabling effective integration of the resulting heterogeneous cues. To address the long-tailed distribution, we further introduce Synthetic Cue Augmentation (SCA), which performs semantically guided local variation transfer from data-rich to underrepresented predicates in both cue spaces. Extensive experiments on VG150, OIV6, and GQA200 demonstrate competitive overall and balanced performance of the proposed framework, with improved long-tailed predicate recognition and zero-shot generalization.

1 INTRODUCTION

Scene Graph Generation (SGG) represents an image as a structured graph of entities and their directed relationships, typically formulated as subject-predicate-object triplets such as <person, riding, horse> (Chang et al., 2023). By extending object detection from categories and locations to relational semantics, SGG provides structured representations for downstream visual reasoning tasks. Among existing paradigms, two-stage methods, including MOTIFS (Zellers et al., 2018) and VCTree (Tang et al., 2019), remain widely studied. They first localize entities with an object detector and then predict relations over candidate object pairs (Jung et al., 2023; Lin et al., 2020; Zheng et al., 2023; Yang et al., 2025).

Despite substantial progress, relation prediction remains challenging under long-tailed and semantically ambiguous predicate distributions. Generic relations such as *on* and *in* dominate the training data, whereas more fine-grained interactions are represented by far fewer samples. Consequently, models tend to favor frequent predicates, while the limited samples of tail classes make it difficult to learn sufficiently robust representations (Tang et al., 2020; Li et al., 2024a). This challenge is further amplified by semantically related predicates such as *on*, *standing on*, and *walking on*, whose visual patterns can substantially overlap, making fine-grained discrimination difficult under sparse supervision (Zheng et al., 2023).

We further observe that relational evidence in SGG is inherently heterogeneous. Visual-spatial cues characterize relations through local appearance and relative geometry, whereas semantic-context cues describe relational patterns from object semantics and contextual information (Lin et al., 2020; Yu et al., 2021; Li et al., 2023; Zhang et al., 2017; Lu et al., 2016). These two forms of evidence capture different aspects of object relations and naturally lend themselves to separate representation spaces. Moreover, semantic cues are commonly derived from object categories or embeddings,

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

offering limited contextual information. Drawing on the dual-branch learning and differentiated supervision of DHL (Zheng et al., 2024), We therefore model these two forms of evidence in separate branches and enrich the latter with pretrained vision-language encoders through cross-modal image semantics and encoded predicate descriptions. Predictions from the two branches are subsequently coordinated at the decision level.

Another challenge in detection-based SGG is generalization to unseen triplet compositions. Models often rely on subject-predicate-object co-occurrence statistics and may therefore struggle with novel compositions, even when the constituent entity and predicate categories have been observed during training. Pretrained vision-language foundation models provide complementary information because large-scale image-text pretraining captures semantic associations beyond the local visual and geometric evidence used in conventional detection-based SGG (Radford et al., 2021; Li et al., 2022a). Such representations are less tied to triplet co-occurrence patterns in the target training set and can support better generalization to unseen compositions. We therefore exploit pretrained vision-language representations to strengthen compositional generalization while retaining detector-based instance grounding, unlike generative SGG paradigms that directly predict graphs with large language or multimodal models (Li et al., 2024c; Kim et al., 2024b).

Within this dual-branch design, we develop Dual-Cue Relation Modeling (DCRM), comprising a visual-spatial branch for local appearance and geometry and a semantic-context branch for object and relation semantics. Pretrained vision-language representations further enrich semantic modeling within the semantic-context branch. A predicate-adaptive router then dynamically coordinates the resulting predictions. To mitigate long-tailed data scarcity, we introduce Synthetic Cue Augmentation (SCA), which performs semantically guided local variation transfer from data-rich predicates to underrepresented ones independently in the two cue spaces.

The main contributions of this work are summarized as follows:

1. We propose Dual-Cue Relation Modeling (DCRM), which separately models heterogeneous visual-spatial and semantic-context evidence in two decoupled relation branches and coordinates their predictions at the decision level through a predicate-adaptive router.

2. We introduce a hybrid SGG framework that combines detector-based instance grounding with pretrained cross-modal representations and text-encoded predicate descriptions, enriching semantic relation modeling and improving compositional generalization.

3. We propose Synthetic Cue Augmentation (SCA), which performs semantically guided local variation transfer from data-rich to underrepresented predicates in both cue spaces, thereby alleviating long-tailed bias and promoting more balanced predictions.

2 Related Work

Current research in SGG mainly focuses on developing effective scene graph generation frameworks (Li et al., 2024b; Yang et al., 2024; Kim & Ko, 2025; Xu et al., 2025) and mitigating prediction bias under long-tailed predicate distributions (Jeon et al., 2024; Kim et al., 2024a; Yang et al., 2025; Gao et al., 2025). Accordingly, we briefly review the most relevant studies from these two perspectives, while additional representative methods are discussed in Appendix A.

2.1 Relation Modeling Frameworks and Representation Learning

Existing two-stage SGG frameworks mainly improve relation modeling through context modeling and message passing, relation representation learning, or multi-branch prediction architectures. Early recurrent, tree-structured, and graph-based methods model contextual dependencies through sequential or graph propagation, including MOTIFS (Zellers et al., 2018), VCTree (Tang et al., 2019), BGNN (Li et al., 2021), and HetSGG (Yoon et al., 2023), among others. Transformer-based frameworks further capture long-range object-relation interactions through attention mechanisms, as exemplified by VETO (Sudhakaran et al., 2023), SQUAT (Jung et al., 2023), and IS-GGT (Kundu & Aakur, 2023). Another line of work focuses more directly on learning discriminative relation representations. PE-Net (Zheng et al., 2023) learns predicate prototypes in an embedding space, DRM (Li et al., 2024b) jointly models predicate- and triplet-level representations, and Adaptive Feature Learning (Yang et al., 2024) improves predicate representations through feature-assisted learning.

Multi-branch frameworks, such as DHL (Zheng et al., 2024), employ coarse- and fine-grained learning branches with differentiated supervision and knowledge distillation to improve long-tailed predicate recognition. Representative methods are further discussed in Appendix A.

Vision-language pretraining has also been increasingly used to introduce transferable cross-modal semantics into relation modeling. Foundation models provide transferable cross-modal representations learned from large-scale image-text data, inspiring SGG methods for generative, weakly supervised, and open-vocabulary settings, such as PGSG (Li et al., 2024c), LLM4SGG (Kim et al., 2024b), and RAHP (Liu et al., 2025). More closely related to unbiased scene graph generation, CaCao (Yu et al., 2023) employs a visually prompted language model to generate diverse fine-grained predicates for underrepresented relations, while FMSGG (Sun et al., 2026) introduces relationship-oriented prompting to isolate relation embeddings from foundation-model features and improve the separability of relation descriptions. In contrast, our framework incorporates foundation-model image representations and encoded predicate descriptions to enrich semantic-context modeling, while visual-spatial evidence is modeled separately and the resulting predictions are adaptively coordinated at the decision level.

2.2 Long-Tailed and Unbiased Scene Graph Generation

Long-tailed predicate distributions cause SGG models to favor frequent and semantically broad relations, while limited supervision makes fine-grained tail predicates difficult to distinguish. Existing methods broadly address this problem from three perspectives: data and supervision rebalancing, loss and prediction adjustment, and feature-space augmentation. Data- and supervision-level approaches improve the effective training distribution by correcting noisy annotations or exploiting unlabeled object pairs. NICE (Li et al., 2022b) corrects noisy predicate labels, while ST-SGG (Kim et al., 2024a) introduces class-adaptive pseudo-labeling to exploit unannotated triplets, and SSC-SGG (Yang et al., 2025) further incorporates semi-supervised clustering for low-frequency relations. Another line adjusts the training objective or prediction distribution. TDE (Tang et al., 2020) applies causal intervention to reduce contextual bias, PCPL (Yan et al., 2020) exploits predicate correlations for adaptive reweighting, CogTree (Yu et al., 2021) organizes predicates hierarchically for coarse-to-fine learning, PPDL (Li et al., 2022d) adjusts supervision through predicate probability distributions, while Probabilistic Debiasing (Biswas & Ji, 2023) and DKBL (Chen et al., 2023) calibrate biased predictions from probabilistic and knowledge-transfer perspectives, respectively.

Feature augmentation provides another direction by enriching the representations of underrepresented predicates. A-PFG (Wang et al., 2023) constructs attributes for predicate features and introduces VAE to learn a shared latent space between predicate features and attributes, from which tail-class features are synthesized for classifier fine-tuning. CFA (Li et al., 2023) instead synthesizes new triplet features by recombining intrinsic and extrinsic feature components. In contrast, our SCA transfers semantically guided local variations from data-rich to underrepresented predicates, with the variation structures modeled independently in the visual-spatial and semantic-context spaces.

3 Method

3.1 Problem Formulation and Framework Overview

Given an input image I, scene graph generation predicts a set of directed subject-predicate-object relations. Let $\mathcal{E}$ denote the candidate relation edge set. The scene graph is formulated as:

$$G = \{\langle s_i, p_{ij}, o_j \rangle \mid (i,j) \in \mathcal{E}\}, \tag{1}$$

where s_i and o_j denote the subject and object instances, respectively, and p_{ij} denotes the predicate.

To characterize relational information from complementary perspectives, we model each candidate pair in two cue-specific representation spaces:

$$G_v = \left\{\langle s_i, \mathbf{p}_{ij}^v, o_j \rangle\right\}_{(i,j)\in\mathcal{E}}, \quad G_s = \left\{\langle s_i, \mathbf{p}_{ij}^s, o_j \rangle\right\}_{(i,j)\in\mathcal{E}}, \tag{2}$$

where $\mathbf{p}_{ij}^v$ and $\mathbf{p}_{ij}^s$ denote the visual-spatial and semantic-context predicate representations. The two relation graphs share identical edge indices and are therefore topologically aligned, while their representations are learned through separate cue branches. The visual-spatial cue mainly captures

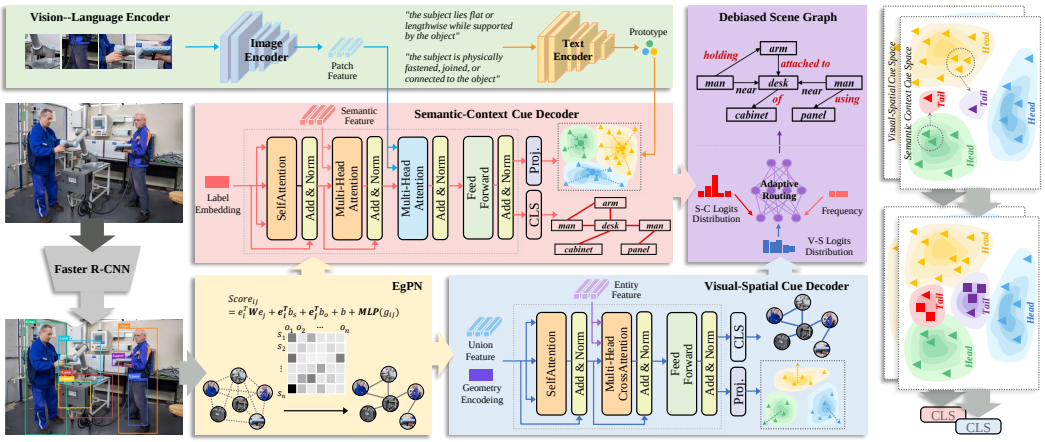


Figure 1: Overview of the proposed framework.

local appearance and geometric information, whereas the semantic-context cue incorporates object semantics and pretrained cross-modal representations.

As illustrated in Figure 1, EgPN first scores directed object pairs and filters out low-confidence candidates to construct a shared relation topology. Based on the retained pairs, Dual-Cue Relation Modeling (DCRM) learns visual-spatial and semantic-context representations in separate branches and adaptively coordinates their predictions at the decision level. Semantically guided learning further improves predicate discrimination, while Synthetic Cue Augmentation (SCA) transfers compatible local variations from data-rich predicates to underrepresented ones in both cue spaces, thereby alleviating the impact of long-tailed predicate distributions.

3.2 Dual-Cue Relation Modeling

DCRM constructs two complementary predicate representations over a shared candidate relation topology and dynamically coordinates their predictions.

Edge Proposal. Before relation modeling, EgPN performs lightweight selection over directed object pairs. For each ordered pair (i, j), its edge score is

$$s_{\text{edge}}(i, j) = e_i^\top W e_j + e_i^\top b_s + e_j^\top b_o + b + \text{MLP}(\phi_{ij}), \tag{3}$$

where e_i and e_j denote object-category embeddings and ϕ_{ij} encodes their relative geometry, mapped by $\text{MLP}(\phi_{ij})$ to a scalar score. The category-dependent terms capture asymmetric subject-object compatibility, while the geometric term provides instance-specific spatial information.

EgPN is optimized using binary cross-entropy:

$$\mathcal{L}_{\text{edge}} = -\frac{1}{|\mathcal{P}|} \sum_{(i,j)\in\mathcal{P}} \left[y_{ij}^e \log \sigma(s_{\text{edge}}(i, j)) + (1 - y_{ij}^e) \log \left(1 - \sigma(s_{\text{edge}}(i, j))\right) \right], \tag{4}$$

where y_{ij}^e indicates whether pair (i, j) has an annotated foreground relation. High-scoring pairs form the shared candidate edge set $\mathcal{E}$, which defines the common topology for the two cue spaces.

Visual-Spatial Cue Modeling. For each candidate pair $(i, j) \in \mathcal{E}$, the detector-derived object feature is constructed as $\mathbf{f}_i^v = F_v([r_i; g_i])$, where r_i and g_i denote the RoI appearance feature and box-geometry encoding. The relation query is initialized from the union representation as $q_{ij}^{v,0} = F_{\text{pair}}^v(u_{ij})$, where u_{ij} combines union-region appearance and pairwise spatial information.

The visual pathway first applies self-attention among candidate relation queries to capture dependencies across relation pairs. The resulting states then attend to the object features $F^v = \{\mathbf{f}_i^v\}$ and are further transformed by a feed-forward block:

$$Q^{v,l+1} = \mathcal{T}_v(Q^{v,l}, F^v), \tag{5}$$

where $\mathcal{T}_v$ summarizes relation self-attention, object-feature cross-attention, and feed-forward transformation. After the final layer, the visual-spatial predicate representation is obtained as $\mathbf{p}_{ij}^{v} = W_v q_{ij}^{v,L}$. It integrates union appearance, relative geometry, and detector-derived object information to provide localized visual-spatial evidence for predicate prediction.

Semantic-Context Cue Modeling. For the same candidate pair, semantic object features are constructed as $\mathbf{f}_i^s = F_s([e_i; g_i])$, and the initial pair query is defined as $q_{ij}^{s,0} = F_{\text{pair}}^s([\mathbf{f}_i^s; \mathbf{f}_j^s])$.

A frozen vision-language pretrained encoder further provides image representations Z. To associate them with individual candidate relations, we pool over the subject, object, union, and global regions:

$$B_{ij} = \{\text{Pool}(Z, b_i), \text{Pool}(Z, b_j), \text{Pool}(Z, b_i \cup b_j), \text{Pool}(Z, I)\}, \quad \mathcal{B} = \{B_{ij}\}_{(i,j)\in\mathcal{E}}. \tag{6}$$

These features provide subject/object information, joint interaction semantics, and global context. The semantic pathway performs relation self-attention and then interacts sequentially with semantic object features $F^s = \{\mathbf{f}_i^s\}$ and pair-specific cross-modal features:

$$Q^{s,l+1} = \mathcal{T}_s(Q^{s,l}, F^s, \mathcal{B}), \tag{7}$$

where $\mathcal{T}_s$ contains relation self-attention, two cross-attention operations, and a feed-forward transformation. The final semantic-context representation $\mathbf{p}_{ij}^{s} = W_s q_{ij}^{s,L}$ combines object-category semantics with region-conditioned and global cross-modal information. Thus, the two cue spaces represent identical candidate relations from different sources while retaining separate structures.

Adaptive Cue Routing. The two representations are independently classified as $l_{ij,c}^{v} = C_v(\mathbf{p}_{ij}^{v})_c$ and $l_{ij,c}^{s} = C_s(\mathbf{p}_{ij}^{s})_c$. Instead of fixed fusion weights, the router estimates cue contributions from relation-level reliability ψ_{ij}^{rel} and class prior η_c^{pre}. The former summarizes branch confidence, decision sharpness, and prediction discrepancy from logits and probabilities, while the latter combines predicate semantics with frequency-aware group information. Their concatenation is mapped by a lightweight MLP to the visual and semantic routing weights for each candidate predicate.

The routing weights and fused logit are as:

$$[\alpha_{ij,c}^{v}, \alpha_{ij,c}^{s}] = \text{Softmax}\left(R([\psi_{ij}^{rel}; \eta_c^{pre}])\right), \quad l_{ij,c} = \alpha_{ij,c}^{v} l_{ij,c}^{v} + \alpha_{ij,c}^{s} l_{ij,c}^{s}. \tag{8}$$

The router therefore adaptively coordinates the two cues using relation-level reliability and predicate-specific class priors.

3.3 Semantically Guided Cue Learning and Optimization

The base framework is jointly optimized through edge proposal learning, predicate-text alignment, cue-specific contrastive learning, and predicate classification.

Predicate Text Alignment. For predicate category c, a frozen text encoder pretrained within the same vision-language model encodes a set of fine-grained descriptions $\{d_{c,\ell}\}_{\ell=1}^{L_c}$ into the normalized text prototype as:

$$\mathbf{p}_c^{\text{text}} = \text{Norm}\left(\frac{1}{L_c}\sum_{\ell=1}^{L_c} E_{\text{text}}(d_{c,\ell})\right). \tag{9}$$

The semantic-context predicate representation is projected as $\bar{\mathbf{p}}_{ij}^{s} = \text{Norm}(W_{\text{text}}\mathbf{p}_{ij}^{s})$ into the text embedding space, where the prototypes serve as semantic anchors for distinguishing closely related predicates. Its alignment logit with predicate c is defined as $l_{ij,c}^{\text{text}} = \cos(\bar{\mathbf{p}}_{ij}^{s}, \mathbf{p}_c^{\text{text}})/\tau_{\text{text}}$. For foreground edges $\mathcal{E}^+$ and foreground predicate set $\mathcal{C}_p^+$, the alignment objective is

$$\mathcal{L}_{\text{text}} = -\frac{1}{|\mathcal{E}^+|}\sum_{(i,j)\in\mathcal{E}^+} \log\frac{\exp(l_{ij,y_{ij}}^{\text{text}})}{\sum_{c\in\mathcal{C}_p^+}\exp(l_{ij,c}^{\text{text}})}. \tag{10}$$

Predicate Classification. The routed output and the two cue predictions are jointly supervised:

$$\mathcal{L}_{\text{cls}} = -(1-p_y)^{\gamma}\log p_y + \text{CE}(l^v, y) + \text{WCE}(l^s, y; w), \tag{11}$$

where p_y denotes the routed target-class probability and w denotes predicate-group weights. The visual branch uses standard CE, while the semantic branch uses WCE to encourage class-balanced recognition.

Cue-Specific Contrastive Learning. We further impose supervised contrastive constraints independently in the two cue spaces. Relation representations with the same predicate label are treated as positives, while those from different categories are treated as negatives. The resulting objective, $\mathcal{L}_{\text{con}} = \mathcal{L}^{v}_{\text{con}} + \mathcal{L}^{s}_{\text{con}}$, encourages intra-class compactness and inter-class separation without forcing the two cue spaces into a shared geometry.

The complete base-training objective is formulated as:

$$\mathcal{L} = \mathcal{L}_{\text{edge}} + \mathcal{L}_{\text{text}} + \mathcal{L}_{\text{con}} + \mathcal{L}_{\text{cls}}. \tag{12}$$

The text-alignment and contrastive projection heads are used only during training and introduce no additional inference-time branch.

3.4 Synthetic Cue Augmentation

Tail predicates contain relatively few instances, limiting the coverage of their intra-class variations. In contrast, data-rich predicates exhibit richer patterns across different subject-object compositions, appearances, and scene conditions. As illustrated in Figure 2, a broadly defined predicate such as *on* may exhibit diverse local variation patterns, some of which can be compatible with more specific relations such as *standing on*, *riding*, and *parked on*. Such compatibility does not imply relabeling between predicates; rather, it suggests that locally observed variations in frequently represented predicates may be directly transferable to semantically related underrepresented predicates. Based on this observation, SCA selectively transfers compatible local variations to underrepresented predicates independently in the two cue spaces. For a target predicate t, SCA first identifies semantically related data-rich predicates using the text prototypes introduced above.



Figure 2: Local variation modes within the predicate *on*.

The semantic affinity and donor set are defined as:

$$s^{\text{sem}}_{t,h} = \cos(\mathbf{p}^{\text{text}}_{t}, \mathbf{p}^{\text{text}}_{h}), \quad \mathcal{D}_t = \text{TopK}_{h\in\mathcal{H}}\, s^{\text{sem}}_{t,h}, \tag{13}$$

where $\mathcal{H}$ denotes the set of data-rich predicates. The predicate-level donor set $\mathcal{D}_t$ is shared by the two cue spaces, restricting variation transfer to semantically related predicates.

Although the donor predicates are shared, their intra-class variation structures are modeled independently in the visual-spatial and semantic-context spaces. For cue $q \in \{v, s\}$, the representations of donor predicate h are grouped into local modes $\{\mathcal{M}^{q}_{h,k}\}_{k=1}^{K^{q}_{h}}$. The center of each mode and its empirical residual bank are defined as:

$$\mathbf{m}^{q}_{h,k} = \frac{1}{|\mathcal{M}^{q}_{h,k}|} \sum_{\mathbf{z}^{q}_{n}\in\mathcal{M}^{q}_{h,k}} \mathbf{z}^{q}_{n}, \quad \mathcal{R}^{q}_{h,k} = \left\{ \mathbf{z}^{q}_{n} - \mathbf{m}^{q}_{h,k} \mid \mathbf{z}^{q}_{n} \in \mathcal{M}^{q}_{h,k} \right\}. \tag{14}$$

Each residual therefore corresponds to a local intra-class variation empirically observed in a data-rich predicate. Since the two cue spaces encode different relation information, their local variation structures are discovered independently.

For an instance i of target predicate t, SCA selects a compatible donor mode separately in each cue space by combining predicate-level semantic affinity with cue-space similarity:

$$\gamma^{q}_{i,h,k} = \lambda_{\text{sem}} s^{\text{sem}}_{t,h} + \lambda_{\text{cue}} \cos(\mathbf{z}^{q}_{i}, \mathbf{m}^{q}_{h,k}), \quad (h^{*}_{q}, k^{*}_{q}) = \arg\max_{h\in\mathcal{D}_t,k} \gamma^{q}_{i,h,k}. \tag{15}$$

The semantic term constrains the donor predicate range, while the cue-space similarity determines which local variation pattern is most compatible with the target instance.

A residual $\mathbf{r}^q \in \mathcal{R}^q_{h^*_q,k^*_q}$ is then transferred to the target representation as:

$$\widetilde{\mathbf{z}}^q_i = \mathbf{z}^q_i + \mathrm{clip}_{\tau^q_t}(\eta \mathbf{r}^q), \quad \mathrm{clip}_\tau(\mathbf{r}) = \mathbf{r} \min\left(1, \frac{\tau}{\|\mathbf{r}\|_2}\right), \tag{16}$$

where η controls the augmentation magnitude and τ^q_t limits excessive displacement from the reliable feature support of the target predicate. The augmented representations are explicitly assigned the target predicate labels and combined with real samples to adapt the classifiers and the router. By independently modeling and transferring local variations in the two cue spaces, SCA enriches the feature coverage of underrepresented predicates while preserving cue heterogeneity and the original inference architecture.

4 Experiments

We evaluate on three widely used SGG benchmarks: VG150, OIV6, and GQA200 (Krishna et al., 2017; Kuznetsova et al., 2020; Hudson & Manning, 2019). Dataset statistics and evaluation protocols are provided in Appendices B.1, B.2, and C. We report R@K, mR@K, and widely used F@K (Kim et al., 2024a; Jeon et al., 2024) for balance on VG150/GQA200; OIV6 reports R@50, weighted relation/phrase mAP, and official score. Following common SGG practice(Li et al., 2021; Dong et al., 2022), we use pretrained Faster R-CNN detectors with ResNeXt-101-FPN backbones and keep them frozen during training. Further implementation details are provided in Appendix B.3.

4.1 Comparison with Representative Methods

Following the above protocols, we compare the proposed framework with representative SGG methods on all three benchmarks. Table 1 reports overall and class-balanced performance on VG150, Table 2 evaluates zero-shot triplet compositions, Table 3 reports results on OIV6, and Table 8 reports results on GQA200 in Appendix C.2. We examine whether the proposed framework retains competitive overall recall, improves class-balanced relation recognition, generalizes to unseen triplet compositions, and shows consistent effectiveness across benchmarks.

Two observations emerge from Table 1. First, the base dual-cue model maintains competitive conventional recall while preserving a balanced trade-off between R@K and mR@K. Second, SCA consistently improves mR@K and F@K across PredCls, SGCls, and SGDet, although with some reduction in R@K. This pattern indicates that SCA shifts the model toward a more class-balanced operating point rather than uniformly increasing conventional recall. For a more intuitive view of this trade-off, Appendix B.4 visualizes the results of Table 1 in the R-mR plane.

Table 2 further shows that the proposed framework achieves strong zero-shot recall, particularly under PredCls and SGCls. These results suggest that the learned relation representations generalize beyond subject-predicate-object compositions observed during training.

Results on OIV6 and GQA200 further support the effectiveness of the proposed framework across benchmarks. On OIV6, our method achieves the highest relation- and phrase-level weighted mAP and official weighted score while maintaining competitive recall. On GQA200, SCA again improves mR@K and F@K over the base model across a substantially larger predicate vocabulary. Together, these results support consistent effectiveness across different label spaces and evaluation protocols.

4.2 Ablation Study

We conduct controlled ablations under PredCls and SGCls to examine the contributions of the two cue branches, the foundation-model components, and predicate-adaptive routing. All variants follow the same relation-training protocol. Table 4 evaluates these components and their coordination.

Three observations emerge from Table 4. First, both branches are effective relation predictors but exhibit different performance profiles: the semantic-context branch favors class-balanced recognition, whereas the visual-spatial branch shows stronger zero-shot recall. Second, the lower zR@K of the semantic-context branch may be related to its explicit use of object-category information, which can increase dependence on category combinations observed during training. In contrast, the visual-spatial branch relies mainly on appearance and geometric evidence and is therefore less

Table 1: Performance comparison with representative SGG methods on the VG150. The **best result** is highlighted in **bold**, while the second- and third-best results are underlined.

	Method	PredCls			SGCls			SGDet		
		R@50/100	mR@50/100	F@50/100	R@50/100	mR@50/100	F@50/100	R@50/100	mR@50/100	F@50/100
Model-Agnostic	Motifs (Zellers et al., 2018)	65.5/67.2	16.5/17.8	26.4/28.1	39.0/39.7	8.7/9.3	14.2/15.1	**32.1**/**36.9**	5.7/6.6	9.7/11.2
	+TDE (Tang et al., 2020)	45.0/50.6	24.2/27.9	31.5/36.0	27.1/29.5	13.1/14.9	17.7/19.8	17.3/20.8	9.2/11.1	12.0/14.5
	+PCPL (Yan et al., 2020)	54.7/56.5	24.3/26.1	33.7/35.7	35.3/36.1	12.0/12.7	17.9/18.8	27.8/31.7	10.7/12.6	15.5/18.0
	+BPL-SA (Guo et al., 2021)	50.7/52.5	29.7/31.7	37.5/39.5	30.1/31.0	16.5/17.5	21.3/22.4	23.0/26.9	13.5/15.6	17.0/19.7
	+CogTree (Yu et al., 2021)	35.6/36.8	26.4/29.0	30.3/32.4	21.6/22.2	14.9/16.1	17.6/18.7	20.0/22.1	10.4/11.8	13.7/15.4
	+PPDL (Li et al., 2022d)	47.2/47.6	32.2/33.3	38.3/39.2	28.4/29.3	17.5/18.2	21.7/22.5	21.2/23.9	11.4/13.5	14.8/17.3
	+NICE (Li et al., 2022b)	55.1/57.2	29.9/32.3	38.8/41.3	33.1/34.0	16.6/17.9	22.1/23.5	27.8/31.8	12.2/14.4	17.0/19.8
	+CFA (Li et al., 2023)	54.1/56.6	35.7/38.2	43.0/45.6	34.9/36.1	17.0/18.4	22.9/24.4	24.7/31.8	13.2/15.5	17.2/20.8
	+TBE (Biswas & Ji, 2023)	51.5/55.1	24.7/30.7	33.4/39.4	32.2/33.8	14.5/17.4	20.0/23.0	23.9/27.1	9.4/11.7	13.5/16.3
	+DKBL (Chen et al., 2023)	57.2/58.8	29.7/32.2	39.1/41.6	32.7/33.4	18.2/19.4	23.4/24.5	27.0/30.7	12.5/15.1	17.1/20.2
	+I-Trans+ST-SGG (Kim et al., 2024a)	50.5/52.8	32.5/35.1	39.5/42.2	31.2/32.1	18.0/19.3	22.8/24.1	25.7/29.8	12.9/15.8	17.2/20.7
	+DPL (Jeon et al., 2024)	54.4/56.3	33.7/37.4	41.6/44.9	32.6/33.8	18.5/20.1	23.6/25.2	24.5/28.7	13.0/15.6	17.0/20.2
	+SDBG+BLRA (Gao et al., 2025)	51.0/52.9	31.9/33.8	39.2/41.2	31.2/32.3	17.6/18.3	22.5/23.4	23.8/27.8	13.8/16.4	17.5/20.6
	VCTree (Tang et al., 2019)	**66.2**/68.1	15.4/16.6	25.0/26.7	40.5/41.4	7.5/8.0	12.7/13.4	31.5/36.2	6.6/7.7	10.9/12.7
	+TDE (Tang et al., 2020)	44.8/49.2	26.2/29.6	33.1/37.0	28.8/32.0	15.2/17.5	19.9/22.6	17.3/20.9	9.5/11.4	12.3/14.8
	+PCPL (Yan et al., 2020)	56.9/58.7	22.8/24.5	32.6/34.6	40.6/41.7	15.2/16.1	22.1/23.2	26.6/30.3	10.8/12.6	15.4/17.8
	+BPL-SA (Guo et al., 2021)	50.0/51.8	30.6/32.6	38.0/40.0	34.0/35.0	20.1/21.2	25.3/26.4	21.7/25.5	13.5/15.7	16.6/19.4
	+CogTree (Yu et al., 2021)	44.0/45.4	27.6/29.7	33.9/35.9	30.9/31.7	18.8/19.9	23.4/24.5	18.2/20.4	10.4/12.1	13.2/15.2
	+PPDL (Li et al., 2022d)	47.6/48.0	33.3/33.8	39.2/39.7	32.1/33.0	21.8/22.4	26.0/26.7	20.1/22.9	11.3/13.3	14.5/16.8
	+NICE (Li et al., 2022b)	55.0/56.9	30.7/33.0	39.4/41.8	37.8/39.0	19.9/21.3	26.1/27.6	27.0/30.8	11.9/14.1	16.5/19.3
	+CFA (Li et al., 2023)	54.7/57.2	34.5/37.2	42.3/45.1	**42.4**/**43.5**	19.1/20.8	26.3/28.1	27.1/32.1	13.1/15.5	17.7/20.9
	+TBE (Biswas & Ji, 2023)	59.5/61.0	28.1/30.7	38.2/40.8	40.7/41.6	17.3/19.4	24.3/26.5	27.7/30.1	10.4/11.9	15.1/17.1
	+DKBL (Chen et al., 2023)	60.1/61.8	28.7/31.3	38.8/41.6	38.8/39.7	21.2/22.6	27.4/28.8	26.9/30.7	11.8/14.2	16.4/19.4
	+I-Trans+ST-SGG (Kim et al., 2024a)	52.5/54.3	32.7/35.6	40.3/43.0	36.3/37.3	21.0/22.4	26.6/28.0	20.7/24.9	12.6/15.1	15.7/18.8
	+DPL (Jeon et al., 2024)	54.0/55.8	35.3/37.9	42.7/45.1	32.6/33.9	19.2/21.3	24.2/26.2	27.0/31.4	13.2/15.9	17.7/21.1
	+SDBG+BLRA (Gao et al., 2025)	50.3/52.2	34.7/36.8	41.1/43.2	34.6/35.9	19.5/20.5	24.9/26.1	22.6/26.3	15.2/17.8	18.2/21.2
Specific	MSDN (Li et al., 2017)	64.6/66.6	15.9/17.5	25.5/27.7	38.4/39.8	9.3/9.7	15.0/15.6	31.9/36.6	6.1/7.2	10.2/12.0
	Graph R-CNN (Yang et al., 2018)	65.4/67.2	16.4/17.2	26.2/27.4	37.0/38.5	9.0/9.5	14.5/15.2	29.7/32.8	5.8/6.6	9.7/11.0
	KERN (Chen et al., 2019)	65.8/67.6	17.7/19.2	27.9/29.9	36.7/37.4	9.4/10.0	15.0/15.8	27.1/29.8	6.4/7.3	10.4/11.7
	GPS-Net (Lin et al., 2020)	65.2/67.1	15.2/16.6	24.7/26.6	37.8/39.2	8.5/9.1	13.9/14.8	31.3/35.9	6.7/8.6	11.0/13.9
	PE-Net (Zheng et al., 2023)	65.2/67.3	31.4/33.5	42.4/44.7	39.1/40.2	18.2/19.3	24.8/26.1	30.4/34.8	12.3/14.3	17.5/20.3
	VETO (Sudhakaran et al., 2023)	64.2/66.3	22.8/24.7	33.6/36.0	35.7/36.9	11.1/11.9	16.9/18.0	27.5/31.5	8.1/9.5	12.5/14.6
	SQUAT (Jung et al., 2023)	55.7/57.9	30.9/33.4	39.7/42.4	33.1/34.5	17.5/18.8	22.9/24.3	24.5/28.9	14.1/16.5	17.9/21.0
	HetSGG (Yoon et al., 2023)	57.1/59.4	32.3/34.5	41.3/43.6	37.6/38.5	15.8/17.7	22.3/24.3	32.0/34.5	11.5/13.5	16.9/19.4
	DRM (Li et al., 2024b)	43.2/44.4	**41.9**/**43.5**	42.5/43.9	23.3/23.9	19.9/20.7	21.5/22.2	18.6/21.7	**18.9**/**21.0**	**18.7**/**21.3**
	SSC-SGG (Yang et al., 2025)	59.7/62.0	31.5/34.0	41.2/43.9	37.4/38.5	17.8/18.9	24.1/25.4	28.6/33.0	12.3/14.4	17.2/20.1
	Ours	66.0/**68.5**	30.1/33.3	41.3/44.8	41.5/43.0	18.0/19.8	25.1/27.1	26.7/30.5	12.3/14.2	16.8/19.4
	+SCA	60.0/63.4	34.3/38.2	**43.7**/**47.6**	34.8/36.3	**23.1**/**25.0**	**27.8**/**29.6**	24.5/28.4	13.9/16.4	17.7/20.8

Table 2: Zero-shot recall comparison on VG.

Method	PredCls	SGCls	SGDet
	zR@50/100	zR@50/100	zR@50/100
Motifs (Zellers et al., 2018)	7.3/9.4	0.7/1.7	0.0/0.3
+TDE (Tang et al., 2020)	14.4/18.2	3.4/4.5	2.3/2.9
+ST-SGG (Kim et al., 2024a)	18.0/21.0	-/-	-/-
+CAModule ()	16.7/20.3	5.1/7.6	2.1/**3.7**
VCTree (Tang et al., 2019)	7.7/10.3	0.7/2.1	0.0/0.5
+TDE (Tang et al., 2020)	14.3/17.6	3.2/4.0	**2.6**/3.2
+ST-SGG (Kim et al., 2024a)	**19.8**/21.0	-/-	-/-
VTransE (Zhang et al., 2017)	11.3/14.7	2.5/3.3	0.8/1.5
+TDE (Tang et al., 2020)	13.3/17.6	2.9/3.8	2.0/2.7
PE-Net (Zheng et al., 2023)	17.1/20.9	5.4/6.5	2.3/3.6
FCSGG (Liu et al., 2021)	8.3/10.7	1.0/1.2	0.6/1.0
Ours	18.4/**21.9**	**6.6**/**7.7**	2.0/2.6

Table 3: Performance comparison on OIV6.

Method	R@50	wmAP		Score$_{wtd}$
		rel	*phr*	
Motifs (Zellers et al., 2018)	71.6	29.9	31.6	38.9
+ST-SGG (Kim et al., 2024a)	71.8	31.1	32.5	39.8
VCTree (Tang et al., 2019)	74.1	34.2	33.1	41.7
BGNN (Li et al., 2021)	75.0	33.5	34.2	42.1
FGPL-A (Lyu et al., 2023)	73.4	35.9	36.8	43.8
PE-Net (Zheng et al., 2023)	76.5	36.6	37.4	44.9
SQUAT (Jung et al., 2023)	75.8	34.9	35.9	43.5
SGTR (Li et al., 2022c)	59.9	37.0	38.7	42.3
RelTR (Cong et al., 2023)	71.7	34.2	37.5	43.0
HetSGG (Yoon et al., 2023)	**76.8**	34.6	35.5	43.3
Ours	76.4	**39.9**	**40.1**	**47.3**

directly tied to such combinations. Foundation-model components consistently improve compositional generalization: their combination raises the semantic-context branch zR@50/100 from 10.7/14.9 to 13.0/16.8, while under the same adaptive dual-cue architecture, zR@50/100 increases from 14.5/18.3 to 18.4/21.9. Although the full model does not surpass the visual-only branch in zR@K, it achieves stronger R@K, mR@K, and F@K overall. Third, while early fusion remains competitive, predicate-adaptive routing achieves higher mR@K and F@K than both early fusion and

Table 4: Ablation of visual-spatial and semantic-context cues, foundation-model components, and cue coordination strategies under PredCls and SGCls.

Visual cue	Semantic cue	Foundation Model		Cue Coordination	PredCls				SGCls		
		$\mathbf{f}^{\text{image}}_{\text{BLIP}}$	$\mathbf{f}^{\text{text}}_{\text{BLIP}}$		R@50/100	mR@50/100	F@50/100	zR@50/100	R@50/100	mR@50/100	F@50/100
✓	×	×	×	×	60.8/63.6	22.8/25.6	33.2/36.5	20.4/23.9	37.1/39.4	12.0/12.9	18.1/19.4
×	✓	×	×	×	59.5/63.9	28.4/31.8	38.4/42.4	10.7/14.9	36.7/38.8	12.7/13.8	18.9/20.4
×	✓	✓	×	×	61.4/64.8	27.9/30.9	38.4/41.8	12.1/16.0	36.9/39.1	12.7/13.5	18.9/20.1
×	✓	×	✓	×	60.1/64.3	30.1/33.7	40.1/44.2	11.2/15.0	36.4/38.6	18.3/19.5	24.4/25.9
×	✓	✓	✓	×	62.2/65.6	30.1/33.5	40.6/44.4	13.0/16.8	39.8/41.2	17.7/19.1	24.5/26.1
✓	✓	✓	✓	Early	64.0/66.5	27.6/30.0	38.6/41.3	18.2/22.0	38.6/40.9	16.2/19.3	22.8/26.2
✓	✓	✓	✓	Average	67.0/69.0	26.7/29.0	38.2/40.8	18.2/21.7	41.9/42.7	15.1/17.3	22.2/24.6
✓	✓	×	×	Adaptive	63.2/65.7	27.4/30.3	38.2/41.5	14.5/18.3	36.9/39.7	12.9/14.6	19.1/21.3
✓	✓	✓	✓	Adaptive	66.0/68.5	30.1/33.3	41.3/44.8	18.4/21.9	41.5/43.0	18.0/19.8	25.1/27.1

fixed averaging, providing a better balance between overall and class-balanced recognition. Given the page limit, further analyses are provided in the appendix. B.4 examines the R-mR trade-off and SCA fine-tuning dynamics. B.5 studies sensitivity to the donor and mode numbers. B.6 provides a controlled comparison with representative debiasing methods using the same baseline. B.7 reports results with common pretrained vision-language encoders.

4.3 Visualization Analysis

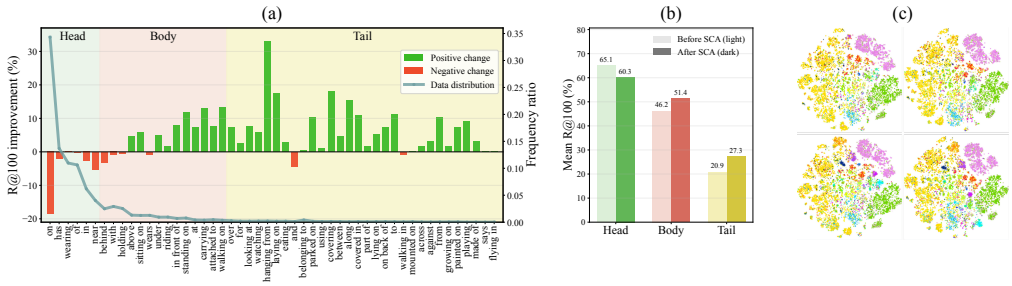


Figure 3: Effect of SCA under PredCls. (a) Change in per-predicate R@100. (b) Mean R@100 of each frequency group. (c) T-sne of real and generated features in the two cue spaces.

Figure 3(a) shows heterogeneous predicate-level responses to SCA: many fine-grained predicates obtain clear gains, while a smaller subset shows reduced R@100, indicating that the adaptation does not uniformly benefit all classes. Aggregating these changes by predicate frequency, Figure 3(b) shows higher mean R@100 for the body and tail groups and a decrease for the head group. This trend is consistent with the higher mR@K and F@K in Table 1, indicating more balanced recognition across predicate frequencies. Figure 3(c) further shows that the generated features remain close to real samples in both cue spaces while enriching nearby local structures without forming clearly isolated clusters.

Figure 4 further illustrates how the two cue branches contribute differently across predicate categories. The semantic-context branch shows stronger performance on tail predicates, whereas the fused model performs better on the head and body groups, indicating complementary prediction tendencies across the two branches. This variation also motivates adaptive coordination between them. Qualitative examples in Appendix B.8 further show that the model can recover informative fine-grained predicates rather than consistently favoring frequent generic relations.

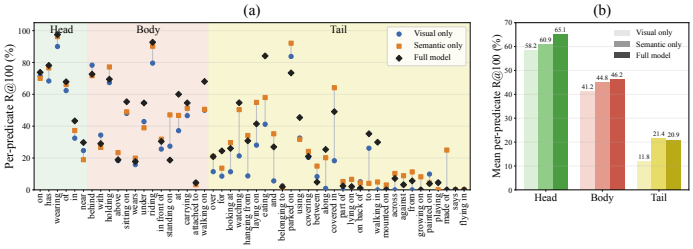


Figure 4: Predicate performance of single cue branch and the full model under PredCls. (a) Per-predicate R@100 grouped by head, body, and tail classes. (b) Mean R@100 for each group.

5 Conclusion

This work presents a dual-cue framework for scene graph generation that separately models visual-spatial and semantic-context evidence and coordinates them through predicate-adaptive routing. By incorporating pretrained vision-language representations into the semantic branch, the framework enriches relation semantics while retaining detector-based instance grounding. Synthetic Cue Augmentation further improves long-tailed predicate recognition by enriching the intra-class variation of underrepresented predicates in both cue spaces. Experiments on VG150, OIV6, and GQA200 demonstrate competitive overall and class-balanced performance, while the zero-shot results on VG150 further show generalization to unseen triplet compositions.

AI USE STATEMENT

We used GPT to assist with language polishing and grammatical correction. We also used GPT to generate fine-grained textual descriptions of predicate categories, which were manually reviewed and refined by the authors before being encoded into semantic prototypes. In addition, AI tools assisted with part of code development. The authors reviewed the AI-assisted work and verified the resulting code and experimental implementations. All research decisions, experimental results, and final manuscript content remain the responsibility of the authors.

REPRODUCIBILITY STATEMENT

All experiments are conducted on publicly available datasets, including Visual Genome, Open Images V6, and GQA200, following the standard dataset splits and evaluation protocols adopted in prior SGG studies. To facilitate inspection of the implementation and experimental procedure during the review process, we provide selected key implementation code together with representative training and evaluation logs in the supplementary material. These materials provide additional details on the implementation of the main components and the experimental settings. The complete source code will be made publicly available upon acceptance of our work.

REFERENCES

Bashirul Azam Biswas and Qiang Ji. Probabilistic debiasing of scene graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10429–10438, 2023.

Xiaojun Chang, Pengzhen Ren, Pengfei Xu, Zhihui Li, Xiaojiang Chen, and Alex Hauptmann. A comprehensive survey of scene graphs: Generation and application. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):1–26, 2023.

Tianshui Chen, Weihao Yu, Riquan Chen, and Liang Lin. 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.

Zhiqing Chen, Yawei Luo, Jian Shao, Yi Yang, Chunping Wang, Lei Chen, and Jun Xiao. Dark knowledge balance learning for unbiased scene graph generation. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 4838–4847, 2023.

Peng Chu, Xiao Bian, Shaopeng Liu, and Haibin Ling. Feature space augmentation for long-tailed data. In *Proceedings of the European Conference on Computer Vision*, pp. 694–710, 2020.

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

Xingning Dong, Tian Gan, Xuemeng Song, Jianlong Wu, Yuan Cheng, and Liqiang Nie. Stacked hybrid-attention and group collaborative learning for unbiased scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19427–19436, 2022.

Jiasong Feng, Lichun Wang, Hongbo Xu, Kai Xu, and Baocai Yin. Ensemble predicate decoding for unbiased scene graph generation. *Neurocomputing*, 648:130496, 2025.

Lianli Gao, Xinyu Lyu, Yuyu Guo, Yuxuan Hu, Yuan-Fang Li, Lu Xu, Heng Tao Shen, and Jingkuan Song. Informative scene graph generation via debiasing. *International Journal of Computer Vision*, 133(7):4196–4219, 2025.

Yuyu Guo, Lianli Gao, Xuanhan Wang, Yuxuan Hu, Xing Xu, Xu Lu, Heng Tao Shen, and Jingkuan Song. From general to specific: Informative scene graph generation via balance adjustment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16383–16392, 2021.

Pengxiao Han, Changkun Ye, Jieming Zhou, Jing Zhang, Jie Hong, and Xuesong Li. Latent-based diffusion model for long-tailed recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 2639–2648, 2024.

Drew A. Hudson and Christopher D. Manning. 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.

Jaehyeong Jeon, Kibum Kim, Kanghoon Yoon, and Chanyoung Park. Semantic diversity-aware prototype-based learning for unbiased scene graph generation. In *Proceedings of the European Conference on Computer Vision*, pp. 379–395, 2024.

Deunsol Jung, Sanghyun Kim, Won Hwa Kim, and Minsu Cho. Devil's on the edges: Selective quad attention for scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18664–18674, 2023.

Hyeongjin Kim and Byoung Chul Ko. Semantic scene graph generation based on an edge dual scene graph and message passing neural network. *Image and Vision Computing*, 160:105572, 2025.

Kibum Kim, Kanghoon Yoon, Yeonjun In, Jinyoung Moon, Donghyun Kim, and Chanyoung Park. Adaptive self-training framework for fine-grained scene graph generation. In *International Conference on Learning Representations*, 2024a.

Kibum Kim, Kanghoon Yoon, Jaehyeong Jeon, Yeonjun In, Jinyoung Moon, Donghyun Kim, and Chanyoung Park. LLM4SGG: Large language models for weakly supervised scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28306–28316, 2024b.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Fei-Fei Li. Visual Genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017.

Sanjoy Kundu and Sathyanarayanan N. Aakur. IS-GGT: Iterative scene graph generation with generative transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6292–6301, 2023.

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper R. R. Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. The Open Images dataset V4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision*, 128(7):1956–1981, 2020.

Hongsheng Li, Guangming Zhu, Liang Zhang, Youliang Jiang, Yixuan Dang, Haoran Hou, Peiyi Shen, Xia Zhao, Syed Afaq Ali Shah, and Mohammed Bennamoun. Scene graph generation: A comprehensive survey. *Neurocomputing*, 566:127052, 2024a.

Jiankai Li, Yunhong Wang, Xiefan Guo, Ruijie Yang, and Weixin Li. Leveraging predicate and triplet learning for scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28369–28379, 2024b.

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

Lin Li, Long Chen, Yifeng Huang, Zhimeng Zhang, Songyang Zhang, and Jun Xiao. The devil is in the labels: Noisy label correction for robust scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18869–18878, 2022b.

Lin Li, Guikun Chen, Jun Xiao, Yi Yang, Chunping Wang, and Long Chen. Compositional feature augmentation for unbiased scene graph generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 21685–21695, 2023.

Rongjie Li, Songyang Zhang, Bo Wan, and Xuming He. Bipartite graph network with adaptive message passing for unbiased scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11109–11119, 2021.

Rongjie Li, Songyang Zhang, and Xuming He. SGTR: End-to-end scene graph generation with transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19486–19496, 2022c.

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

Wei Li, Haiwei Zhang, Qijie Bai, Guoqing Zhao, Ning Jiang, and Xiaojie Yuan. PPDL: Predicate probability distribution based loss for unbiased scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19447–19456, 2022d.

Yikang Li, Wanli Ouyang, Bolei Zhou, Kun Wang, and Xiaogang Wang. Scene graph generation from objects, phrases and region captions. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1261–1270, 2017.

Yikang Li, Wanli Ouyang, Bolei Zhou, Jianping Shi, Chao Zhang, and Xiaogang Wang. Factorizable net: An efficient subgraph-based framework for scene graph generation. In *Proceedings of the European Conference on Computer Vision*, pp. 335–351, 2018.

Xin Lin, Changxing Ding, Jinquan Zeng, and Dacheng Tao. GPS-Net: Graph property sensing network for scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3746–3753, 2020.

Hengyue Liu, Ning Yan, Masood Mortazavi, and Bir Bhanu. Fully convolutional scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11546–11556, 2021.

Tao Liu, Rongjie Li, Chongyu Wang, and Xuming He. Relation-aware hierarchical prompt for open-vocabulary scene graph generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(5):5576–5584, 2025.

Julian Lorenz, Robin Schön, Katja Ludwig, and Rainer Lienhart. A review and efficient implementation of scene graph generation metrics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 2567–2575, 2024.

Cewu Lu, Ranjay Krishna, Michael Bernstein, and Fei-Fei Li. Visual relationship detection with language priors. In *Proceedings of the European Conference on Computer Vision*, pp. 852–869, 2016.

Xinyu Lyu, Lianli Gao, Pengpeng Zeng, Heng Tao Shen, and Jingkuan Song. Adaptive fine-grained predicates learning for scene graph generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(11):13921–13940, 2023.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pp. 1532–1543, 2014.

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

Jie Shao, Ke Zhu, Hanxiao Zhang, and Jianxin Wu. DiffuLT: Diffusion for long-tail recognition without external knowledge. In *Advances in Neural Information Processing Systems*, volume 37, pp. 123007–123031, 2024.

Gopika Sudhakaran, Devendra Singh Dhami, Kristian Kersting, and Stefan Roth. Vision relation transformer for unbiased scene graph generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 21882–21893, 2023.

Shuzhou Sun, Jing Liu, Li Liu, Shuaifeng Zhi, Huali Xu, Zhen Liu, Janne Heikkilä, and Yongxiang Liu. Foundation models based scene graph generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026. doi: 10.1109/TPAMI.2026.3712094.

Kaihua Tang, Hanwang Zhang, Baoyuan Wu, Wenhan Luo, and Wei Liu. 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.

Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. Unbiased scene graph generation from biased training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3716–3725, 2020.

Lei Wang, Zejian Yuan, and Badong Chen. Learning to generate an unbiased scene graph by using attribute-guided predicate features. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(2):2581–2589, 2023.

Zhaodi Wang, Yangyan Zeng, Biao Leng, and Xiaokang Zhou. CDC: Enhancing scene graph generation for IoST-driven social behavioral modeling with cooperative dual classifier. *IEEE Transactions on Computational Social Systems*, 13(1):1120–1133, 2026.

Danfei Xu, Yuke Zhu, Christopher B. Choy, and Fei-Fei Li. Scene graph generation by iterative message passing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5410–5419, 2017.

Kai Xu, Lichun Wang, Shuang Li, Tong Gao, and Baocai Yin. Scene adaptive context modeling and balanced relation prediction for scene graph generation. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 21(3):76:1–76:19, 2025.

Shaotian Yan, Chen Shen, Zhongming Jin, Jianqiang Huang, Rongxin Jiang, Yaowu Chen, and Xian-Sheng Hua. PCPL: Predicate-correlation perception learning for unbiased scene graph generation. In *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 265–273, 2020.

Jianwei Yang, Jiasen Lu, Stefan Lee, Dhruv Batra, and Devi Parikh. Graph R-CNN for scene graph generation. In *Proceedings of the European Conference on Computer Vision*, pp. 670–685, 2018.

Jiarui Yang, Chuan Wang, Liang Yang, Yuchen Jiang, and Angelina Cao. Adaptive feature learning for unbiased scene graph generation. *IEEE Transactions on Image Processing*, 33:2252–2265, 2024.

Jiarui Yang, Chuan Wang, Jun Zhang, Shuyi Wu, Jinjing Zhao, Zeming Liu, and Liang Yang. Semi-supervised clustering framework for fine-grained scene graph generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(9):9220–9228, 2025.

Kanghoon Yoon, Kibum Kim, Jinyoung Moon, and Chanyoung Park. Unbiased heterogeneous scene graph generation with relation-aware message passing neural network. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(3):3285–3294, 2023.

Kanghoon Yoon, Kibum Kim, Jaehyeong Jeon, Yeonjun In, Donghyun Kim, and Chanyoung Park. RA-SGG: Retrieval-augmented scene graph generation framework via multi-prototype learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(9):9562–9570, 2025.

Jing Yu, Yuan Chai, Yujing Wang, Yue Hu, and Qi Wu. CogTree: Cognition tree loss for unbiased scene graph generation. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, pp. 1274–1280, 2021.

Qifan Yu, Juncheng Li, Yu Wu, Siliang Tang, Wei Ji, and Yueting Zhuang. Visually-prompted language model for fine-grained scene graph generation in an open world. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 21560–21571, 2023.

Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. Neural motifs: Scene graph parsing with global context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5831–5840, 2018.

Hanwang Zhang, Zawlin Kyaw, Shih-Fu Chang, and Tat-Seng Chua. Visual translation embedding network for visual relation detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5532–5540, 2017.

Hao Zhang, Xingning Dong, Jinfei Gao, Liang Hao, Pei Shen, and Tian Gan. MBC-ATA: Maximum binary classification and anchor-based triplet augmentation for unbiased scene graph generation. In *Proceedings of the 6th ACM International Conference on Multimedia in Asia*, pp. 130:1–130:5, 2024.

Chaofan Zheng, Xinyu Lyu, Lianli Gao, Bo Dai, and Jingkuan Song. Prototype-based embedding network for scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22783–22792, 2023.

Chaofan Zheng, Lianli Gao, Xinyu Lyu, Pengpeng Zeng, Abdulmotaleb El Saddik, and Heng Tao Shen. Dual-branch hybrid learning network for unbiased scene graph generation. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(3):1743–1756, 2024.

Shanjin Zhong, Yang Cao, Qiaosen Chen, and Jie Gong. Learning with semantic ambiguity for unbiased scene graph generation. *PeerJ Computer Science*, 11:e2639, 2025.

A Additional Related Work

A.1 Representative Relation Reasoning Frameworks

Context-based two-stage SGG methods improve relation prediction by encoding object-level dependencies before predicate classification. IMP passes messages between object and relation nodes (Xu et al., 2017), MOTIFS uses recurrent context to capture frequent object layouts (Zellers et al., 2018), and VCTree learns dynamic tree structures for hierarchical visual context (Tang et al., 2019). Graph-based methods further make relation reasoning explicit: Factorizable Net shares computation over object-relation subgraphs (Li et al., 2018), Graph R-CNN uses object-category compatibility and union-region features to construct relation candidates (Yang et al., 2018), GPS-Net incorporates edge direction and node priority (Lin et al., 2020), and BGNN and HetSGG improve message passing through adaptive or relation-aware propagation (Li et al., 2021; Yoon et al., 2023). More recent methods reduce unnecessary relation interactions or strengthen edge-centric structure through selective quad attention, sparse graph generation, semantic edge modeling, or transformer-based global relation reasoning (Jung et al., 2023; Kundu & Aakur, 2023; Kim & Ko, 2025; Sudhakaran et al., 2023).

Multi-branch and multi-decoder designs have also been introduced to address the uneven difficulty of predicate recognition under long-tailed distributions. (Feng et al., 2025) employs three decoders with different frequency-oriented responsibilities: the main decoder is trained on all predicates, while two auxiliary decoders progressively focus on lower-frequency categories, with one emphasizing body and tail predicates and the other concentrating on tail predicates. (Zhang et al., 2024) reformulates multi-class predicate prediction through Maximum Binary Classification, decomposing the original classification problem into a set of predicate-specific binary classification tasks, such that each classifier determines whether a candidate relation belongs to a particular predicate category. (Zheng et al., 2024) employs a coarse-grained branch trained with standard cross-entropy and a fine-grained branch optimized through curriculum reweighting. It further introduces semantic-context refinement, knowledge distillation, and branch curriculum scheduling to coordinate head- and tail-predicate learning. (Xu et al., 2025) instead combines a regular prediction branch with a debiasing branch using rebalancing strategies, and progressively adjusts their contributions during training. Collectively, these designs alleviate long-tailed predicate learning by introducing differentiated learning behaviors or decision functions for predicates with varying frequencies and recognition difficulties.

These methods promote diverse predicate learning through frequency aware supervision, specialized classifiers, and different training and coordination strategies. Our dual-cue formulation is organized from a different perspective. The two branches are primarily distinguished by the form of relational evidence being modeled: the visual-spatial branch focuses on localized appearance and relative geometry, whereas the semantic-context branch models object semantics and contextual relation information. Frequency-aware grouping is used as an auxiliary learning and coordination prior rather than as the principal mechanism that defines the two branches. The resulting cue-specific predictions are subsequently coordinated at the decision level.

Vision-language and large language models have also been incorporated into SGG through several different paradigms. PGSG (Li et al., 2024c) reformulates open-vocabulary SGG as an image-to-sequence generation problem, directly using a generative vision-language model to produce scene-graph sequences from images and then recovering structured entities and relations through relation construction. LLM4SGG (Kim et al., 2024b) targets weakly supervised SGG and employs large language models to extract subject-predicate-object triplets from image captions and align the extracted entities and predicates with the target vocabulary, thereby constructing improved weak supervision for downstream relation learning. RAHP (Liu et al., 2025) focuses on open-vocabulary relation recognition through hierarchical prompting. It combines entity-aware and region-aware prompts, where the region-aware descriptions are generated by an LLM, and uses a vision-language model to dynamically select relevant textual prompts according to visual content.CaCao (Yu et al., 2023) addresses long-tailed predicate recognition through cross-modal predicate boosting. It learns a visually prompted language model to generate diverse fine-grained predicates under limited supervision and uses the generated predicate information to augment existing SGG models in a plug-and-play manner. FMSGG (Sun et al., 2026) more directly incorporates foundation-model representations into the relation prediction pipeline. It uses relationship prompts constructed from detector-derived

pair, spatial, and category information to isolate relation embeddings from CLIP image features in a single forward pass. It further combines these foundation-model representations with triplet-aware adjustment and evidential smoothing for long-tailed and zero-shot relation prediction.

Our framework uses foundation-model knowledge in a different role. Rather than formulating SGG as direct sequence generation, using LLMs primarily to construct weak supervision, or relying on prompt-based visual-textual matching for open-vocabulary relation recognition, we retain the conventional detector-based SGG pipeline and incorporate pretrained cross-modal information specifically into the semantic-context branch. Region-conditioned image representations provide contextual semantic evidence, while text-encoded predicate descriptions provide class-level semantic guidance. These representations are modeled separately from detector-derived visual-spatial evidence, and the resulting cue-specific predictions are coordinated at the decision level.

A.2 Representative Long-Tailed SGG Methods

Long-tailed SGG methods commonly address predicate imbalance through supervision rebalancing, prediction adjustment, representation learning, and feature-space augmentation. Supervision-level approaches modify the effective training signal to improve learning for underrepresented relations. NICE (Li et al., 2022b) corrects noisy predicate annotations, PCPL (Yan et al., 2020) exploits predicate correlations for adaptive reweighting, and PPDL (Li et al., 2022d) adjusts supervision according to predicate probability distributions. ST-SGG (Kim et al., 2024a) further introduces class-adaptive pseudo-labeling to exploit unlabeled relation pairs, while SSC-SGG (Yang et al., 2025) explores semi-supervised learning for low-frequency predicates.

Prediction-adjustment methods instead operate on biased decision distributions or transfer knowledge across predicate categories. TDE (Tang et al., 2020) introduces causal intervention to reduce contextual bias inherited from object co-occurrence, while Probabilistic Debiasing (Biswas & Ji, 2023) calibrates biased predicate predictions from a probabilistic perspective. CogTree (Yu et al., 2021) organizes predicate categories hierarchically for coarse-to-fine recognition, and DKBL (Chen et al., 2023) transfers knowledge across predicates to improve the recognition of underrepresented relations.

Representation-oriented approaches improve long-tailed relation recognition by imposing additional structure on the predicate feature space. PE-Net (Zheng et al., 2023) learns predicate prototypes in an embedding space to improve relation discrimination. DRM (Li et al., 2024b) jointly models predicate-level and triplet-level representations, allowing predicate semantics and subject-predicate-object composition information to be learned at different relational granularities. Other semantic representation methods further enhance the organization and discriminability of relation features (Jeon et al., 2024). These approaches aim to improve the separability and semantic structure of predicate representations, particularly for visually or semantically ambiguous relations.

Feature-space augmentation provides another direction by directly increasing the diversity of underrepresented predicate representations. Generic long-tailed recognition methods synthesize minority-class features through feature-space generation, latent generative modeling, or diffusion-based augmentation (Chu et al., 2020; Han et al., 2024; Shao et al., 2024). In SGG, A-PFG (Wang et al., 2023) generates additional predicate features under attribute guidance to construct a more balanced feature distribution, while CFA (Li et al., 2023) increases compositional diversity by decomposing relation representations into intrinsic and extrinsic components and recombining them to construct additional training samples. These feature-level strategies enrich the training representations available to low-frequency predicates.

B Experiments on Visual Genome

B.1 Dataset

We follow the widely used preprocessing protocol for Visual Genome (Krishna et al., 2017; Xu et al., 2017; Zellers et al., 2018; Li et al., 2021; Yoon et al., 2023). The dataset contains approximately 108K images and is divided into 70% for training and 30% for testing; 5K images are held out from the training portion for validation. Following the standard VG150 setting, we retain the 150 most frequent object categories and the 50 most frequent predicate categories. After preprocessing, an im-

age contains 11.6 objects and 6.2 annotated relations on average. This setting is highly long-tailed: generic predicates account for a large fraction of the annotations, whereas visually and semantically informative fine-grained predicates occur much less frequently.

B.2 Metrics and Evaluation Protocols

We report Recall@K (R@K), which measures the fraction of annotated ground-truth relations recovered among the top-K predictions for each image. Because R@K is dominated by frequent predicates, mean Recall@K (mR@K) first computes recall independently for every predicate category and then averages across categories (Lorenz et al., 2024). Thus, R@K emphasizes overall recognition, whereas mR@K gives equal weight to head and tail predicates. We report their harmonic mean,

$$F@K = \frac{2\, R@K\, \mathrm{mR}@K}{R@K + \mathrm{mR}@K}, \tag{17}$$

which rewards class-balanced improvements without overlooking overall recall. This measure is particularly relevant to SCA because it enriches underrepresented predicates rather than uniformly increasing all class scores.

Zero-shot Recall@K (zR@K) is evaluated on subject-predicate-object compositions that do not occur in the training set but appear at test time. The constituent object and predicate categories remain within the closed vocabulary; the novelty lies in their composition. We use zR@K to assess transfer beyond memorized triplet co-occurrences.

For evaluation protocols, we use the three conventional SGG tasks (Xu et al., 2017). Predicate Classification (PredCls) provides ground-truth object boxes and labels and evaluates predicate recognition for each candidate pair. Scene Graph Classification (SGCls) provides ground-truth boxes but requires prediction of both object and predicate labels. Scene Graph Detection (SGDet) provides only the image and requires detection of objects and their pairwise relations. For SGDet, the detector retains the top 80 entity proposals and applies per-class non-maximum suppression at IoU 0.5. A relation is correct only when its subject, object, and predicate categories are correct and both predicted boxes have IoU ≥ 0.5 with the corresponding ground-truth boxes.

B.3 Implementation Details

Following prior two-stage SGG work (Zellers et al., 2018; Tang et al., 2019; Li et al., 2021; Yoon et al., 2023; Jung et al., 2023), we use a pretrained Faster R-CNN with a ResNeXt-101-FPN backbone (Tang et al., 2020; Li et al., 2021; Dong et al., 2022). The detector remains frozen during relation training. Images are resized to a shorter side of 600 pixels with the longer side capped at 1,000 pixels. EgPN scores directed object pairs and retains at most 756 candidate relation pairs per image according to their scores, yielding a shared candidate topology for the two cue branches.

The relation feature dimension is 512, and both cue decoders contain two layers with eight attention heads. The semantic-context branch uses frozen BLIP with a ViT-B/16 vision encoder to extract subject, object, union, and global region features, while subject-object category context is initialized with GloVe embeddings (Pennington et al., 2014). Three descriptions for each predicate are generated using ChatGPT 5.5 and fixed throughout all experiments (see Table 5). They are encoded by the frozen BLIP text encoder, and their normalized embeddings are averaged to form the text prototype.

For adaptive routing, the reliability descriptor summarizes branch confidence, entropy-based sharpness, prediction margins, and cross-branch logit discrepancy, while the class prior combines predicate embeddings with the head/body/tail grouping of GCL (Dong et al., 2022). The router is a two-hidden-layer MLP with dimensions 128 and 64, followed by softmax normalization to obtain visual and semantic routing weights. Representative predicate descriptions are provided in Table 5; they are encoded once and introduce no text-encoder computation at inference.

For SCA, K-means is applied independently in the visual-spatial and semantic-context spaces to discover local variation modes. For each target predicate, the three most semantically related donor predicates are retained, i.e., $|\mathcal{D}_t| = 3$. Among the local modes associated with these donors, the three most compatible submodes are selected for variation transfer. We set $\lambda_{\mathrm{sem}} = \lambda_{\mathrm{cue}} = 0.5$ when evaluating mode compatibility. A residual sampled from the selected modes is transferred to

a real target feature with scale $\eta = 0.35$ and constrained by τ_t^q, defined as the maximum distance from real target samples to their class center in cue space q.

We train the base model with SGD using a learning rate of 1×10^{-3}, batch size 4, gradient accumulation over 4 iterations, and focal-loss parameter $\gamma = 2$. EgPN, both cue decoders, predicate classifiers, and the router are jointly optimized, while the detector and BLIP encoders remain frozen. During SCA adaptation, cue representations are frozen, while the cue-specific predicate classifiers and the predicate-adaptive router are jointly fine-tuned with real and synthetic features using a learning rate of 2×10^{-6}. SCA introduces no additional inference-time modules.

Table 5: Representative fine-grained predicate descriptions for five closely related VG predicates.

Predicate	Three complementary descriptions
on	(1) The subject is physically supported by the upper surface of the object.
	(2) The subject touches and rests on the object from above.
	(3) The object acts as the surface carrying or supporting the subject.
standing on	(1) The subject is upright with feet or base supported by the object.
	(2) The subject maintains a standing posture on top of the object's surface.
	(3) The object bears the subject's weight while the subject stands upright.
walking on	(1) The subject moves on foot across the surface of the object.
	(2) The subject steps over or along the object while supported by its surface.
	(3) The object is the ground or path on which the subject is walking.
sitting on	(1) The subject is seated with its weight supported by the object.
	(2) The subject rests in a sitting posture on the surface of the object.
	(3) The object functions as a seat or support for the subject's seated body.
above	(1) The subject is spatially higher than the object, without requiring physical contact.
	(2) The subject appears vertically over the object in the image layout.
	(3) The subject occupies a position that is higher than the object's position.

B.4 Recall-Mean Recall Trade-off on VG150

Figure 5 visualizes the VG150 results reported in Table 1 by plotting R@100 against mR@100 for the three SGG tasks. Across PredCls, SGCls, and SGDet, SCA consistently moves the base model toward higher mR@100, with some reduction in R@100, while crossing into higher F@100 contours in all three settings. This shift indicates a better balance between conventional and class-balanced recognition.

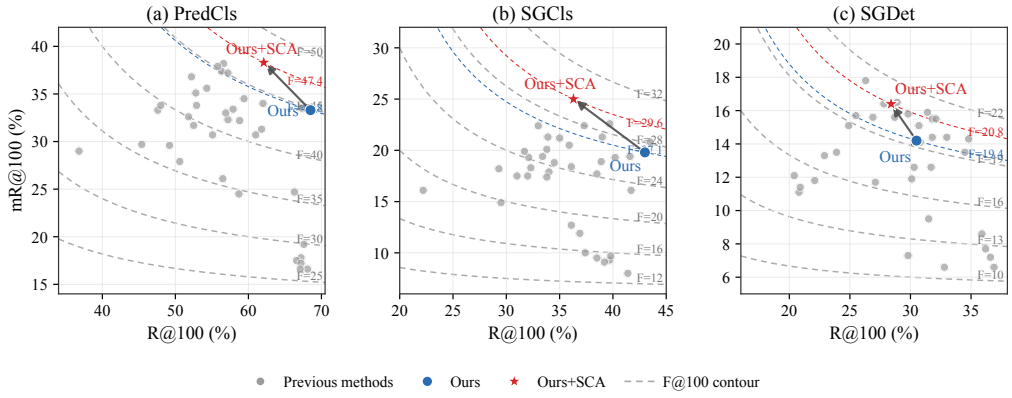


Figure 5: Trade-off between R@100 and mR@100 on VG150 across the three SGG tasks. Dashed curves denote equal-F@100 contours.

Figure 6 traces how the balance between conventional and class-balanced recognition evolves during SCA fine-tuning. As synthetic cue-space features progressively enrich underrepresented predicates, mR@100 rises steadily while R@100 decreases gradually. Their harmonic mean improves through the early and middle stages and reaches its maximum at 1,130 iterations, identifying the most favorable operating point for the two objectives. Further optimization continues to favor class-balanced recall but slightly reduces F@100, so we select the peak checkpoint for evaluation. The smooth

trajectory shows that SCA provides a controllable rebalancing process rather than an abrupt shift in the prediction distribution.

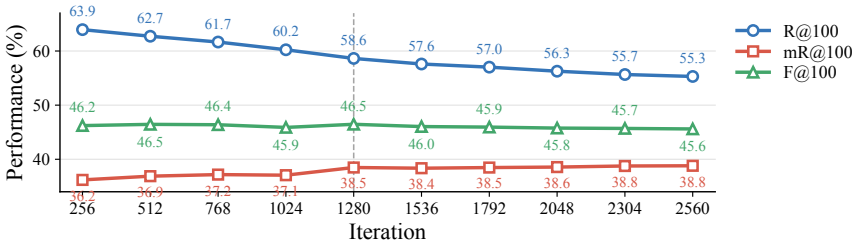


Figure 6: PredCls performance over SCA fine-tuning iterations on VG150.

B.5 Hyperparameter Sensitivity of SCA

We study the number of semantically related donor predicates D and the number of selected local submodes $\mathcal{M}$ on VG150 under the SGCls setting. Figure 7(a) shows that SCA performs favorably over a moderate range of donor and mode numbers, with the best F@100 obtained around $D = 2$-3 and $\mathcal{M} = 3$-7. In contrast, overly small or expanded donor-mode spaces lead to lower performance.

All SCA hyperparameters and checkpoints are selected on the validation set according to F@100. The selected configuration is then evaluated on the test set, and the resulting performance is reported in the main tables.

To further explain this difference, panels (b) and (c) compare the adaptation dynamics of the best and worst settings. The worst setting exhibits a slightly faster decline in R@100 and, more importantly, a substantially weaker increase in mR@100 during later adaptation. This indicates that an excessively large donor-mode space not only increases the loss in conventional recall but also limits the sustained benefit to class-balanced recognition. Overall, these results suggest that SCA benefits from a moderate candidate space that balances variation diversity and transfer reliability.

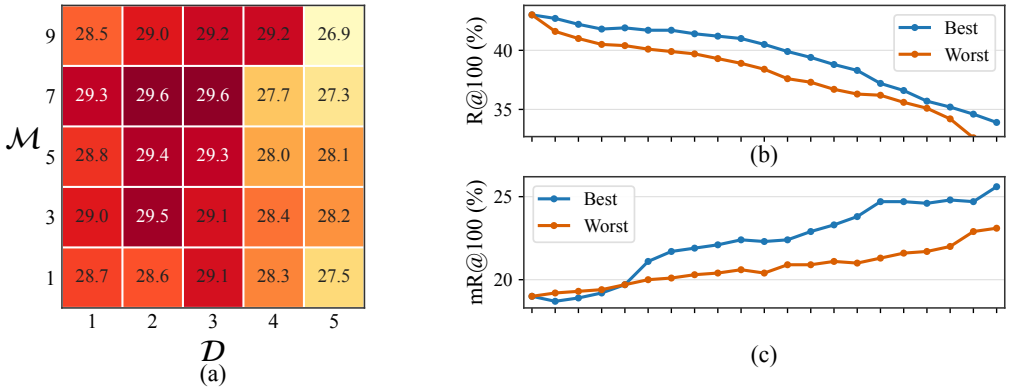


Figure 7: Sensitivity of SCA to the numbers of donor predicates D and local modes $\mathcal{M}$ on the VG150 validation set under SGCls. (a) Validation F@100 across donor-mode configurations. (b) Validation R@100 and (c) validation mR@100 adaptation curves for the best and worst settings.

B.6 Comparison with Debiasing Strategies

We further conduct a controlled comparison using the proposed dual-cue model as a common baseline, with only the long-tailed learning strategy changed. We compare simple resampling, CFA (Li et al., 2023), A-PFG (Wang et al., 2023), DKT (Li et al., 2024b), and SCA under the PredCls setting. R@K, mR@K, and F@K are reported to evaluate their trade-off between overall and class-balanced relation recognition.

Table 6: Controlled comparison of debiasing strategies on VG150 under PredCls. All variants use Ours as the common baseline. Metrics with multiple debiasing methods applied are not annotated.

Method	PredCls		
	R@50/100	mR@50/100	F@50/100
Ours	**66.0**/**68.5**	30.1/33.3	41.3/44.8
+TDE (Tang et al., 2020)	33.8/41.5	24.5/30.5	28.4/35.2
+Resample	56.3/58.4	31.7/35.9	40.6/44.8
+CFA (Li et al., 2023)	53.3/57.2	36.7/40.2	43.3/47.2
+A-PFG (Wang et al., 2023)	51.8/54.7	40.7/43.1	**45.6**/**48.2**
+A-PFG +SCA	52.8/55.0	42.0/44.6	46.8/49.3
+DKT (Li et al., 2024b)	43.5/46.3	**42.4**/**43.9**	42.9/45.1
+SCA	60.0/63.4	34.3/38.2	43.7/47.6

As shown in Table 6, the compared strategies exhibit different operating points between R@K and mR@K. Methods such as A-PFG and DKT achieve stronger class-balanced recall but incur larger reductions in conventional recall. In contrast, SCA substantially improves mR@K over the base model while retaining markedly higher R@K, resulting in competitive F@K. Moreover, SCA consistently outperforms simple resampling in R@K, mR@K, and F@K, indicating that transferring local variations provides benefits beyond merely increasing the number of training samples. Although A-PFG achieves slightly higher standalone F@K than SCA, combining SCA with A-PFG yields further improvements over A-PFG alone. This result suggests that SCA captures complementary local variation information that is not fully covered by generative feature augmentation. Figure 8 provides further insight into the performance gap between SCA and A-PFG in Table 6. Unlike methods that learn an additional generative model, SCA directly transfers local variations estimated to be compatible with underrepresented predicates. This direct strategy remains effective, as SCA consistently outperforms simple resampling, but imperfect compatibility may still introduce noise.

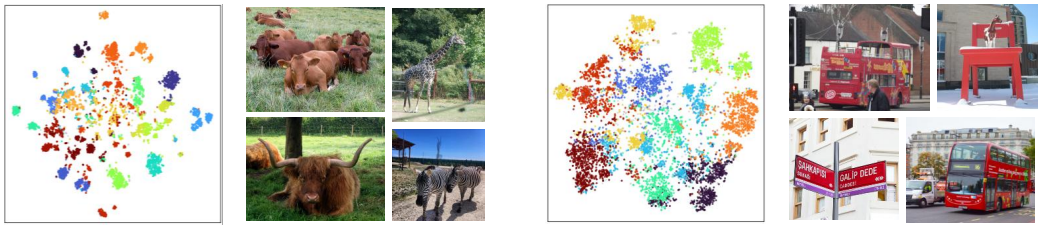


Figure 8: Failure cases of SCA variation transfer. (a) Semantic-context feature space and (b) retrieved variations for *standing on*. (c) Visual-spatial feature space and (d) retrieved variations for *parked on*.

Figures 8(a) and (c) illustrate the different structures of the two cue spaces. The semantic-context space is more compact and forms clearer local clusters, whereas the visual-spatial space is more dispersed, reflecting their different sources of relational evidence. These structural differences also lead to distinct failure modes. In Figure 8(b), one local variation selected for *standing on* contains semantically similar animal-ground compositions, but some instances correspond more closely to relations such as *lying on*, showing that semantic similarity does not always guarantee relation-level compatibility. In Figure 8(d), one variation selected for *parked on* contains samples with similar appearance or layout, but such visual similarity may also admit unrelated patterns, such as red signs or other objects. These cases show that direct variation transfer can occasionally introduce mismatched local modes, which may limit the gains of SCA despite its clear advantage over resampling.

B.7 Robustness to Foundation Models

Table 7 evaluates the robustness of the proposed framework to different pretrained vision-language encoder configurations. All three configurations are integrated into the same dual-cue architecture but exhibit different performance profiles. The CLIP-based configuration achieves higher conventional recall, whereas the BLIP-based configuration yields higher class-balanced and zero-shot recall. These results indicate that the choice of pretrained encoders affects the trade-off among overall,

class-balanced, and compositional recognition. We adopt the BLIP-based configuration in the main experiments because it achieves higher mR@K and zR@K while maintaining competitive R@K.

Table 7: Robustness to different pretrained vision-language encoder configurations on VG150 under the PredCls setting.

Pretraining Model	Image Encoder	Text Encoder	PredCls		
			R@50/100	mR@50/100	zR@50/100
None	None	None	63.2/65.7	27.4/30.3	14.5/18.3
CLIP	ViT-B/16	Transformer	67.1/69.3	28.7/31.6	17.8/21.6
BLIP	ViT-B/16	Transformer	66.0/68.5	30.1/33.3	18.4/21.9
BLIP-2	EVA-CLIP-g/14	Q-Former	66.5/69.0	24.5/27.9	17.0/21.2

B.8 Qualitative PredCls Results

Figure 9 presents representative PredCls predictions across scenes with different objects, interactions, and spatial layouts. Given ground-truth object boxes and labels, these examples isolate the model's relation reasoning from object detection. Across diverse interactions and spatial configurations, the model recovers informative, fine-grained predicates rather than collapsing them into frequent generic relations. These examples qualitatively support the improvements in mR@K and F@K reported in the quantitative experiments, showing that the model mitigates predicate bias while retaining accurate relation recognition.

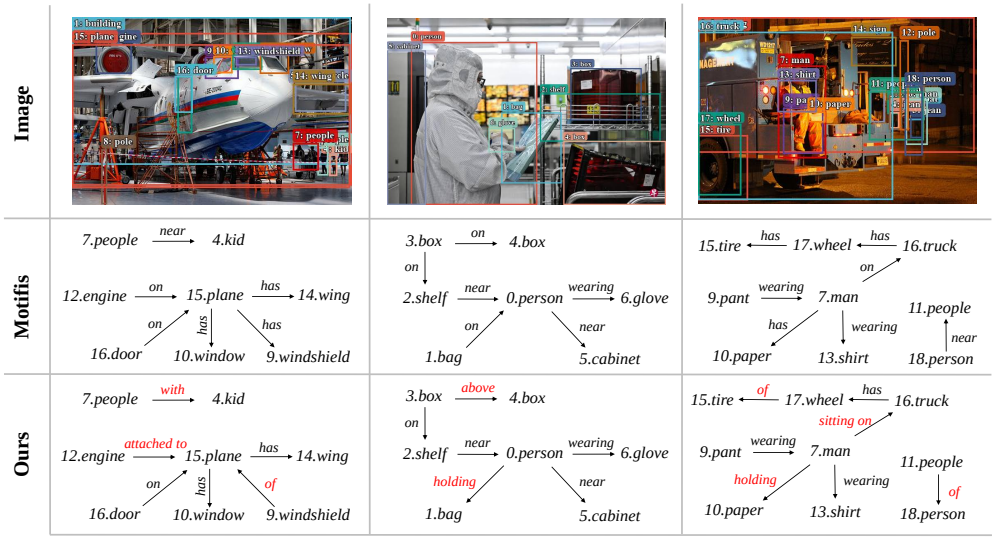


Figure 9: Qualitative PredCls results on Visual Genome. Given ground-truth object boxes and labels, each example compares the ground-truth and predicted scene graphs. Relation edges are labeled by predicate categories, illustrating the subject-predicate-object triplets recovered under the PredCls protocol.

C Experiments on OIV6 and GQA200

C.1 Open Images V6

We use the standard OIV6 scene-graph split following prior work (Kuznetsova et al., 2020; Li et al., 2021; Yoon et al., 2023; Jung et al., 2023). After preprocessing, the dataset contains 126,368 training images, 1,813 validation images, and 5,322 test images, with 301 object categories and 31 predicate categories. Its scale, label vocabulary, and evaluation criteria differ substantially from VG150, providing a complementary evaluation under a different data distribution.

Following the official protocol, we report R@50, relation-weighted mean average precision ($\mathrm{wmAP}_{\mathrm{rel}}$), phrase-weighted mean average precision ($\mathrm{wmAP}_{\mathrm{phr}}$), and the official weighted score

(Kuznetsova et al., 2020). $\mathrm{wmAP}_{\mathrm{rel}}$ evaluates subject, predicate, and object predictions while requiring both predicted entity boxes to overlap their ground-truth boxes by an IoU greater than 0.5; per-class AP is weighted according to predicate frequency. $\mathrm{wmAP}_{\mathrm{phr}}$ instead treats the subject-object pair as a single enclosing phrase region. The official aggregate score is

$$\mathrm{Score}_{\mathrm{wtd}} = 0.2\,\mathrm{R@50} + 0.4\,\mathrm{wmAP}_{\mathrm{rel}} + 0.4\,\mathrm{wmAP}_{\mathrm{phr}}. \tag{18}$$

OIV6 is evaluated under SGDet: the model generates entity and relation proposals and predicts the corresponding object and predicate classes. We retain the top 80 detector proposals and apply per-class non-maximum suppression at an IoU threshold of 0.5, following the official relationship-detection protocol.

Table 3 reports the comparison on OIV6. Our method achieves the highest relation- and phrase-level weighted mAP and the highest official weighted score while maintaining competitive recall. The gains across both precision-oriented measures indicate that the improvement is not tied to a single matching criterion. Together with the VG150 results, these results show that the proposed framework remains effective across different dataset scales, label spaces, and evaluation protocols.

C.2 GQA200

GQA was introduced for compositional visual reasoning and was constructed from Visual Genome images and scene graphs (Hudson & Manning, 2019). The original dataset contains approximately 113K images and more than 3.8 million relation annotations. Following the standard GQA200 preprocessing protocol (Dong et al., 2022), the benchmark is defined with the 200 most frequent object categories and 100 most frequent predicate categories. The processed split contains 57,623 training images, 5,000 validation images, and 8,209 test images. Compared with VG150, GQA200 follows the same evaluation setting while substantially expanding the object and predicate vocabularies.

GQA200 uses the same metrics and evaluation protocols as VG150. We report R@50/100, mR@50/100, and F@50/100 under PredCls, SGCls, and SGDet.

Table 8 reports the comparison on GQA200. The base model maintains strong conventional recall, while SCA further improves class-balanced recall and F@K across the three SGG tasks. These results extend the VG150 observations to a 100-predicate setting, supporting the effectiveness of SCA under a substantially larger predicate vocabulary.

Table 8: Performance comparison with representative SGG methods on GQA200. The **best result** is highlighted in **bold**, while the second- and third-best results are underlined.

Method	PredCls			SGCls			SGDet		
	R@50/100	mR@50/100	F@50/100	R@50/100	mR@50/100	F@50/100	R@50/100	mR@50/100	F@50/100
Motifs (Zellers et al., 2018)	65.3/66.8	16.4/17.1	26.2/27.2	**34.2/34.9**	8.2/8.6	13.2/13.8	28.9/33.1	6.4/7.7	10.5/12.5
+GCL (Dong et al., 2022)	44.5/46.2	36.7/38.1	40.2/41.8	23.2/24.0	17.3/18.1	19.8/20.6	18.5/21.8	16.8/18.8	**17.6**/20.2
+DPL (Jeon et al., 2024)	50.3/52.3	31.6/33.9	38.8/41.1	25.0/25.9	13.3/14.4	17.4/18.5	15.0/19.0	11.1/13.1	12.8/15.5
+BRP (Xu et al., 2025)	60.1/61.7	28.7/29.6	38.8/40.1	30.9/31.7	13.6/14.0	18.9/19.4	26.0/30.1	12.2/14.0	16.6/19.1
+MBRL (Zhong et al., 2025)	55.5/57.2	31.9/33.9	40.5/42.6	28.8/29.6	15.9/16.6	20.5/21.3	25.0/28.7	12.6/15.8	16.8/**20.4**
VCTree (Tang et al., 2019)	63.8/65.7	16.6/17.4	26.3/27.5	34.1/34.8	7.9/8.3	12.8/13.4	28.3/31.9	6.5/7.4	10.6/12.0
+GCL (Dong et al., 2022)	44.8/46.6	35.4/36.7	39.5/41.1	23.7/24.5	17.3/18.0	20.0/20.8	17.6/20.7	15.6/17.8	16.5/19.1
+DPL (Jeon et al., 2024)	46.1/50.6	31.4/34.3	37.4/40.9	22.1/23.7	13.3/14.9	16.6/18.3	14.4/18.1	11.9/13.7	13.0/15.6
+BRP (Xu et al., 2025)	58.4/60.1	29.2/30.2	38.9/40.2	30.9/31.6	13.9/14.2	19.2/19.6	23.5/27.0	11.9/13.3	15.8/17.8
SHA (Dong et al., 2022)	63.3/65.2	19.5/21.1	29.8/31.9	32.7/33.6	8.5/9.0	13.5/14.2	25.5/29.1	6.6/7.8	10.5/12.3
+GCL (Dong et al., 2022)	42.7/44.5	**41.0/42.7**	**41.8/43.6**	21.4/22.2	**20.6/21.3**	**21.0/21.7**	14.8/17.9	**17.8/20.1**	16.2/18.9
VETO (Sudhakaran et al., 2023)	64.5/66.0	21.2/22.1	31.9/33.1	30.4/31.5	8.6/9.1	13.4/14.1	26.1/29.0	7.0/8.1	11.0/12.7
PE-Net (Zheng et al., 2023)	54.3/56.0	26.2/27.1	35.4/36.5	26.2/27.0	11.2/11.5	15.7/16.1	19.5/22.9	10.3/11.9	13.5/15.7
+RA-SGG (Yoon et al., 2025)	48.3/50.1	35.4/36.8	40.9/42.4	19.9/20.8	16.4/17.2	18.0/18.8	16.3/19.0	12.9/15.0	14.4/16.8
SAC-Net (Xu et al., 2025)	**65.6/67.2**	16.8/18.1	26.8/28.5	31.9/32.6	8.2/8.7	13.1/13.7	**29.3/33.4**	6.7/7.7	10.9/12.5
+BRP (Xu et al., 2025)	59.9/61.4	28.3/29.4	38.4/39.8	30.7/31.4	13.6/13.9	18.8/19.3	25.1/29.4	13.3/15.3	17.4/20.1
CDC (Wang et al., 2026)	36.9/40.1	40.3/**42.7**	38.5/41.4	19.0/20.5	18.6/19.8	18.8/20.1	16.2/19.8	17.1/19.2	16.6/19.5
Ours	63.9/66.5	25.2/27.1	36.1/38.5	33.1/34.5	11.7/12.5	17.3/18.4	28.5/32.1	10.3/11.7	15.1/17.1
+SCA	52.6/55.9	31.0/35.7	39.0/**43.6**	30.8/31.9	14.8/15.9	20.0/21.2	25.2/29.3	13.5/15.5	**17.6**/20.3

C.3 COMPLEXITY ANALYSIS

We further analyze the computational cost of the proposed framework. Table 9 compares its model size with representative two-stage SGG methods using the same ResNeXt-101-FPN detector backbone (Zellers et al., 2018; Tang et al., 2019; Zheng et al., 2023; Chen et al., 2019). The detector and task-specific SGG architecture contain 401.8M parameters, while the frozen BLIP ViT-B/16 visual encoder contributes an additional 85.8M parameters, resulting in a total inference-time model size

of 487.6M parameters. The 401.8M task-specific architecture is dominated by two components: the Faster R-CNN detector, which accounts for approximately 166M parameters, and the relation head, which contains 235.9M parameters. Within the relation head, most parameters are concentrated in relation and union feature extraction, accounting for approximately 147.4M parameters, and in cue context modeling and decoding, accounting for approximately 87.6M parameters. The fine-grained predicate descriptions are encoded offline in advance, and their fixed text prototypes are directly reused during inference. SCA is used only during post-training adaptation and is removed after training, introducing no additional inference-time parameters or modules.

For a fair comparison of inference efficiency, all methods are evaluated on a single NVIDIA GeForce RTX 4090 GPU with batch size 1 and FP32 precision, without mixed-precision inference, model quantization, or TensorRT acceleration. Before measurement, we warm up each model on 30 images to exclude initialization and CUDA startup overhead. Inference time is then measured over 1,000 images from the VG test set, with CUDA synchronization performed before and after each timed forward pass.

The reported latency includes object detection, relation candidate construction, relation prediction, and post-processing, while excluding data loading, CPU-to-GPU transfer, and metric computation. FPS is computed from the average per-image inference latency.

Table 9: Parameter count and inference speed of representative two-stage SGG methods.

Method	Params (M)	FPS
Motifs (Zellers et al., 2018)	369.9	6.83
VCTree (Tang et al., 2019)	361.5	5.82
PE-Net (Zheng et al., 2023)	411.0	6.55
KERN (Chen et al., 2019)	405.2	6.05
Ours	487.6	14.47

D Limitation and Future Work

Although adaptive cue coordination improves the overall balance between the two relation branches, it does not consistently attain the stronger branch prediction for every predicate. A fixed predicate- or frequency-dependent routing rule could provide more explicit coordination, but such rules may not transfer reliably across datasets with different predicate vocabularies and frequency distributions. Developing more precise and distribution-adaptive cue coordination is therefore an interesting direction for future work. In addition, the current study focuses on closed-vocabulary SGG and evaluates generalization to unseen compositions of known categories. Extending the dual-cue framework to open-vocabulary relation prediction, where previously unseen predicate or object categories must also be recognized, represents another promising direction.