

Frequency-Aware Elastic Prototype Boundary Learning for Long-Tailed Scene Graph Generation

Binghao Wang

Jiangsu University of Science and Technology

Xueying Sun

sunxueying@just.edu.cn

Jiangsu University of Science and Technology

Hanzhu Dai

Jiangsu University of Science and Technology

Yuqi Shi

Jiangsu University of Science and Technology

Qiang Zhang

Jiangsu University of Science and Technology

Research Article

Keywords: scene graph generation, long-tailed relation prediction, prototype boundary learning, visual-semantic fusion, elastic boundary-aware distance calibration

Posted Date: May 7th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-9516264/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Frequency-Aware Elastic Prototype Boundary Learning for Long-Tailed Scene Graph Generation

Abstract

Scene graph generation (SGG) addresses the task of detecting objects in an image and predicting the relationships among them. Although prototype-based methods have recently achieved clear progress on long-tailed SGG, fine-grained low-frequency predicates remain difficult to recognize because their relation features often exhibit larger intra-class variation and more dispersed distributions, making them easily confused with semantically similar high-frequency coarse-grained predicates under a unified prototype-matching rule. To alleviate this issue, we propose a frequency-aware elastic prototype boundary learning framework, termed SGE-Net. Under fixed relation prototypes, the framework learns relation-category-specific boundary scales through explicit frequency compensation and frequency-adaptive virtual sampling, so that relation prediction can exploit not only prototype-center matching but also category-dependent decision-boundary information. During inference, we further introduce elastic boundary-aware distance calibration, enabling the boundary information learned during training to better distinguish relation categories that are easily confused under prototype matching. In addition, we combine visual and semantic features with dynamic gating to provide more reliable relation features for the above boundary learning. Experiments and analyses on Visual Genome and Open Images V6 demonstrate that the proposed method achieves consistent gains in both long-tailed relation prediction and overall evaluation metrics.

Keywords: scene graph generation; long-tailed relation prediction; prototype boundary learning; visual-semantic fusion; elastic boundary-aware distance calibration

1 Introduction

Visual scenes in the real world contain rich spatial and semantic interactions. For tasks such as embodied intelligence (Wu et al. 2024), vision-language understanding (Bordes et al. 2024), and autonomous driving (Sun et al. 2025), recognizing objects alone is usually insufficient; one must further reason about the relationships among them. Scene graph generation (SGG) (Johnson et al. 2015) provides an effective way to represent objects and relations in a structured form for complex scene understanding. It treats objects as nodes and relations as edges, converting an image into a structured semantic graph, and has shown substantial value in image captioning (Yuan et al.

2022; Zeng et al. 2022), visual question answering (Guo et al. 2021; Hildebrandt et al. 2020; Zeng et al. 2022), image retrieval (Guo et al. 2021; Sudhakaran et al. 2023; Touvron et al. 2023), and robotic vision and interaction (Liu et al. 2023, 2024; Singh et al. 2023).

Most early studies (Chen et al. 2019; Lin et al. 2020) directly predicted relations based on detected objects and their pairwise combinations. These methods established the basic paradigm of SGG, but limited relation modeling often leads to unstable predictions when visual ambiguity is pronounced. To alleviate this issue, later works emphasized contextual dependencies among objects and explicitly modeled interactions through message passing (Xu et al. 2017), sequential modeling (Zellers et al. 2018), and self-attention mechanisms (Cong et al. 2023; Li et al. 2024), thereby producing more discriminative relation representations. However, under long-tailed distributions, tail predicates remain difficult to recognize because they are not only sensitive to appearance variations but also easily confused with semantically similar high-frequency predicates. Building on this line, recent prototype-based methods (Zheng et al. 2023) further reformulate relation prediction as alignment and matching between relation features and semantic prototypes, offering a new perspective for mitigating appearance diversity.

Nevertheless, this paradigm (Zheng et al. 2023) still has difficulty distinguishing fine-grained low-frequency predicates from coarse-grained high-frequency predicates. Subsequent studies extended prototype-space modeling from multiple perspectives, including multi-prototype representations (Chen et al. 2024), distribution modeling (Li et al. 2025), and multimodal enhancement (Yang et al. 2026). Although these methods improve the expressiveness of the prototype space, they leave a key question underexplored: under severe long-tailed distributions, high-frequency coarse-grained predicates can dominate local matching decisions, whereas low-frequency fine-grained predicates often exhibit more dispersed visual features. Consequently, fine-grained relations with inadequately clustered visual features are easily confused with nearby high-frequency prototypes. To alleviate this issue, should the effective decision region around each prototype adaptively adjust its tolerance margin according to predicate frequency? As shown in Fig. 1(a), the predicates in Visual Genome exhibit a pronounced long-tailed distribution, and the frequency gaps among *on*, *walking on*, and *painted on* are especially representative. Furthermore, as illustrated in Fig. 1(b)(c), these fine-grained low-frequency relations exhibit larger feature dispersion because of limited training samples. Since they are strongly related to the high-frequency predicate *on*, their scattered relation features are more likely to fall within the dominant decision region of *on* in the prototype space and are therefore more easily predicted as the coarse high-frequency relation under existing prototype-matching frameworks. This suggests that the key challenge in long-tailed relation prediction lies not only in whether prototype centers are separable, but also in whether a frequency-aware tolerance margin can be established for low-frequency categories.

Based on these observations, we propose a frequency-aware elastic prototype boundary learning framework, named SGE-Net. Unlike existing methods that mainly focus on the separability of prototype centers, our framework further emphasizes

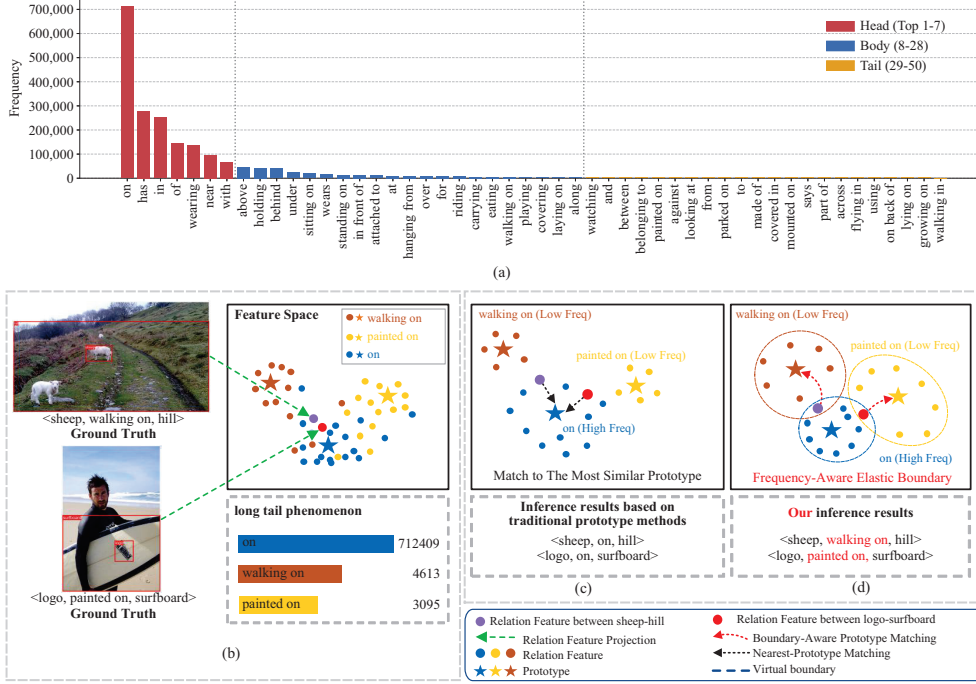


Fig. 1 Motivation of SGE-Net. (a) Predicates in Visual Genome exhibit a pronounced long-tailed distribution, with substantial frequency gaps between high-frequency coarse-grained relations and low-frequency fine-grained relations. (b) Taking *on*, *walking on*, and *painted on* as examples, fine-grained low-frequency relations differ semantically but their relation features tend to lie near the high-frequency coarse-grained relation *on* in feature space. (c) Under existing prototype-matching frameworks, these ambiguous relation features are therefore more likely to be predicted as the high-frequency coarse-grained relation *on*. (d) Based on fixed relation prototypes, SGE-Net learns frequency-aware elastic boundaries to provide larger tolerance margins for low-frequency categories, which empirically helps accommodate their dispersed feature space during matching.

relation-category decision boundaries around fixed relation prototypes, extending long-tailed relation prediction from prototype-center matching alone to joint modeling with boundary information. Specifically, under fixed relation prototypes, we learn relation-category-specific boundary scales through explicit frequency compensation and frequency-adaptive virtual sampling, allocating larger elastic boundaries to low-frequency predicates to better accommodate their higher feature variation. During inference, we further introduce elastic boundary-aware distance calibration, which acts as a category-level distance adjustment derived from the learned boundary scale. In addition, we combine visual and semantic features with dynamic gating to construct more reliable relation features, thereby forming a more stable feature space for prototype matching and subsequent boundary learning. As shown in Fig. 1(d), this frequency-aware tolerance mechanism provides an interpretable way to incorporate boundary information into prototype matching.

The primary contributions of this study are as follows:

- We develop a frequency-aware elastic prototype boundary learning framework, termed SGE-Net, for long-tailed scene graph generation. To alleviate the confusion between fine-grained low-frequency relations and semantically similar high-frequency relations, the framework learns relation-category-specific boundary scales under fixed relation prototypes through explicit frequency compensation and frequency-adaptive virtual sampling, while combining visual and semantic features with dynamic gating to provide more reliable relation features for boundary learning.
- We propose an elastic boundary-aware distance calibration strategy that exploits the boundary information learned during training at inference time to better distinguish relation categories that are easily confused under prototype matching, further improving fine-grained relation discrimination.
- We evaluate the proposed method on Visual Genome and Open Images V6. The experimental results indicate that it achieves consistent gains in both long-tailed relation recognition and overall evaluation metrics.

2 Related Work

2.1 Scene Graph Generation and Long-Tailed Relation Prediction

Scene graph generation is concerned with identifying entities in an image and reasoning about the relations among them. Earlier studies (Chen et al. 2019; Lin et al. 2020) directly predicted relations from detected object pairs and established the basic SGG paradigm. However, these methods model inter-entity dependencies only weakly, so their predictions are often unstable when relation reasoning depends on broader contextual cues or severe visual ambiguity. To address this limitation, later works placed more emphasis on contextual modeling. IMP (Xu et al. 2017) explicitly models information exchange between objects and relations through iterative message passing. Motifs (Zellers et al. 2018) captures stable statistical co-occurrence patterns among objects with sequential modeling, thereby enhancing global relation reasoning. VCTree (Tang et al. 2019) organizes object relations using dynamic tree structures to model hierarchical contextual dependencies more effectively. SGTR (Li et al. 2024) and RelTR (Cong et al. 2023) employ self-attention to model global relations in parallel, improving relation inference in complex scenes. Overall, these methods substantially strengthen relation representation and reasoning, but tail predicates remain difficult to recognize under long-tailed distributions and are still easily confused with high-frequency predicates.

To address long-tailed bias, existing studies mainly proceed from two directions. One line of work alleviates head-tail imbalance by adjusting the training distribution. For example, NICE (Li et al. 2022) cleans noisy annotations to improve the quality of long-tailed supervision; BGNN (Li et al. 2021) combines image-level and instance-level resampling to increase the learning opportunities of tail predicates; CogTree (Yu et al. 2021) reorganizes predicate relations with a cognitive hierarchy so that tail predicates receive more attention during training; and SQUAT (Jung et al. 2023) dynamically adjusts loss weights to weaken the dominance of high-frequency predicates. Another line operates at inference time to correct the model bias toward frequent predicates.

TDE (Tang et al. 2020) separates contextual bias from genuine relation cues through counterfactual causal inference; DLFE (Chiou et al. 2021) treats missing annotations as a positive-unlabeled learning problem to recover more stable prediction probabilities. These studies suggest that the difficulty of long-tailed SGG is not solely caused by sample imbalance, but is also closely related to visual ambiguity, semantic overlap, and confusion among fine-grained relations. In general, however, these methods mainly mitigate long-tailed bias through distribution adjustment or inference correction, while discussion of how relation-category decision boundaries in prototype space should vary with frequency remains limited. Motivated by this gap, we further study boundary learning around fixed relation prototypes.

2.2 Prototype-Based Relation Modeling

Recent studies have begun to model relation representations with semantic prototypes, reformulating relation prediction as alignment and matching between relation features and prototypes. PE-Net (Zheng et al. 2023) introduces semantic prototypes to reduce the interference caused by diverse object appearances and to improve the separability of relation representations, thereby opening a new direction for long-tailed relation prediction. Subsequent studies have extended prototype-space modeling from several perspectives. MPL (Chen et al. 2024) constructs multiple sub-prototypes for the same predicate to alleviate semantic overlap across different object combinations. DPL (Li et al. 2025) models prototype uncertainty through distributions, improving adaptability to complex semantic variation. SPLN (Zhang et al. 2024) introduces language priors to describe semantic space from multiple views, compensating for the limitations of a single prototype. FAPL (Xie et al. 2026) adjusts prototype learning according to frequency, mitigating the imbalance in learning intensity across predicates. Although these methods improve the expressiveness of prototype space, they still underexplore how the decision boundaries around prototypes should adapt to predicates with different frequencies. For fine-grained low-frequency predicates, relation features can still remain closer to coarse-grained high-frequency predicates even after being projected into prototype space, which leads to incorrect matching. Unlike these methods, which primarily strengthen prototype representations themselves, we further focus on how the decision boundary around fixed relation prototypes should be adjusted according to relation frequency, so that prototype learning concerns not only prototype centers but also the way their decision boundaries vary with predicate frequency.

2.3 Multimodal Relation Feature Learning

Reliable relation features are the basis of relation prediction. Accordingly, prior studies have introduced semantic priors and multimodal auxiliary information to improve relation modeling in complex scenes. MAAL-MLF (Yang et al. 2026) enhances relation representations with geometry-aware cues to address the weakness of pure appearance features for fine-grained interaction modeling. SDSGG (Chen et al. 2024) uses large language models to generate scene-related relation descriptions, providing additional semantic priors for relation prediction. 3DGraphLLM (Zemskova and Yudin 2025)

combines semantic graphs with large language models to enhance structured understanding in complex environments. These studies suggest that multimodal information can effectively enrich relation features. In contrast, SGE-Net integrates visual-semantic relation representation with frequency-aware prototype-boundary learning, so that multimodal cues provide a more stable feature basis for subsequent boundary estimation.

3 Method

3.1 Problem Definition

Scene graph generation (SGG) aims to transform a visual input into a graph-structured semantic representation. Given an image I , the problem can be viewed as two closely related subtasks: identifying the entity set $E = \{e_1, e_2, \dots, e_n\}$ in the image, where e_n denotes the n -th entity, and predicting the relation set $\mathcal{R}$ among these entities. Here, $\mathcal{R}$ consists of relation triplets $r_{i \rightarrow j}$, where $r_{i \rightarrow j}$ denotes the predicate defined over the directed subject-object pair (i, j) . The generation of a scene graph $\mathcal{G}$ can be decomposed as

$$P(\mathcal{G} | I) = P(\mathcal{B}, C | I) \cdot P(\mathcal{R} | \mathcal{B}, C, I) \quad (1)$$

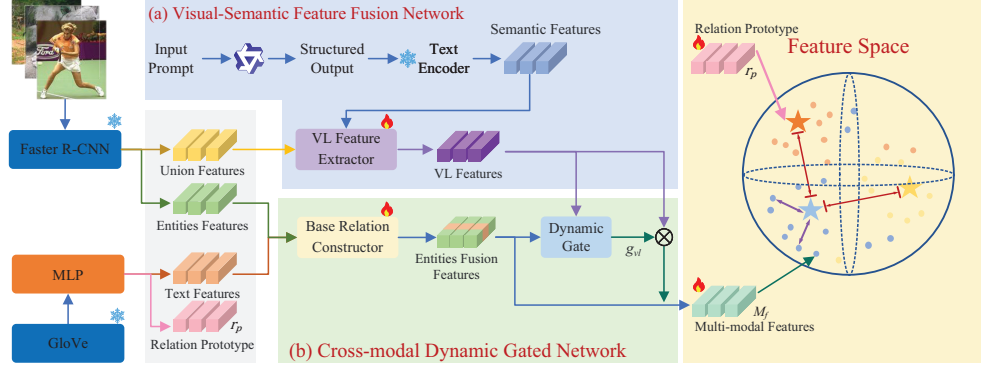
where $\mathcal{B}$ denotes the set of detected object boxes in the image, and $C = \{c_1, c_2, \dots, c_n\}$ denotes the corresponding entity categories. The term $P(\mathcal{B}, C | I)$ corresponds to object detection, while $P(\mathcal{R} | \mathcal{B}, C, I)$ corresponds to relation prediction. The object detection stage is handled by a pre-trained Faster R-CNN. This study mainly focuses on the second component, namely, learning relation representations, prototype structures, and decision boundaries that are more suitable for long-tailed relation prediction given the detected entities.

3.2 Overall Architecture

As shown in Fig. 2, SGE-Net is built upon relation prototypes and feature-space construction, and further focuses on the decision boundaries of different relation categories in feature space. Specifically, we first establish stable correspondence between relation features and relation prototypes, and then refine the decision boundaries of different relation categories on top of this basis. Guided by this objective, the overall framework is designed as a two-stage architecture.

The upper part of Fig. 2 corresponds to Stage 1. In this stage, subject-object union visual features and pre-constructed relation-predicate semantic features are jointly used as input, and visual-semantic relation features are obtained through cross-modal interaction, as illustrated in Fig. 2(a). These features are then combined with the basic relation feature derived from subject and object features under a dynamic gate, as shown in Fig. 2(b), yielding a more stable relation representation. On top of this representation, Stage 1 further forms a feature space suitable for prototype matching, so that relation features and relation prototypes establish clearer initial correspondence, thereby providing a reliable basis for boundary learning in Stage 2 under fixed relation prototypes.

Stage1: Cross-modal Context adaptive Weighted Fusion



Stage2: Multimodal Relation Feature Optimization and Elastic Prototype Boundary Learning

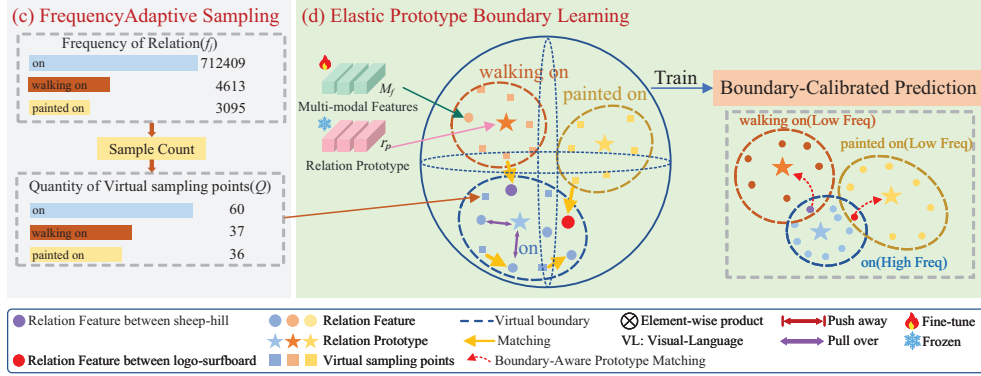


Fig. 2 Overall architecture of SGE-Net. The framework consists of two stages. In Stage 1, cross-modal interaction between subject-object union visual features and relation-predicate semantic features, together with subject/object features and dynamic gating, is used to construct relation representations and form a feature space suitable for prototype matching. In Stage 2, based on fixed relation prototypes, frequency-adaptive virtual sampling and explicit frequency compensation are used to learn relation-category-specific decision boundaries, and the learned boundary scale is further exploited during inference for elastic boundary-aware distance calibration.

The lower part of Fig. 2 corresponds to Stage 2. Within the feature space constructed in Stage 1, we fix the relation prototypes and learn relation-category-specific boundary scales through frequency-adaptive sampling, as shown in Fig. 2(c), and explicit frequency compensation, so that relation categories with different frequencies can obtain more suitable decision boundaries within the same prototype space. The corresponding elastic prototype boundary learning process is illustrated in Fig. 2(d). During inference, the learned boundary information is further used for elastic boundary-aware distance calibration, enabling the boundary knowledge formed during training to directly affect relation prediction at test time.

Table 1 Prompt template and structured output example used to construct relation-predicate semantic features. The left side shows the input prompt, and the right side shows the structured output for the target predicate generated offline by Qwen3-14B.

Input	Output
Input Template Instruction: Analyze the target predicate with the given object vocabulary. Provide three sections: Strong Visual Example , Strong Spatial Example , and Rare/Uncommon Example . Each section should contain one triplet (<i>Subject, Predicate, Object</i>) and three descriptions: [Visual Cues], [Spatial Logic], and [Physical Commonsense]. Constraints: Objects must be selected from the VG vocabulary, and the response must be written in English. In-context Example: Predicate: <i>riding</i> • (<i>person, riding, horse</i>) • (<i>person, riding, bike</i>) • (<i>dog, riding, surfboard</i>) Target Predicate: <i>hanging from</i>	Structured Output Strong Visual Example: (<i>person, hanging from, branch</i>) [Visual Cues] Arms extended to grasp the branch; body suspended below it. [Spatial Logic] The subject is below the object and aligned with the support point. [Physical Commonsense] Grip and friction support the hanging state. Strong Spatial Example: (<i>lamp, hanging from, ceiling</i>) [Visual Cues] The lamp is suspended by a cable or chain with the base facing downward. [Spatial Logic] The object is vertically attached to an overhead support. [Physical Commonsense] Suspension is maintained by tensile support from above. Rare/Uncommon Example: (<i>bag, hanging from, handle</i>) [Visual Cues] The bag hangs from the handle and may tilt slightly. [Spatial Logic] The handle serves as the support point and the bag remains below it. [Physical Commonsense] The handle bears the load and maintains balance.

3.3 Multimodal Relation Feature Construction

As shown in the upper part of Fig. 2, Stage 1 focuses on constructing multimodal relation features. Its goal is not to directly learn relation-category decision boundaries, but to obtain more stable and more discriminative relation representations through cross-modal feature interaction and adaptive fusion, while jointly aligning them with relation prototypes to support the formation and refinement of feature space during training. In this way, Stage 1 both constructs and optimizes multimodal relation features, and provides the basis for subsequent prototype-space construction and frequency-aware boundary learning under fixed relation prototypes in Stage 2. The construction of relation-predicate semantic features is illustrated in Table 1, and the internal feature interaction and relation-feature formation in Stage 1 are shown in Fig. 3(a)–(c).

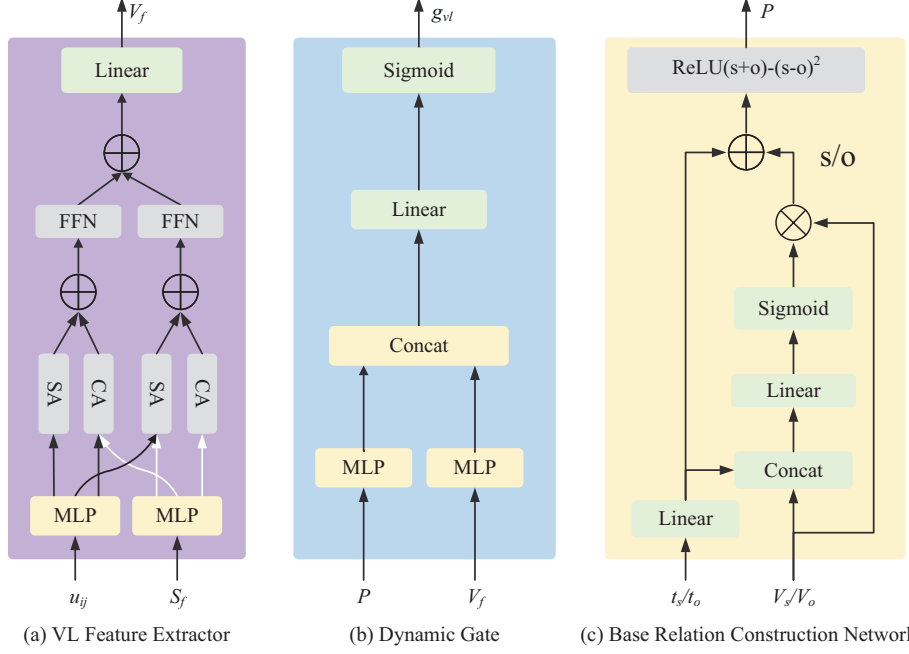


Fig. 3 Illustration of multimodal relation feature construction. (a) Cross-modal interaction between subject-object union visual features and relation-predicate semantic features. (b) Dynamic-gate estimation from the basic relation feature and the visual-semantic feature. (c) Construction of the basic relation feature from subject/object information and its adaptive fusion with the visual-semantic feature.

3.3.1 Visual-Semantic Feature Fusion

Table 1 shows how relation-predicate semantic features are constructed. The input side contains task instructions, constraints, in-context examples, and the target predicate, while the output side contains a structured description generated offline by Qwen3-14B around the target predicate. Based on these structured descriptions, we build an offline predicate-semantic library $S_f = [s_0, \dots, s_N] \in \mathbb{R}^{(N+1) \times d}$ using a frozen text encoder, where s_j denotes the semantic description vector of the j -th predicate category. This semantic library is generated once before model training and kept fixed thereafter. During both training and inference, a subject-object pair never directly uses the ground-truth predicate embedding. Instead, its union-region visual feature queries the entire predicate-semantic library to obtain an instance-adaptive semantic summary, which avoids label leakage while still injecting predicate-level semantic priors into the relation representation.

After obtaining the predicate-semantic library, we further fuse it with the subject-object union visual features, as shown in Fig. 3(a), to construct a visual-semantic relation feature. For a given subject-object pair, we use Faster R-CNN to extract the union-region visual feature and denote it by u_{ij} . We then map the visual feature and the predicate-semantic library into a unified representation space:

$$V'_{ij} = MLP_v(u_{ij}), \quad S'_f = MLP_s(S_f) \quad (2)$$

where u_{ij} denotes the subject-object union visual feature, S_f denotes the predicate-semantic library, and V'_{ij} and S'_f are the projected visual and semantic features, respectively. This step provides a common representational basis for subsequent instance-level semantic retrieval and cross-modal interaction.

The current subject-object pair then retrieves an instance-adaptive semantic summary from the entire predicate-semantic library:

$$a_{ij} = \text{Softmax} \left(\frac{(W_q V'_{ij}) (W_k S'_f)^T}{\sqrt{d}} \right), \quad \bar{S}_{ij} = a_{ij} (W_v S'_f) \quad (3)$$

where a_{ij} denotes the retrieval weights over all predicate categories, $\bar{S}_{ij}$ denotes the retrieved semantic summary for the current pair, and W_q , W_k , and W_v are learnable projection matrices. This retrieval step ensures that the semantic cue used by the current pair is inferred from the whole predicate inventory rather than injected from the target label.

With V'_{ij} and $\bar{S}_{ij}$, we perform bidirectional fusion by treating visual and retrieved semantic features as two complementary sources:

$$\begin{aligned} f_{vs} &= SA(V'_{ij}) \oplus CA(V'_{ij}, \bar{S}_{ij}), \\ f_{sv} &= SA(\bar{S}_{ij}) \oplus CA(\bar{S}_{ij}, V'_{ij}), \\ V_f &= \text{Linear}(FFN(f_{vs}) \oplus FFN(f_{sv})). \end{aligned} \quad (4)$$

where $SA(\cdot)$ and $CA(\cdot, \cdot)$ represent self-attention and cross-attention, respectively, $FFN(\cdot)$ is a feed-forward network, $\text{Linear}(\cdot)$ denotes a linear projection, and V_f is the resulting visual-semantic feature. This process preserves intra-modal structure while introducing bidirectional cross-modal interaction, allowing visual context and retrieved predicate semantics to complement each other within a unified representation. The resulting feature then serves as the basis for constructing a relation feature together with local subject and object information.

3.3.2 Dynamic-Gated Relation Construction

As shown in Fig. 3(b)(c), after obtaining the visual-semantic feature, we further incorporate subject-object information to construct a relation feature and use dynamic gating to adapt the contribution of semantic priors to each instance. Given the subject and object, we build their features from label word embeddings and ROI features. For the subject branch, we first generate a gate weight from the semantic and visual inputs:

$$g_s = \sigma(\text{Linear}([\text{Linear}(t_s), V_s])) \quad (5)$$

where t_s and V_s represent the subject word embedding and subject visual feature, respectively, $[\cdot, \cdot]$ indicates concatenation, and $\sigma(\cdot)$ is the Sigmoid activation function. g_s is the gate weight for the subject branch, controlling the contribution of visual cues to the current entity feature.

Accordingly, the semantic and gated visual information of the subject are fused as

$$s = \text{Linear}(t_s) \oplus (g_s \otimes V_s) \quad (6)$$

where $\otimes$ and $\oplus$ stand for element-wise multiplication and element-wise addition, respectively, and s is the fused subject feature. This formulation combines subject semantics with gated visual evidence, allowing the entity feature to retain both category semantics and visual cues.

The object feature o is obtained in the same way. Once both subject and object features are available, we follow PE-Net (Zheng et al. 2023) and use a nonlinear composition function to form the basic relation feature of the entity pair:

$$P = \text{ReLU}(s + o) - (s - o)^2 \quad (7)$$

where o is the object feature and P represents the basic relation feature jointly formed from the subject and object. This feature reflects the local relational context of the current entity pair. A similar nonlinear composition strategy was also adopted in earlier work (Zhang et al. 2018).

However, P is derived from the entity-pair composition, whereas V_f comes from visual-semantic interaction over the union region. Their sources and distributions are therefore not fully aligned. If these two types of features are directly used to estimate the gate coefficient, the estimation can be affected by feature-scale mismatch. We thus project both the basic relation feature and the visual-semantic feature into a unified space before computing the dynamic gate:

$$P^n = \text{MLP}(P), \quad V_f^n = \text{MLP}(V_f) \quad (8)$$

where P^n and V_f^n denote the projected basic relation feature and projected visual-semantic feature, respectively. This step allows the two features to be compared and combined in a more consistent representational space.

The two projected features are then concatenated to characterize the matching state between the current relational context and the visual-semantic information, from which the dynamic gate coefficient is generated:

$$g_{vl} = \sigma(\text{Linear}([P^n, V_f^n])) \quad (9)$$

where g_{vl} denotes the dynamic gate weight for the visual-semantic feature. This weight varies with the current entity-pair context and is used to regulate the actual contribution of the semantic prior to different instances.

Finally, the visual-semantic feature is adaptively fused into the basic relation feature according to the current relation context:

$$M_f = P \oplus (g_{vl} \otimes V_f^n) \quad (10)$$

where M_f denotes the final multimodal relation feature. Unlike direct addition or fixed-weight fusion, this formulation can adapt the contribution of the semantic prior to the current visual context, thereby reducing the interference of irrelevant semantics. The resulting relation feature serves as the input to subsequent prototype alignment and boundary learning.

3.4 Prototype Space Construction, Elastic Boundary Learning, and Elastic Boundary-Aware Distance Calibration

As shown in Fig. 2, multimodal fusion and dynamic gating first produce a relatively stable multimodal relation feature M_f . On this basis, we focus on the following question: how can different relation categories form decision boundaries in feature space that are compatible with their frequency distributions? Under long-tailed settings, relation categories differ substantially in sample frequency, semantic granularity, and feature distribution. If a unified prototype-matching criterion is still adopted, fine-grained low-frequency relations remain prone to confusion with semantically similar high-frequency relations. To alleviate this issue, we design three progressive components: prototype-space construction, elastic boundary learning, and elastic boundary-aware distance calibration. Specifically, Stage 1 jointly constrains relation features and relation prototypes to construct and refine feature space, gradually strengthening the correspondence between relation features and their target prototypes; Stage 2 learns frequency-aware elastic boundaries under fixed relation prototypes, so that different relation categories can obtain decision boundaries compatible with their distributions; and during inference, the learned boundary information is used for elastic boundary-aware distance calibration, allowing relation-category decision boundaries to contribute at test time and provide clearer evidence for fine-grained low-frequency relations. We explain these three components in the following.

3.4.1 Stage 1: Prototype Space Construction

After obtaining the multimodal relation feature M_f , the model must first establish a sufficiently discriminative prototype space. To this end, we construct and refine the feature space through joint constraints between relation features and relation prototypes r_p , where relation prototypes are obtained by projecting GloVe embeddings of relation words through an MLP. This process strengthens the correspondence between relation features and their target prototypes while keeping different prototypes sufficiently separated, thus improving the overall separability of prototype space. After this stage, relation prototypes can serve as stable semantic anchors in the unified feature space, providing reliable references for frequency-aware relation-category boundary learning under fixed relation prototypes in Stage 2.

Directional Alignment. We first align the direction of a relation feature with its target prototype using cosine similarity:

$$\mathcal{L}_{e_sim} = -\log \frac{\exp(\langle \bar{M}_f, \bar{r}_p^t \rangle / \tau)}{\sum_{j=0}^N \exp(\langle \bar{M}_f, \bar{r}_p^j \rangle / \tau)} \quad (11)$$

where $\bar{\cdot}$ denotes vector normalization, $\langle \cdot, \cdot \rangle$ denotes inner product, t is the ground-truth category index, and τ is the temperature parameter. This term enforces directional consistency between the relation feature and the target prototype. Directional similarity alone is insufficient for a stable prototype space, so we additionally impose intra-class compactness and inter-class separation in Euclidean space. The squared Euclidean distance from the relation feature to each prototype is defined as

$$g_j = \|M_f - r_p^j\|_2^2 \quad (12)$$

where g_j denotes the squared Euclidean distance from relation feature M_f to the j -th prototype r_p^j . It reflects the absolute proximity of the current sample to each prototype in prototype space.

We further distinguish the target prototype from its closest non-target prototypes so that intra-class aggregation and inter-class separation can be jointly described. Let the target-prototype distance be $g^+ = g_t$, and let g^- be the average of the k_1 smallest negative distances after sorting:

$$g^- = \frac{1}{k_1} \sum_{j=0}^{k_1-1} b'_j \quad (13)$$

where b'_j denotes the j -th element in the sorted set of negative distances, and k_1 is the number of nearest negative prototypes used in the computation. g^- characterizes the average distance between the current sample and its closest non-target prototypes.

The Euclidean feature-alignment loss is then written as

$$\mathcal{L}_{euc} = \max(0, g^+ - g^- + \gamma_1) \quad (14)$$

where $g^+ = g_t$ refers to the distance to the target prototype, and γ_1 is the margin parameter. This term encourages the relation feature to remain close to its target prototype and stay separated from nearby non-target prototypes in Euclidean space.

Euclidean and Cosine Separation. Besides constraining sample features against their target prototypes, we also expect the prototypes themselves to remain distinguishable. We thus introduce cosine and Euclidean separation constraints among prototypes. Let the prototype similarity matrix be

$$S = \bar{R}_p \cdot \bar{R}_p^T = (s_{ij}) \in \mathbb{R}^{(N+1) \times (N+1)} \quad (15)$$

where $\bar{R}_p$ denotes the normalized prototype matrix and s_{ij} is the cosine similarity between the i -th and j -th prototypes. This matrix describes the overall correlation among prototypes in semantic space.

If some prototypes remain overly close in semantic space, subsequent boundary learning will be built upon an entangled prototype structure. We therefore regularize only the off-diagonal similarities by

$$\mathcal{L}_{r.sim} = \frac{1}{N(N+1)} \sum_{i=0}^N \sum_{\substack{j=0 \\ j \neq i}}^N s_{ij}^2 \quad (16)$$

This term suppresses overlap among different relation prototypes in semantic space without penalizing the trivial self-similarity on the diagonal.

To further constrain prototype distribution from the distance perspective, we define the Euclidean distance between prototypes as

$$d_{ij} = \|r_p^i - r_p^j\|_2^2 \quad (17)$$

where d_{ij} denotes the squared Euclidean distance between prototype i and prototype j . This quantity further measures prototype separation in Euclidean space.

We are particularly concerned with the prototype neighborhoods that are easiest to confuse. We therefore compute the mean of the nearest k_2 distances for each prototype, denoted by d^- :

$$d^- = \frac{1}{(N+1)k_2} \sum_{i=0}^N \sum_{j=1}^{k_2} d'_{ij} \quad (18)$$

where d'_{ij} denotes the sorted prototype distances and k_2 denotes the number of nearest other-class prototypes considered for each prototype. d^- reflects the part of prototype distance relations that is most prone to confusion.

The corresponding Euclidean separation constraint is then given by

$$\mathcal{L}_{r_euc} = \max(0, -d^- + \gamma_2) \quad (19)$$

where γ_2 is the prototype-separation margin. This term promotes adequate separation among different relation prototypes in Euclidean space.

Combining the above constraints, the total Stage-1 loss is

$$\mathcal{L}_{stage1} = \mathcal{L}_{e_sim} + \mathcal{L}_{e_euc} + \mathcal{L}_{r_sim} + \mathcal{L}_{r_euc} \quad (20)$$

where the first two terms enforce directional alignment and distance constraints between relation features and target prototypes, while the latter two reduce semantic overlap among prototypes and enlarge Euclidean separation. Together, they align relation features with relation prototypes and form the prototype-matching space required for subsequent boundary learning.

After Stage 1 is completed, the learned relation prototypes are fixed and used as static anchors for Stage-2 boundary learning.

3.4.2 Stage 2: Frequency-Aware Elastic Boundary Learning

As shown in the lower part of Fig. 2, after Stage 1 constructs the prototype space, relation features and relation prototypes have established relatively stable correspondence, and the fixed relation prototypes provide reliable references for subsequent boundary learning. On this basis, Stage 2 no longer aims to continue adjusting prototype centers; instead, it learns decision boundaries around these fixed relation prototypes according to category frequency. Specifically, we generate virtual samples around fixed relation prototypes through implicit adaptive virtual sampling and adjust boundary ranges via explicit frequency compensation, so that relation categories with different frequencies can form more appropriate decision boundaries within a unified feature space.

As shown in Fig. 2(c), Stage 2 first allocates virtual sampling scales according to relation-category frequency. The ‘‘Sample Count’’ module denotes the procedure that determines the virtual sampling budget for each relation category around its fixed relation prototype, and its result for the j -th relation category is denoted by Q_j , i.e., the number of virtual samples assigned to category j . As shown in Fig. 1(a), head and tail relations often differ by orders of magnitude in occurrence frequency; for example, the

high-frequency relation *on* differs by more than two orders of magnitude from the low-frequency relations *walking on* and *painted on*. If a uniform sampling scale is adopted, the variation around high-frequency prototypes cannot be sufficiently covered; if sampling points are allocated directly according to raw counts, sampling becomes overly biased toward head categories. We therefore determine sampling scales according to logarithmic frequency, which preserves frequency differences while yielding smoother sampling allocation. We first compute the mean logarithmic frequency of all relation categories:

$$\mu_{\log} = \frac{1}{N+1} \sum_{j=0}^N \log f_j \quad (21)$$

where f_j denotes the occurrence frequency of the j -th relation category in the training set, and $\mu_{\log}$ denotes the mean logarithmic frequency over all relation categories. This quantity provides a unified normalization reference for sampling-scale allocation.

Based on this, the number of virtual samples assigned to relation category j is defined as

$$Q_j = \max \left(1, \text{Round} \left(\frac{\log f_j}{\mu_{\log}} \times Q_{base} \right) \right) \quad (22)$$

where Q_{base} is the base sample count and $\text{Round}(\cdot)$ denotes the rounding operation. This formulation gives high-frequency relations denser sampling coverage around fixed relation prototypes while retaining essential sampling budget for low-frequency relations, thereby providing a more balanced virtual-sample distribution for boundary learning.

Determining only the number of sampling points is not enough to characterize the decision boundaries of different relation categories. Low-frequency fine-grained relations often exhibit less compact visual representations (i.e., larger intra-class variation). Therefore, instead of assigning a uniformly narrow semantic boundary, we introduce a larger elastic boundary (σ_j) for some low-frequency predicates as a frequency-guided modeling prior. This enlarged scale is intended to better accommodate their scattered features, and we later examine this design through ablation and predicate-wise boundary-scale visualization. Considering that boundary adjustment in long-tailed relations should vary with category frequency, we further introduce an explicit frequency compensation term driven by category frequency to expand the boundary range of each tail relation category. This compensation should satisfy several intuitive requirements: it should be zero for the highest-frequency category, increase as frequency decreases, and vary smoothly overall. We therefore define the frequency compensation for relation category j as

$$\Delta_j = \lambda_t \log \left(\frac{f_{max}}{f_j} \right) \quad (23)$$

where f_{max} denotes the frequency of the most frequent relation category in the training set, and λ_t is an annealing coefficient whose upper limit is set to 0.05 in implementation. Here, $\frac{f_{max}}{f_j}$ describes the frequency difference between category j and the most frequent category, while the logarithm compresses the dynamic range under

the long-tailed distribution so that boundary compensation changes more smoothly with frequency.

After obtaining the frequency compensation term, we further use a Gaussian distribution to describe the neighborhood range around a fixed relation prototype. We do not assume that actual relation features strictly follow a Gaussian distribution. Rather, we introduce a continuous and sampling-friendly form to represent an adjustable boundary range around a fixed relation prototype. A Gaussian can provide a smooth description of a prototype region through both center and scale, and also allows the above frequency compensation to be naturally imposed on the boundary range. Similarly, Yang et al. (2021) used Gaussian distributions in scene graph generation to describe semantic uncertainty in relation features. Different from that work, we focus more on how frequency differences affect relation-category decision boundaries. Based on this idea, we represent the fixed relation prototype of relation category j as a Gaussian distribution $\mathcal{N}(r_p^j, \sigma_j^2)$, and combine the compensation term with the prototype’s own scale prediction to obtain the log-variance:

$$\log \sigma_j^2 = MLP_{var}(r_p^j) + \Delta_j, \quad \sigma_j = \exp\left(\frac{1}{2} \log \sigma_j^2\right) \quad (24)$$

where $MLP_{var}(\cdot)$ is the scale-prediction network, and σ_j denotes the boundary scale associated with the j -th prototype. This equation combines the scale prediction of the prototype itself with category-frequency compensation to determine the boundary range of relation category j around the fixed relation prototype while keeping the variance parameterization mathematically consistent. In practice, the explicitly injected Δ_j biases the model toward assigning a wider variance σ_j^2 to lower-frequency classes, which can be interpreted as a frequency-aware buffer zone for more dispersed feature distributions. In implementation, we clip the compensated log-scale to the range $[-3.5, 0.5]$ before exponentiation. This bounded parameterization prevents σ_j from becoming excessively small or overly large: when σ_j is too small, the reparameterized virtual samples collapse around the prototype center and fail to capture meaningful local variation; when σ_j is too large, the virtual samples become overly dispersed, leading to overly loose boundaries and weakened discrimination among nearby relation categories.

Once the boundary range is determined, we further generate virtual samples around the fixed relation prototype through reparameterization:

$$z_{j,k} = r_p^j + \sigma_j \odot \epsilon_{j,k}, \quad \epsilon_{j,k} \sim \mathcal{N}(0, I), \quad k = 1, \dots, Q_j \quad (25)$$

where $z_{j,k}$ is the k -th virtual sample of category j , $\odot$ represents element-wise multiplication, and $\epsilon_{j,k}$ is Gaussian noise. This reparameterized formulation keeps the sampling process differentiable, so gradients from the boundary objective can be back-propagated to the boundary-scale parameters through the virtual samples. These virtual samples do not replace the prototype center; instead, they serve as auxiliary references for the boundary range around fixed relation prototypes, enabling Stage 2 to learn relation-category decision boundaries that better reflect frequency differences within prototype neighborhoods.

During training, the model should preserve the basic ability to classify relation categories around prototype centers, while also ensuring that samples fall within the

boundary range corresponding to their target categories. We therefore first define a basic classification loss around prototype centers:

$$\mathcal{L}_{ce}^{(2)} = -\log \frac{\exp\left(-\alpha \|M_f - r_p^t\|_2\right)}{\sum_{j=0}^N \exp\left(-\alpha \|M_f - r_p^j\|_2\right)} \quad (26)$$

where α is a learnable parameter controlling distance scaling. This term maintains the basic category discrimination centered on prototypes.

On this basis, we further examine whether the current sample truly lies within the effective boundary range of its target relation category. To keep the boundary-learning objective differentiable, we approximate the nearest target-boundary distance with a temperature-controlled soft minimum over the target virtual sample set:

$$d_{\text{soft}} = -\tau_b \log \sum_{k=1}^{Q_t} \exp\left(-\|M_f - z_{t,k}\|_2 / \tau_b\right) \quad (27)$$

where $z_{t,k}$ denotes the k -th virtual sample in the target category- t sample set, Q_t is the number of samples assigned to that category, and τ_b is the soft-min temperature. d_{soft} is a differentiable approximation of the nearest distance from the current sample to the elastic boundary of the target category.

The corresponding boundary loss is then defined as

$$\mathcal{L}_{\text{bound}} = [\max(0, d_{\text{soft}} - R)]^2 \quad (28)$$

where R is the hyperparameter controlling the boundary radius. When a sample lies outside the target boundary, this term pulls it toward the target boundary; when the sample is already within the boundary, no additional penalty is imposed. Specifically, R serves as a tolerance margin that helps avoid over-compressing the feature space and accommodates part of the inherent intra-class variation. This objective therefore focuses on whether the sampled relation feature can enter the target boundary while preserving the prototype-centered decision structure. Because Eqs. (24)–(28) are connected through reparameterized virtual samples and the soft minimum, the boundary loss can update the boundary-scale parameters end-to-end and thereby realizes implicit adjustment of relation-category boundary scales during training.

By combining the basic classification term and the boundary constraint, the total Stage-2 loss is

$$\mathcal{L}_{\text{stage2}} = \lambda^{(2)} \mathcal{L}_{\text{bound}} + \mathcal{L}_{ce}^{(2)} \quad (29)$$

where $\lambda^{(2)}$ balances the boundary term and the basic classification term. This formulation indicates that Stage 2 preserves prototype-center-based discrimination while additionally exploiting boundary constraints to adjust the decision boundaries of different relation categories in prototype space.

3.4.3 Inference: Elastic Boundary-Aware Distance Calibration

In the second stage shown in Fig. 2, the scale information learned from boundary learning is further used for relation prediction at test time. Training determines the decision

boundaries that different relation categories should have in feature space; if testing still relies solely on the original prototype distance, the boundary information learned during training cannot effectively enter final decision making, and a mismatch remains between training objectives and inference criteria. To address this, we further design an elastic boundary-aware distance calibration strategy, which preserves the original geometric distance while incorporating relation-category decision-boundary information into prototype matching. The original distance between the relation feature and the j -th prototype is

$$d_j = \|M_f - r_p^j\|_2 \quad (30)$$

where d_j denotes the original Euclidean distance from relation feature M_f to prototype r_p^j . This distance reflects the basic matching result without boundary information.

Since relation matching is ultimately based on scalar distance comparison, we further aggregate the per-dimension scale information of the j -th prototype into a category-level boundary range:

$$\hat{\sigma}_j = (\text{Mean}(\sigma_j))^\rho \quad (31)$$

where σ_j corresponds to the per-dimension boundary scales of prototype j , $\text{Mean}(\cdot)$ denotes averaging over dimensions, ρ is the scale adjustment coefficient, and $\hat{\sigma}_j$ denotes the category-level boundary range associated with relation category j . This category-level scalarization provides a unified and efficient metric for subsequent distance calibration and keeps inference compact. The resulting boundary range is then used for calibrated prototype matching.

The key is how to combine this boundary range with the original distance. If only d_j is used, the boundary-range differences among relation categories still cannot enter final decision making; if only the normalized distance $\frac{d_j}{\hat{\sigma}_j}$ is used, the matching result becomes overly dependent on boundary range and weakens the discriminative role of original prototype distances. We therefore define the calibrated distance as

$$\tilde{d}_j = \lambda d_j + (1 - \lambda) \frac{d_j}{\hat{\sigma}_j} \quad (32)$$

where λ is set to 0.3 in implementation, and $\tilde{d}_j$ denotes the calibrated matching distance. Equation (32) is the key inference design: the first term λd_j preserves the proximity relationship between the feature and the prototype as measured by the original Euclidean distance, whereas the second term $(1 - \lambda) \frac{d_j}{\hat{\sigma}_j}$ calibrates the matching distance according to the relation-category boundary range learned during training. It is worth noting that the scaled term $\frac{d_j}{\hat{\sigma}_j}$ plays a role analogous to a category-level distance normalization. For low-frequency predicates, a larger $\hat{\sigma}_j$ yields a smaller distance penalty. This mathematical relaxation does not imply that fine-grained relations have broader semantic definitions; rather, it can be interpreted as a category-level confidence adjustment based on the learned boundary scale. When the Euclidean distances between a relation feature and multiple candidate prototypes are close, this calibration can provide supplementary decision evidence for dispersed tail-class features. The calibrated distance is then used for final relation-category prediction. Fig. 4 shows the

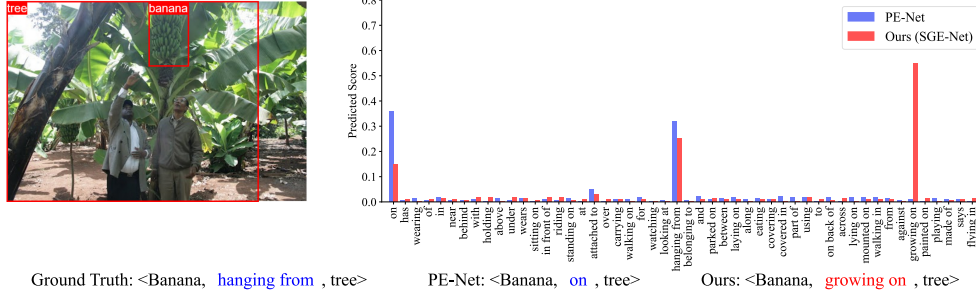


Fig. 4 Prediction distribution comparison between SGE-Net and PE-Net on the <Banana, tree> entity pair. Compared with the baseline, which is more inclined to output the high-frequency predicate *on*, SGE-Net is more likely to produce the fine-grained prediction *growing on* under the joint effect of boundary learning and elastic boundary-aware distance calibration.

prediction-distribution comparison on the <Banana, tree> entity pair, where the joint effect of boundary learning and elastic boundary-aware distance calibration makes the model more likely to predict the fine-grained relation consistent with the underlying semantics.

Overall, boundary learning and elastic boundary-aware distance calibration form a coherent decision process: the former gradually establishes more reasonable relation-category decision boundaries during training, while the latter continues to exploit relation-prototype boundary information to refine prototype matching during testing, thereby alleviating confusion between fine-grained relation categories and semantically similar high-frequency relations. In this way, the relation-category decision boundaries formed during training are translated into decision evidence at test time, providing a clearer basis for distinguishing fine-grained relations.

4 Experiments

4.1 Datasets

To assess the proposed method under both long-tailed relation prediction and more general relation-detection settings, we perform experiments on Visual Genome and Open Images V6. Visual Genome exhibits a representative long-tailed relation distribution and is therefore suitable for examining performance on infrequent fine-grained relations. Open Images V6, by contrast, adopts a stricter evaluation protocol for relation detection and is used to further assess overall performance in a more general setting.

Visual Genome. On Visual Genome, the dataset contains 108,077 images, with each image annotated by 38 object instances and 22 relation triplets on average. To maintain comparability with earlier studies, we use the standard VG150 split (Tang et al. 2020), which retains the 150 most frequent object categories and the 50 most frequent relation predicates. Following the conventional 70%/30% train-test partition, we further select 5,000 images from the training split to form the validation set.

Open Images V6. We also evaluate the proposed method on Open Images V6 (Kuznetsova et al. 2020). This benchmark includes 126,368 training images, 1,813

validation images, and 5,322 test images, covering 601 object categories and 31 relation predicates.

4.2 Evaluation Metrics

Visual Genome. To provide a thorough assessment of long-tailed SGG performance, we follow previous studies (Chen et al. 2024; Zheng et al. 2023) and present results on three widely used tasks: Predicate Classification (PredCls), Scene Graph Classification (SGCls), and Scene Graph Detection (SGDet). PredCls focuses solely on relation-predicate prediction, with ground-truth boxes and object labels provided. SGCls requires object classification together with relation prediction under ground-truth boxes. SGDet, in contrast, requires both object detection and relation prediction without relying on object-level priors.

Given the pronounced long-tailed distribution in Visual Genome, we mainly use mean Recall@K (mR@K) together with its aggregate metric Mean_{mR} for evaluation. Specifically, mR@K retains the top- K predicted relation triplets in each image, computes recall for every relation category, and then averages over all categories. It therefore better reflects recognition performance on infrequent relation categories. Mean_{mR} is computed as the arithmetic mean of mR@50 and mR@100, and serves as an overall indicator of long-tailed relation-prediction performance.

Open Images V6. Consistent with earlier work (Suhail et al. 2021; Sudhakaran et al. 2023), we report Recall@50 (R@50), relation weighted mean Average Precision (wmAP_{rel}), phrase weighted mean Average Precision (wmAP_{phr}), and the official composite metric (score_{wtd}). R@50 counts the number of ground-truth relations recovered within the top 50 predictions for each image. Both wmAP_{rel} and wmAP_{phr} are derived from average precision by first ranking all predictions according to confidence and then applying weighted averaging; see (Zhang et al. 2019) for details. The difference is that wmAP_{rel} requires the subject, object, and relation to all be correct, whereas wmAP_{phr} evaluates the subject-relation-object phrase as a whole and thus places greater emphasis on complete interaction semantics. The official composite metric is computed as $\text{score}_{wtd} = 0.2 \times \text{R@50} + 0.4 \times \text{wmAP}_{rel} + 0.4 \times \text{wmAP}_{phr}$. As it jointly accounts for recall, relation-detection precision, and phrase-detection precision, it provides a more comprehensive view of overall relation-detection quality on Open Images V6.

4.3 Implementation Details

In line with prior studies (Suhail et al. 2021; Sudhakaran et al. 2023), we use a pre-trained Faster R-CNN (ResNeXt-101-FPN) as the detector, with its parameters kept fixed throughout training. For fair comparison, we follow the standard detector, backbone architecture, dataset splits, and evaluation protocols adopted in prior SGG studies and the PE-Net setting. Entity semantic features are provided by GloVe (Pennington et al. 2014). Relation-predicate semantic features are generated offline by Qwen3-14B (Team 2025) and encoded by a pre-trained BERT (Devlin et al. 2019), after which they are used as an external offline, fixed predicate-semantic prior resource. Specifically, we use carefully designed prompts to generate 3 structured descriptions

Table 2 Performance comparison with state-of-the-art SGG methods on the Visual Genome dataset. * denotes reproduced results obtained using the codebase provided by the corresponding Motifs work (Zellers et al. 2018). The best result in each metric is marked in bold and the second best is underlined.

Model	Predicate Classification		Scene Graph Classification		Scene Graph Detection	
	mR@20/50/100	Mean _{mR}	mR@20/50/100	Mean _{mR}	mR@20/50/100	Mean _{mR}
Motif* Zellers et al. (2018)	11.7/15.2/16.2	15.7	6.7/8.3/8.8	8.6	5.0/6.8/7.9	7.4
VCtree Tang et al. (2019)	12.4/15.4/16.6	16.0	6.3/7.5/8.0	7.8	4.9/6.6/7.7	7.2
GPS-Net-SIL Wang et al. (2023)	<u>19.1</u> /23.8/25.2	24.5	11.7/13.2/14.0	13.6	6.9/9.1/10.4	9.8
RepSGG Liu and Bhanu (2024)	16.8/22.2/24.4	23.3	10.5/14.5/17.3	15.9	<u>7.2</u> /10.0/12.3	11.2
VCtree-HIERCOM Jiang et al. (2025)	17.6/26.3/31.8	29.1	<u>11.8</u> /16.9/20.0	<u>18.5</u>	5.8/8.7/11.1	9.9
Motif-NICEST Li et al. (2024)	-/29.5/31.6	30.6	-/15.7/16.5	16.1	-/10.4/12.4	11.4
Motif-SBG Li et al. (2024)	-/32.1/34.4	33.3	-/17.5/18.6	18.1	-/ 13.8 / <u>16.1</u>	15.0
Motif-Resam-ST-SGG Kim et al. (2024)	-/28.1/31.5	29.8	-/16.9/18.0	17.5	-/11.6/14.2	12.9
VCtree-MEET Sudhakaran et al. (2023)	-/25.5/ <u>34.5</u>	30.0	-/14.5/18.6	16.6	-/8.2/11.5	9.9
Motif-NICE Li et al. (2022)	-/29.9/32.3	31.1	-/16.6/17.9	17.3	-/12.2/14.4	13.3
VCtree-Inf Biswas and Ji (2023)	-/28.1/30.7	29.4	-/17.3/19.4	18.4	-/10.4/11.9	11.2
PE-Net Zheng et al. (2023)	-/31.5/33.8	32.7	-/ <u>17.8</u> /18.9	18.4	-/12.4/14.5	13.5
HetSGG Yoon et al. (2023)	-/31.6/33.5	32.6	-/17.2/18.7	18.0	-/12.2/14.4	13.3
Motif-IETrans Zhang et al. (2022)	-/30.9/33.6	32.3	-/16.8/17.9	17.4	-/12.4/14.9	13.7
IS-GGT Kundu and Aakur (2023)	-/26.4/31.9	29.2	-/15.8/18.9	17.4	-/9.1/11.3	10.2
SGE-Net(Ours)	26.1 / 36.5 / 40.1	38.3	13.2 / 17.9 / 20.3	19.1	9.7 / <u>13.4</u> / 16.2	<u>14.8</u>

per predicate, and extract 768-dimensional features using BERT. The structured descriptions are produced once before training, and the resulting semantic library remains fixed during both training and inference. Therefore, no online LLM interaction or additional external supervision is introduced at inference time. The semantic library will be made publicly available for reproducibility. Training is carried out in two successive stages. We first optimize Stage 1 to construct the prototype-aligned feature space. After Stage 1 converges, the relation prototypes are frozen as static anchors, and Stage 2 further updates the multimodal relation encoder and the boundary-scale predictor under $\mathcal{L}_{stage2}$. For reproducibility, the total 60k iterations are split evenly across the two stages. Following common practice in prior SGG studies, the main comparison results are reported from a single run with a fixed random seed. To further examine the stability of SGE-Net, we additionally repeat the PredCls experiment on Visual Genome under four different random seeds and report the corresponding statistics in the stability analysis. Unless otherwise specified, the remaining training settings follow previous studies.

For the key hyperparameters, we set the base sample count Q_{base} to 40, the distance threshold R to 0.3, the loss balancing coefficient $\lambda^{(2)}$ to 0.1, the upper bound of the annealed explicit frequency compensation coefficient to 0.05, the inference-stage distance calibration weight λ to 0.3, and the scale adjustment coefficient ρ to 0.55. The model is trained for 60k iterations with SGD, using an initial learning rate of 10^{-3} and a batch size of 8. All experiments are implemented using PyTorch and executed on one NVIDIA GeForce RTX 4090 GPU.

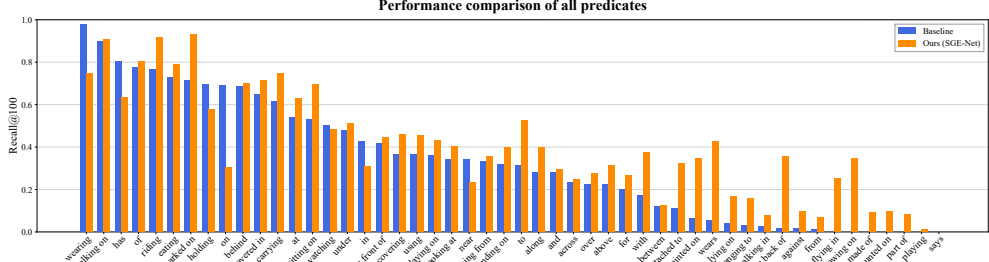


Fig. 5 Recall@100 comparison across all predicate categories under PredCls on Visual Genome between the baseline and SGE-Net (Ours).

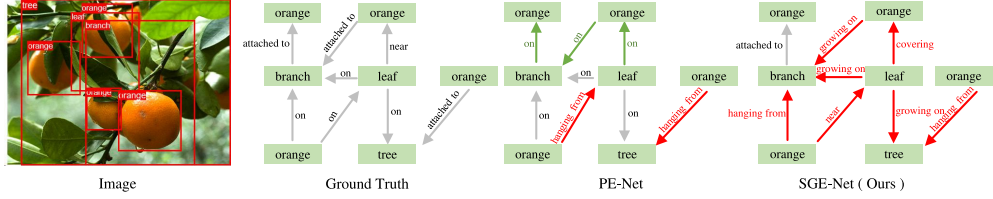


Fig. 6 Visual comparison between our two-stage method and PE-Net on PredCls. Predicates in red denote granularity-mismatch cases in which the prediction is more specific than the ground truth, predicates in black match the ground truth, and predicates in green differ from the ground truth.

4.4 Comparison with State-of-the-Art Methods

Visual Genome. As shown in Table 2, we benchmark SGE-Net against several representative recent SGG methods on Visual Genome. Owing to the pronounced long-tailed distribution of this dataset, we mainly focus on Mean_{mR} . SGE-Net achieves the best Mean_{mR} on Predicate Classification and Scene Graph Classification, with 38.3% and 19.1%, respectively; on Scene Graph Detection, it reaches 14.8%, the second-best result. In particular, on Predicate Classification, SGE-Net improves $mR@100$ to 40.1%, surpassing PE-Net (Zheng et al. 2023) by 6.3%. These results indicate that the proposed method improves tail-relation recognition while also achieving stronger overall performance on long-tailed relation prediction.

To more intuitively analyze predicate-level performance, Fig. 5 shows Recall@100 comparisons across all predicate categories under PredCls. SGE-Net achieves higher Recall@100 than the baseline on many predicates, indicating more consistent predicate-level improvements rather than gains limited to a specific tail subset. The improvements are especially notable for fine-grained and relatively low-frequency predicates such as *painted on*, *growing on*, and *part of*, which are semantically finer-grained and more easily confused with frequent coarse-grained relations.

Fig. 6 presents qualitative comparisons on Visual Genome. For the entities *orange*, *branch*, and *leaf*, PE-Net predicts more general relations such as *on* and *attached to*, whereas SGE-Net further produces more specific predicates such as *growing on*, *hanging from*, and *covering*. This example shows that, in scenes with relatively clear relation semantics, SGE-Net produces more fine-grained relation descriptions than

Table 3 Performance comparison with state-of-the-art methods on Open Images V6. * denotes methods originally proposed for HOI detection. The best result in each metric is marked in bold and the second best is underlined.

Model	R@50	$wmAP_{rel}$	$wmAP_{phr}$	$score_{wtd}$
PE-Net Zheng et al. (2023)	76.5	<u>36.6</u>	37.4	<u>44.9</u>
SQUAT Jung et al. (2023)	75.8	34.9	35.9	43.5
CooK Kim et al. (2024)	75.5	34.6	36.4	43.5
FGPL-A Lyu et al. (2023)	73.4	35.9	36.8	43.8
Motif-SIL Wang et al. (2023)	72.3	30.4	31.8	39.4
VCtree+TDE Tang et al. (2020)	69.3	30.7	32.8	39.3
HOTR* Zou et al. (2021)	52.7	19.4	21.5	26.9
AS-Net* Chen et al. (2021)	55.3	25.9	27.5	34.4
RelTR Cong et al. (2023)	64.5	34.2	<u>37.5</u>	41.6
SGE-Net(Ours)	73.4	38.3	38.6	45.4

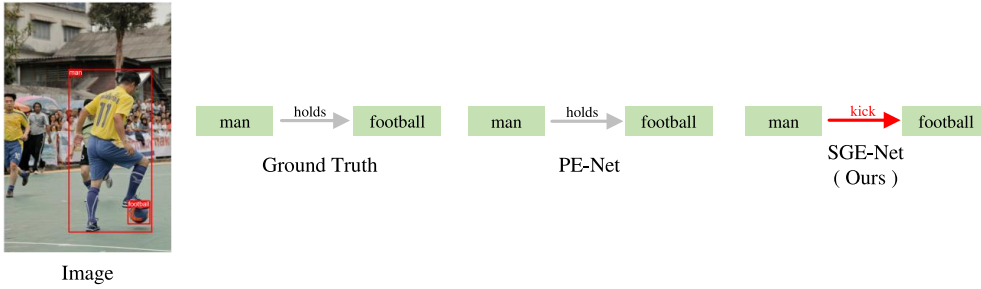


Fig. 7 Visual comparison between our two-stage method and PE-Net on Open Images V6. Predicates in red denote granularity-mismatch cases in which the prediction is more specific than the ground truth, and predicates in black match the ground truth.

PE-Net. Although the prediction does not exactly match the benchmark annotation, it remains semantically plausible and reflects a confusion between related predicates rather than a random error, which is consistent with the predicate-level improvements shown in Fig. 5.

Open Images V6. We further compare SGE-Net with existing methods on Open Images V6, as shown in Table 3. SGE-Net achieves 38.3% on $wmAP_{rel}$, 38.6% on $wmAP_{phr}$, and 45.4% on the official composite metric $score_{wtd}$, which are the best results on all three metrics. Compared with the second-best methods, the improvements are 1.7, 1.1, and 0.5 percentage points, respectively. Although SGE-Net is not the best on R@50, it achieves clear improvements on $wmAP_{rel}$ and $wmAP_{phr}$. This reflects a key characteristic of our method: boundary calibration improves ranking precision more clearly than broad recall. By refining the decision boundaries, the model becomes more confident in its top predictions for fine-grained relations, thereby boosting AP metrics, even if the total number of retrieved relations remains comparable. The results demonstrate that our method achieves better overall relation-detection quality while maintaining competitive recall.

Table 4 Ablation study on the main components of SGE-Net. Baseline denotes PE-Net. VL + Dynamic Gate denotes the use of relation-predicate semantic features with visual-semantic fusion and dynamic gating. Boundary Learning (Train) denotes frequency-aware boundary-scale learning around fixed relation prototypes during training. Infer. reports the inference matching distance, where Euc. denotes the original Euclidean prototype distance and Calib. denotes the elastic boundary-aware calibrated distance.

Method	VL + Dynamic Gate	Boundary Learning	Infer.	PredCls		
				mR@50	mR@100	Mean _{mR}
Baseline	×	×	Euc.	31.5	33.8	32.65
Baseline + VL + Dynamic Gate	✓	×	Euc.	32.4	34.8	33.6
Baseline + Boundary Learning	×	✓	Calib.	<u>35.4</u>	<u>38.8</u>	<u>37.1</u>
SGE-Net (Full)	✓	✓	Calib.	36.5	40.1	38.3

Fig. 7 shows a qualitative comparison on Open Images V6. For the pair *man* and *football*, both the ground truth and PE-Net predict the coarse relation *holds*, whereas SGE-Net further predicts the more specific relation *kick*. This example highlights annotation-granularity mismatch in a scene with relatively clear interactions, where SGE-Net predicts a more specific relation while preserving the basic interaction semantics. This qualitative observation is broadly consistent with the quantitative trend in Table 3, where SGE-Net obtains higher $wmAP_{rel}$, $wmAP_{phr}$, and $score_{wtd}$.

4.5 Ablation Studies

To more clearly explain how each design contributes to model performance, the following ablations start from the overall framework, then move to the key mechanisms and related hyperparameters in Stage 2, and finally compare the fusion strategies and semantic-prior sources used in Stage 1.

Core component analysis. To verify the contribution of the major components, Table 4 compares the effects of visual-semantic fusion with dynamic gating and Stage-2 elastic boundary learning. The baseline denotes PE-Net and achieves Mean_{mR} of 32.65% under the same experimental setting. Adding VL with dynamic gating improves Mean_{mR} to 33.6%, indicating that relation-predicate semantic features and adaptive visual-semantic fusion provide complementary cues for relation representation. In contrast, adding elastic boundary learning and calibrated inference without VL and dynamic gating raises Mean_{mR} to 37.1% and mR@100 to 38.8%, bringing much larger gains. This comparison isolates the contribution of boundary learning and inference-time calibrated matching without introducing the Stage-1 semantic-prior setting. The full model achieves the best results on all three metrics, further indicating that Stage-1 relation-feature refinement and Stage-2 boundary learning are highly complementary.

Training-versus-inference applicability of elastic boundary-aware distance calibration. Table 5 examines whether the elastic boundary-aware calibrated distance should be introduced during training, inference, or both. The checkmarks in

Table 5 Stage-wise ablation of elastic boundary-aware distance calibration. Calib. denotes the elastic boundary-aware calibrated distance. The checkmarks indicate whether Calib. is used in the training classification objective or in inference matching, rather than whether the whole Stage-2 boundary-learning process is enabled.

Setting	Train Calib.	Infer Calib.	PredCls		
			mR@50	mR@100	Mean _{mR}
Exp1: No Calibration	×	×	<u>31.2</u>	<u>33.4</u>	<u>32.3</u>
Exp2: Train Calibration Only	✓	×	30.3	32.7	31.5
Exp3: Train + Inference Calibration	✓	✓	14.0	17.1	15.55
Exp4: Inference Calibration Only	×	✓	36.5	40.1	38.3

this table indicate only whether the calibrated distance is used in the training classification objective or in inference matching, rather than whether the entire Stage-2 boundary-learning process is enabled. Exp1 keeps both training and inference under the original Euclidean prototype distance and can therefore be regarded as continued prototype-space fine-tuning without calibrated matching. Exp2 introduces calibrated distance during training but still evaluates with Euclidean matching, leading to a slight performance drop. This indicates that involving calibrated distance in the training objective may affect boundary-scale estimation, but its influence is limited when final prediction still relies on the original Euclidean prototype distance. Exp3 applies calibrated distance in both training and inference, where performance degrades substantially and Mean_{mR} falls to 15.55%. This suggests that directly using calibrated distance in the training classification objective can make boundary-scale estimation less stable, because σ is then involved in inter-class distance optimization and may be driven more by immediate discrimination than by the local variation range around fixed relation prototypes. When the resulting calibrated distance is further used during inference, the unstable boundary-scale estimates can have a stronger impact on final matching. By contrast, Exp4 learns boundary scales without using calibrated distance in the training classification objective and applies the learned boundary information only during inference, achieving the best performance with mR@100 of 40.1% and Mean_{mR} of 38.3%. Fig. 8 further visualizes the learned predicate-wise boundary-scale distributions under the worst-performing and best-performing calibration settings. As shown in Fig. 8(a), Exp3 yields a sharper head-to-tail change, where head predicates are assigned noticeably smaller σ values and the body/tail groups exhibit stronger fluctuations. In contrast, Fig. 8(b) shows that Exp4 produces a smoother and more progressive predicate-wise σ distribution: head predicates retain non-negligible boundary scales, while tail predicates preserve relatively larger σ values. This trend is more consistent with the intended frequency-aware boundary allocation and provides an intuitive explanation for the superior performance of Exp4 in Table 5. Overall, Table 5 and Fig. 8 indicate that elastic boundary-aware distance calibration is most effective when used as an inference-time matching strategy after boundary scales have been learned.

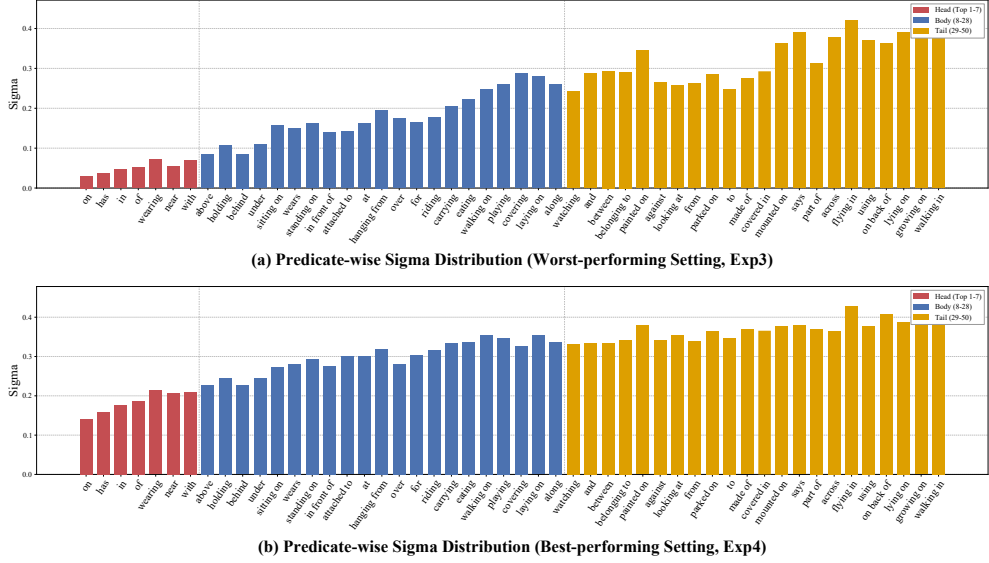


Fig. 8 Predicate-wise boundary-scale (σ) distributions under different calibration settings. Predicates are sorted in descending order of training frequency and grouped into head, body, and tail classes. Subfigure (a) corresponds to Exp3, where calibrated distance is used in both training and inference, while subfigure (b) corresponds to Exp4, where calibrated distance is used only for inference matching, as reported in Table 5.

Elastic boundary-learning mechanism analysis. To verify the necessity of the major components in Stage 2, Table 6 removes the explicit frequency compensation and the implicit adaptive virtual sampling, respectively. Without explicit frequency compensation, Mean_{mR} drops to 36.75%, indicating that a uniform boundary adjustment is insufficient to reflect the frequency differences among relation categories, while frequency-aware compensation helps preserve more suitable decision boundaries for low-frequency fine-grained relations. Without implicit adaptive virtual sampling, Mean_{mR} further decreases to 36.15%, suggesting that explicit frequency compensation alone is not enough to support stable boundary learning, and that virtual sampling around fixed relation prototypes is also essential for forming boundaries. The full model achieves the best performance on all three metrics, indicating that explicit frequency compensation and implicit adaptive virtual sampling work together from two complementary perspectives: boundary adjustment and local neighborhood constraint.

Distance threshold R for elastic virtual boundaries. Table 7 reports the effect of different distance thresholds. When R is too small, the boundary constraint becomes overly strict and the admissible effective range around a fixed relation prototype is too narrow, which is especially unfavorable for covering the reasonable variation of low-frequency fine-grained relations near prototypes. As a result, mR@100 is only 38.1% and Mean_{mR} is 35.5% when $R = 0.03$. As R increases to 0.2 and 0.3, performance steadily improves, and the best results are achieved at $R = 0.3$, indicating that the boundary range at this point can both preserve stable discrimination near the prototype and reserve necessary elastic space for low-frequency relations. When R is

Table 6 Ablation study on the elastic prototype-boundary construction strategy in Stage 2.

Ablation Models (Stage2)	Predicate Classification		
	mR@50	mR@100	Mean _{mR}
w/o Explicit Frequency Compensation	35.3	38.2	36.75
w/o Implicit Adaptive Virtual Sampling	34.7	37.6	36.15
SGE-Net	36.5	40.1	38.3

Table 7 Effect of the distance threshold R of elastic virtual boundaries on predicate-classification performance.

R	Predicate Classification		
	mR@50	mR@100	Mean _{mR}
0.03	32.9	38.1	35.5
0.2	35.8	38.9	37.35
0.3	36.5	40.1	38.3
0.4	36.2	39.8	38.0
0.8	35.7	39.1	37.4

further increased to 0.4 and 0.8, performance declines, suggesting that an overly wide boundary weakens discrimination among neighboring relation categories. We therefore set $R = 0.3$.

Base sample count Q_{base} . Table 8 reflects the influence of the virtual sampling scale on boundary learning. When Q_{base} is small, the sampling points around fixed relation prototypes are too sparse to provide a stable neighborhood reference for boundary learning, so mR@100 is only 35.8% and Mean_{mR} is 33.5% at $Q_{base} = 2$. As Q_{base} increases, performance steadily improves and reaches the optimum at $Q_{base} = 40$, indicating that a moderate sampling scale can better cover effective variation in prototype neighborhoods and is especially helpful for stabilizing the decision boundaries of low-frequency relations. When Q_{base} is further increased to 60 and 80, the performance slightly declines, suggesting that overly dense virtual sampling does not continuously bring gains and may instead introduce redundant boundary constraints. We therefore set $Q_{base} = 40$.

Loss-balancing coefficient $\lambda^{(2)}$. Table 9 shows how the strength of the Stage-2 boundary constraint affects the results. When $\lambda^{(2)}$ is small, the boundary loss has limited contribution in overall optimization, so Stage 2 behaves more like basic classification around prototype centers and the adjustment of low-frequency relation boundaries remains insufficient. For example, Mean_{mR} is 37.0% when $\lambda^{(2)} = 0.01$. Increasing it to 0.08 and 0.1 further improves mR@100 and Mean_{mR}, with $\lambda^{(2)} = 0.1$ giving the best results, indicating a more suitable balance between the basic classification term and the boundary constraint. When it is further increased to 0.12 and 0.2, performance declines, which suggests that overly strong boundary constraints begin

Table 8 Ablation study on the base sample count Q_{base} .

Q_{base}	Predicate Classification		
	mR@50	mR@100	Mean $_{mR}$
2	31.2	35.8	33.5
10	34.8	37.9	36.35
20	35.4	38.8	37.1
30	36.1	39.2	37.65
40	36.5	40.1	38.3
60	36.2	<u>39.8</u>	<u>38.0</u>
80	<u>36.3</u>	39.6	37.95

Table 9 Ablation study on the Stage-2 loss-balancing coefficient $\lambda^{(2)}$.

$\lambda^{(2)}$	Predicate Classification		
	mR@50	mR@100	Mean $_{mR}$
0.01	35.1	38.9	37.0
0.08	<u>36.1</u>	<u>39.7</u>	<u>37.9</u>
0.1	36.5	40.1	38.3
0.12	36.0	39.6	37.8
0.2	35.8	39.3	37.55

to interfere with the basic discrimination provided by prototype centers. We therefore set $\lambda^{(2)} = 0.1$.

Elastic boundary-aware distance-calibration weight λ . Table 10 shows the weight allocation between the original prototype distance and the elastic boundary-aware term. When $\lambda = 0.2$, the calibrated distance still relies more heavily on the boundary-aware term $(1 - \lambda)\frac{d_i}{\sigma_j}$, and the model already achieves relatively strong results, with mR@100 of 39.4% and Mean $_{mR}$ of 37.8%. Increasing λ to 0.3 further improves performance to the best mR@100 and Mean $_{mR}$ values of 40.1% and 38.3%, respectively, indicating a more suitable balance between the original Euclidean distance and elastic boundary-aware distance calibration. In this regime, boundary information can still provide effective supplementary evidence when the original distances of candidate relation categories are close, while the basic geometric matching relationship is better preserved. Further increasing λ to 0.5 and 0.8 reduces the overall performance, especially when $\lambda = 0.8$, suggesting that if the original-distance term dominates too much, the useful contribution of boundary information is weakened. We therefore set $\lambda = 0.3$ for elastic boundary-aware distance calibration.

Annealed upper bound of explicit frequency compensation λ_t . Table 11 reflects how the strength of frequency compensation affects Stage-2 boundary learning. When λ_t is too small, frequency differences cannot be sufficiently transformed into boundary adjustment, so low-frequency relations still fail to receive adequate compensation. For example, mR@100 is 37.7% and Mean $_{mR}$ is 36.3% when $\lambda_t = 0.01$.

Table 10 Ablation study on the elastic boundary-aware distance-calibration weight λ .

λ	Predicate Classification		
	mR@50	mR@100	Mean _{mR}
0.2	<u>36.2</u>	<u>39.4</u>	<u>37.8</u>
0.3	36.5	40.1	38.3
0.5	35.7	36.7	36.2
0.8	33.9	35.7	34.8

Table 11 Ablation study on the annealed upper bound of explicit frequency compensation λ_t .

λ_t	Predicate Classification		
	mR@50	mR@100	Mean _{mR}
0.01	34.9	37.7	36.3
0.05	36.5	40.1	38.3
0.1	<u>35.6</u>	<u>38.9</u>	<u>37.25</u>
0.25	34.3	36.7	35.5
0.5	33.1	35.9	34.5
1.0	32.3	34.5	33.4

Increasing λ_t to 0.05 yields the best results on all three metrics, indicating that a moderate compensation strength is more effective for translating frequency priors into useful boundary adjustment. Further increasing λ_t to 0.1, 0.25, 0.5, and 1.0 leads to gradually declining performance, suggesting that overly strong compensation causes excessive expansion of low-frequency relation boundaries and weakens stable discrimination from semantically similar relation categories. We therefore set the upper bound of λ_t to 0.05.

Hyperparameter-sensitivity visualization. Fig. 9 summarizes the variations of Mean_{mR} under different values of R , Q_{base} , $\lambda^{(2)}$, λ , and λ_t . Most of these hyperparameters perform best around moderate values, whereas overly small or overly large settings cause varying degrees of performance drop. This observation indicates that the performance gain of SGE-Net does not rely on any extreme hyperparameter choice, but on a relatively well-balanced coordination among prototype space, boundary learning, and elastic boundary-aware distance calibration. More specifically, R and Q_{base} mainly affect the coverage and sampling density of the boundary range around fixed relation prototypes; values that are too small fail to provide sufficient constraints, while values that are too large bring overly wide or redundant boundary descriptions. By contrast, $\lambda^{(2)}$, λ , and λ_t correspond to the strength of boundary constraints during training, the strength of elastic boundary-aware distance calibration, and the magnitude of frequency compensation, respectively, and their better results likewise occur

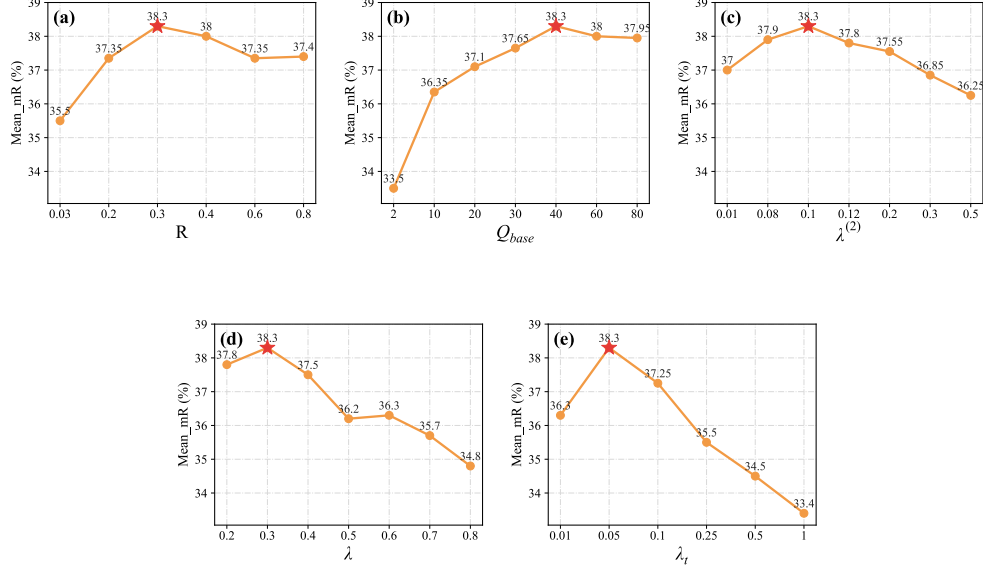


Fig. 9 Hyperparameter-sensitivity analysis. The figure summarizes the influence trends of R , Q_{base} , $\lambda^{(2)}$, λ , and λ_t on Mean_{mR} .

Table 12 Ablation study on multimodal feature-fusion strategies in Stage 1. “Add” denotes element-wise addition between visual features and semantic features.

Ablation Models (Stage1)	Predicate Classification		
	mR@50	mR@100	Mean _{mR}
No Gate(Add)	31.2	33.2	32.2
Fixed Gate	<u>31.6</u>	<u>33.9</u>	<u>32.75</u>
Dynamic Gate	32.4	34.8	33.6

within moderate ranges. Together with the quantitative tables above, this figure further shows that the key factors of the proposed method need to stay in relative balance in order to more effectively improve long-tailed relation prediction.

Multimodal feature-fusion strategy analysis. To verify the necessity of dynamic gating in Stage 1, Table 12 compares different fusion strategies. Directly adding visual features and semantic features achieves only 32.2% Mean_{mR} , indicating that simple superposition is insufficient to distinguish the information truly needed by the current relation instance. Introducing a fixed gate increases Mean_{mR} to 32.75%, which shows that basic weighting of different modalities is helpful, but a fixed fusion ratio still cannot adapt to the differences among entity pairs. In contrast, dynamic gating achieves the best results on all three metrics, especially raising mR@100 to 34.8%. This suggests that adaptively adjusting the fusion ratio between visual and semantic information according to the current relation context is more beneficial for obtaining stable and discriminative multimodal relation features.

Table 13 Ablation study on different semantic-prior sources in Stage 1.

Ablation Models (Stage1)	Predicate Classification		
	mR@50	mR@100	Mean _{mR}
Random tensor	31.0	33.1	32.05
GloVe	<u>31.4</u>	<u>33.6</u>	<u>32.5</u>
Qwen3-14B	32.4	34.8	33.6

Table 14 Computational overhead analysis compared to the baseline PE-Net.

Model	Params (M)	Train Time/Iter (s)	Infer Time/Img (s)
PE-Net	61.4	0.42	0.115
SGE-Net(Ours)	66.6	0.47	0.120

Semantic-prior source analysis. To evaluate the effects of different semantic sources, Table 13 compares random tensors, GloVe, and Qwen3-14B-based structured priors. Random tensors achieve only 32.05% Mean_{mR}, suggesting that inputs without semantic constraints cannot provide effective guidance for cross-modal fusion. Replacing them with GloVe improves Mean_{mR} to 32.5%, indicating that basic word vectors offer useful semantic complement, though their expressiveness remains limited. In contrast, Qwen3-14B-based structured priors achieve the best results on all three metrics, particularly improving mR@100 to 34.8% and Mean_{mR} to 33.6%. This result shows that structured descriptions more closely aligned with relation-predicate semantics help Stage 1 construct higher-quality relation features.

Computational overhead analysis. Table 14 provides a brief analysis of the computational overhead. While introducing semantic retrieval and virtual sampling, SGE-Net maintains reasonable computational costs. Compared to the baseline PE-Net, our method increases the parameter count by approximately 5.2M (mainly from the multimodal fusion modules). During training, the two-stage process increases the per-iteration time by about 12%, while the inference time overhead is marginal (less than 5% increase), as the semantic library and boundaries are pre-computed or fixed. This indicates that the performance gains are achieved at an acceptable computational cost.

4.6 Stability and Reproducibility Analysis

To further examine the stability of SGE-Net under different random initializations, we repeat the PredCls experiment on Visual Genome using four different random seeds while keeping all other settings unchanged. As shown in Table 15, SGE-Net maintains consistently strong performance across different runs, with mean results of 35.90% on mR@50, 39.70% on mR@100, and 37.80% on Mean_{mR}. The corresponding standard deviations are 0.41, 0.29, and 0.35, respectively. In addition, the fluctuation across the

Table 15 Random-seed stability analysis of SGE-Net on PredCls over four runs on Visual Genome.

Seed	mR@50	mR@100	Mean _{mR}
19991001	36.5	40.1	38.3
3047	35.8	39.6	37.7
42	35.6	39.4	37.5
2026	35.7	39.7	37.7
Mean $\pm$ Std	35.90 $\pm$ 0.41	39.70 $\pm$ 0.29	37.80 $\pm$ 0.35

four runs stays within 1.0 point on all three metrics. These results suggest that the reported gain is stable on Visual Genome PredCls under the tested seeds.

4.7 Visualization Results

Qualitative analysis. To provide an intuitive view of how different methods structure the relation features of several semantically related predicates that are prone to confusion, we use t-SNE to project the relation features of six such predicates into a two-dimensional space. The visualization results are shown in Fig. 10. In Motifs, the selected predicates are relatively mixed, and no clear prototype-centered structure is observed. PE-Net improves the overall organization by drawing relation features closer to their corresponding prototypes, but noticeable overlap still remains among semantically related predicates. In contrast, SGE-Net forms a more structured distribution around prototypes, and the corresponding elastic boundaries characterize category-dependent local ranges for different predicates. In particular, for predicates that are semantically close to *on* and therefore prone to confusion, the projected features are more clearly organized around their corresponding prototypes. The dashed ellipses in Fig. 10 provide an intuitive view of the learned tolerance ranges and feature organization. Together with the predicate-wise boundary-scale distributions in Fig. 8, this visualization shows that SGE-Net not only learns frequency-aware boundary scales at the category level, but also forms a more organized feature space around relation prototypes for semantically confusing predicates.

Visual Genome. Figs. 11–13 provide additional qualitative comparisons between PE-Net and SGE-Net on Visual Genome. These examples show that both methods can recognize the basic relations among objects, while SGE-Net sometimes produces more specific relation expressions. For example, in Fig. 11, SGE-Net predicts *part of* and *mounted on*; in Figs. 12 and 13, it also produces predicates such as *laying on*, *walking on*, and *covering*. These examples further illustrate the annotation-granularity mismatch in SGG: SGE-Net often produces more specific relation expressions while preserving the basic relation semantics. The mismatches mainly occur among semantically related predicates rather than as random prediction errors, suggesting that the learned relation space captures meaningful predicate-level structure.

Open Images V6. Fig. 14 provides the corresponding example on Open Images V6. For the pair *man* and *horse*, both the ground truth and PE-Net give the more general relation *interacts with*, whereas SGE-Net further predicts the more specific relation *ride*. This example further illustrates annotation-granularity mismatch, where

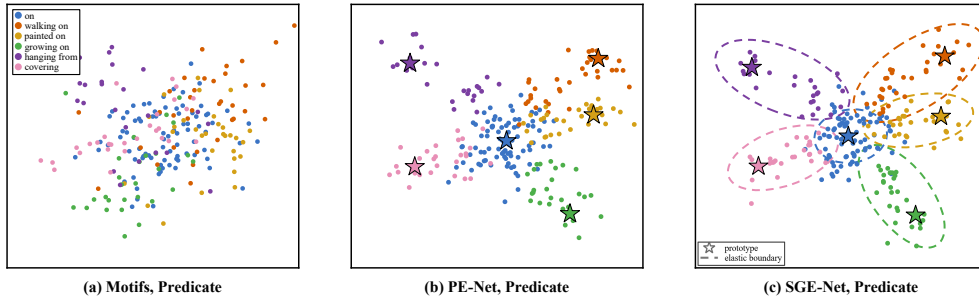


Fig. 10 t-SNE visualization of six semantically related predicates that are prone to confusion under Motifs, PE-Net, and SGE-Net. Each dot denotes a predicate feature in the projected space. The star markers denote prototypes, and the dashed ellipses in SGE-Net provide a qualitative illustration of the learned tolerance ranges.

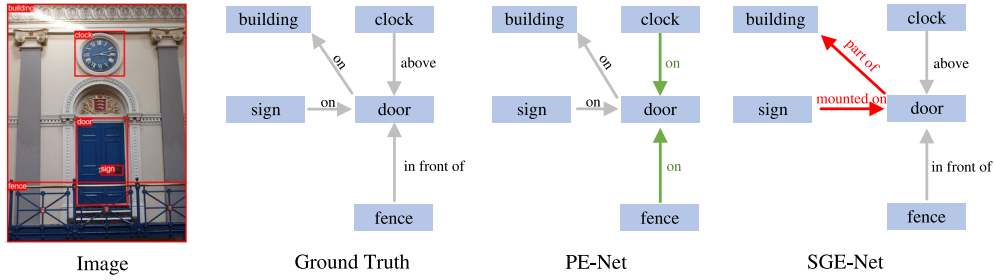


Fig. 11 Qualitative comparison between SGE-Net and PE-Net on Visual Genome (Example 1). Predicates in red denote granularity-mismatch cases in which the prediction is more specific than the ground truth, predicates in black match the ground truth, and predicates in green differ from the ground truth.

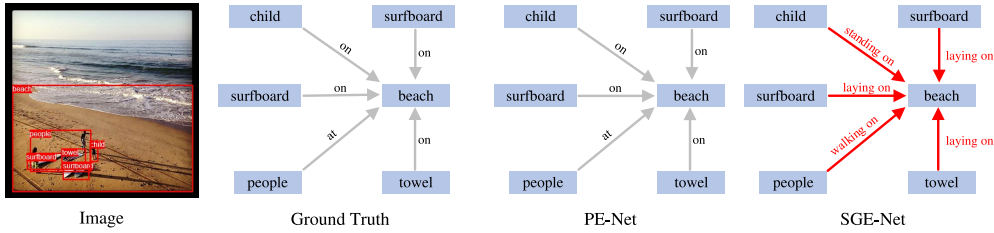


Fig. 12 Qualitative comparison between SGE-Net and PE-Net on Visual Genome (Example 2). Predicates in red denote granularity-mismatch cases in which the prediction is more specific than the ground truth, predicates in black match the ground truth, and predicates in green differ from the ground truth.

SGE-Net predicts a more specific interaction while preserving the basic relation semantics. This observation is broadly consistent with the quantitative trend reported earlier on Open Images V6.

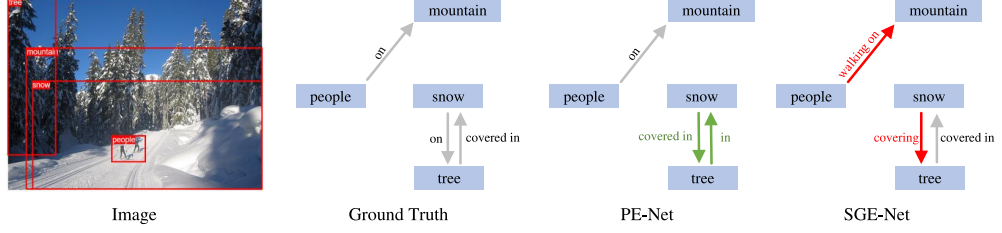


Fig. 13 Qualitative comparison between SGE-Net and PE-Net on Visual Genome (Example 3). Predicates in red denote granularity-mismatch cases in which the prediction is more specific than the ground truth, predicates in black match the ground truth, and predicates in green differ from the ground truth.

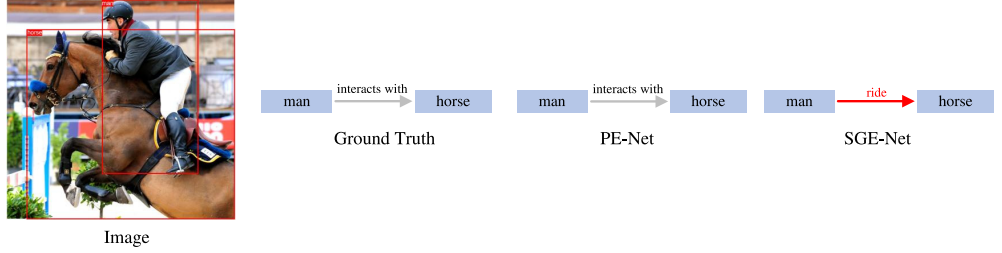


Fig. 14 Qualitative comparison between SGE-Net and PE-Net on Open Images V6. Predicates in red denote granularity-mismatch cases in which the prediction is more specific than the ground truth, while predicates in black match the ground truth.

5 Future Work

Several directions remain valuable for future exploration. First, the offline semantic library can be further enhanced with higher-quality predicate descriptions or adaptive semantic refinement. Second, annotation-granularity-aware evaluation remains valuable for SGG, because fine-grained predictions may not always be fully reflected when benchmark annotations use broader predicates. Third, exploring richer boundary parameterizations, such as anisotropic or multi-dimensional boundaries, may further improve the flexibility of frequency-aware prototype-boundary learning.

6 Conclusion

In this work, we proposed SGE-Net, a two-stage frequency-aware elastic prototype boundary learning framework. Stage 1 constructs a prototype-aligned multimodal relation-feature space by fusing visual and semantic features with dynamic gating, while Stage 2 revisits long-tailed scene graph generation from the perspective of relation-category decision boundaries. Specifically, under fixed relation prototypes, Stage 2 learns relation-category-specific boundary scales through explicit frequency compensation and frequency-adaptive virtual sampling, so that relation categories with different frequencies obtain more suitable decision boundaries within a unified prototype space. During inference, we further design elastic boundary-aware distance

calibration, which translates the learned tolerance margins into a category-level distance adjustment for prototype matching. Experimental results on Visual Genome and Open Images V6 demonstrate consistent improvements on the reported settings, and the four-seed analysis on Visual Genome PredCls further suggests that the gain is stable in that setting. These findings suggest that, in long-tailed relation prediction, beyond the separability of prototype centers, incorporating frequency-aware tolerance margins is a promising direction for alleviating fine-grained relation confusion. More broadly, the idea of learning frequency-aware prototype boundaries may also provide useful inspiration for other long-tailed recognition tasks where rare categories exhibit larger feature variation.

Declarations

Competing interests

The authors declare no competing interests.

References

- Biswas, B.A., Ji, Q.: Probabilistic debiasing of scene graphs. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023, pp. 10429–10438. IEEE, Vancouver, BC, Canada (2023). <https://doi.org/10.1109/CVPR52729.2023.01005> . <https://doi.org/10.1109/CVPR52729.2023.01005>
- Bordes, F., Pang, R.Y., Ajay, A., Li, A.C., Bardes, A., Petryk, S., Mañas, O., Lin, Z., Mahmoud, A., Jayaraman, B., Ibrahim, M., Hall, M., Xiong, Y., Lebensold, J., Ross, C., Jayakumar, S., Guo, C., Bouchacourt, D., Al-Tahan, H., Padthe, K., Sharma, V., Xu, H., Tan, X.E., Richards, M., Lavoie, S., Astolfi, P., Hemmat, R.A., Chen, J., Tirumala, K., Assouel, R., Moayeri, M., Talattof, A., Chaudhuri, K., Liu, Z., Chen, X., Garrido, Q., Ullrich, K., Agrawal, A., Saenko, K., Celikyilmaz, A., Chandra, V.: An introduction to vision-language modeling. CoRR **abs/2405.17247** (2024) <https://doi.org/10.48550/ARXIV.2405.17247> 2405.17247
- Chiou, M., Ding, H., Yan, H., Wang, C., Zimmermann, R., Feng, J.: Recovering the unbiased scene graphs from the biased ones. In: Shen, H.T., Zhuang, Y., Smith, J.R., Yang, Y., César, P., Metze, F., Prabhakaran, B. (eds.) MM '21: ACM Multimedia Conference, Virtual Event, China, October 20 - 24, 2021, pp. 1581–1590. ACM, Virtual Event, China (2021). <https://doi.org/10.1145/3474085.3475297> . <https://doi.org/10.1145/3474085.3475297>
- Chen, M., Liao, Y., Liu, S., Chen, Z., Wang, F., Qian, C.: Reformulating HOI detection as adaptive set prediction. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, Virtual, June 19-25, 2021, pp. 9004–9013. Computer Vision Foundation / IEEE, Virtual Event (2021). <https://doi.org/10.1109/CVPR46437.2021.00889> . <https://doi.org/10.1109/CVPR46437.2021.00889>

- Chen, G., Li, J., Wang, W.: Scene graph generation with role-playing large language models. In: Globersons, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J.M., Zhang, C. (eds.) *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024* (2024). http://papers.nips.cc/paper_files/paper/2024/hash/ee74a6ade401e200985e2421b20bbae4-Abstract-Conference.html
- Chen, L., Song, Y., Cai, Y., Lu, J., Li, Y., Xie, Y., Wang, C., He, G.: Multi-prototype space learning for commonsense-based scene graph generation. In: Wooldridge, M.J., Dy, J.G., Natarajan, S. (eds.) *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2024, February 20-27, 2024, Vancouver, BC, Canada*, pp. 1129–1137. AAAI Press, Vancouver, BC, Canada (2024). <https://doi.org/10.1609/AAAI.V38I2.27874> . <https://doi.org/10.1609/AAAI.V38I2.27874>
- Chen, T., Yu, W., Chen, R., Lin, L.: Knowledge-embedded routing network for scene graph generation. In: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 6163–6171. Computer Vision Foundation / IEEE, Long Beach, CA, USA (2019). <https://doi.org/10.1109/CVPR.2019.00632> . <https://doi.org/10.1109/CVPR.2019.00632>
- Cong, Y., Yang, M.Y., Rosenhahn, B.: Reltr: Relation transformer for scene graph generation. *IEEE Trans. Pattern Anal. Mach. Intell.* **45**(9), 11169–11183 (2023) <https://doi.org/10.1109/TPAMI.2023.3268066>
- Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Burstein, J., Doran, C., Solorio, T. (eds.) *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pp. 4171–4186. Association for Computational Linguistics, Minneapolis, MN, USA (2019). <https://doi.org/10.18653/V1/N19-1423> . <https://doi.org/10.18653/V1/N19-1423>
- Guo, W., Zhang, Y., Yang, J., Yuan, X.: Re-attention for visual question answering. *IEEE Trans. Image Process.* **30**, 6730–6743 (2021) <https://doi.org/10.1109/TIP.2021.3097180>
- Hildebrandt, M., Li, H., Koner, R., Tresp, V., Günnemann, S.: Scene graph reasoning for visual question answering. *CoRR* **abs/2007.01072** (2020) [2007.01072](https://arxiv.org/abs/2007.01072)
- Jung, D., Kim, S., Kim, W.H., Cho, M.: Devil’s on the edges: Selective quad attention for scene graph generation. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pp. 18664–18674. IEEE, Vancouver, BC, Canada (2023). <https://doi.org/10.1109/CVPR.2023.01072>

CVPR52729.2023.01790 . <https://doi.org/10.1109/CVPR52729.2023.01790>

- Johnson, J., Krishna, R., Stark, M., Li, L., Shamma, D.A., Bernstein, M.S., Fei-Fei, L.: Image retrieval using scene graphs. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015, pp. 3668–3678. IEEE Computer Society, Boston, MA, USA (2015). <https://doi.org/10.1109/CVPR.2015.7298990> . <https://doi.org/10.1109/CVPR.2015.7298990>
- Jiang, B., Zhuang, Z., Shivakumar, S.S., Taylor, C.J.: Enhancing scene graph generation with hierarchical relationships and commonsense knowledge. In: IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2025, Tucson, AZ, USA, February 26 - March 6, 2025, pp. 8883–8894. IEEE, Tucson, AZ, USA (2025). <https://doi.org/10.1109/WACV61041.2025.00861> . <https://doi.org/10.1109/WACV61041.2025.00861>
- Kundu, S., Aakur, S.N.: IS-GGT: iterative scene graph generation with generative transformers. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023, pp. 6292–6301. IEEE, Vancouver, BC, Canada (2023). <https://doi.org/10.1109/CVPR52729.2023.00609> . <https://doi.org/10.1109/CVPR52729.2023.00609>
- Kim, H., Kim, S., Ahn, D., Lee, J.T., Ko, B.C.: Scene graph generation strategy with co-occurrence knowledge and learnable term frequency. In: Salakhutdinov, R., Kolter, Z., Heller, K.A., Weller, A., Oliver, N., Scarlett, J., Berkenkamp, F. (eds.) Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024. Proceedings of Machine Learning Research, pp. 24094–24109. PMLR / OpenReview.net, Vienna, Austria (2024). <https://proceedings.mlr.press/v235/kim24n.html>
- Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J.R.R., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Mallocci, M., Kolesnikov, A., Duerig, T., Ferrari, V.: The open images dataset V4. *Int. J. Comput. Vis.* **128**(7), 1956–1981 (2020) <https://doi.org/10.1007/S11263-020-01316-Z>
- Kim, K., Yoon, K., In, Y., Moon, J., Kim, D., Park, C.: Adaptive self-training framework for fine-grained scene graph generation. In: The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024. OpenReview.net, Vienna, Austria (2024). <https://openreview.net/forum?id=WipsLtH77t>
- Liu, H., Bhanu, B.: Repsgg: Novel representations of entities and relationships for scene graph generation. *IEEE Trans. Pattern Anal. Mach. Intell.* **46**(12), 8018–8035 (2024) <https://doi.org/10.1109/TPAMI.2024.3402143>
- Li, L., Chen, L., Huang, Y., Zhang, Z., Zhang, S., Xiao, J.: The devil is in the labels: Noisy label correction for robust scene graph generation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA,

- June 18-24, 2022, pp. 18847–18856. IEEE, New Orleans, LA, USA (2022). <https://doi.org/10.1109/CVPR52688.2022.01830> . <https://doi.org/10.1109/CVPR52688.2022.01830>
- Lin, X., Ding, C., Zeng, J., Tao, D.: Gps-net: Graph property sensing network for scene graph generation. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020, pp. 3743–3752. Computer Vision Foundation / IEEE, Seattle, WA, USA (2020). <https://doi.org/10.1109/CVPR42600.2020.00380> . <https://doi.org/10.1109/CVPR42600.2020.00380>
- Lyu, X., Gao, L., Zeng, P., Shen, H.T., Song, J.: Adaptive fine-grained predicates learning for scene graph generation. *IEEE Trans. Pattern Anal. Mach. Intell.* **45**(11), 13921–13940 (2023) <https://doi.org/10.1109/TPAMI.2023.3298356>
- Li, Y., Wang, T., Wu, K., Wang, L., Guo, X., Wang, W.: Fine-grained scene graph generation via sample-level bias prediction. In: Leonardis, A., Ricci, E., Roth, S., Russakovsky, O., Sattler, T., Varol, G. (eds.) *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part XXVI. Lecture Notes in Computer Science*, pp. 18–35. Springer, Milan, Italy (2024). https://doi.org/10.1007/978-3-031-73347-5_2 . https://doi.org/10.1007/978-3-031-73347-5_2
- Liu, R., Wang, X., Wang, W., Yang, Y.: Bird’s-eye-view scene graph for vision-language navigation. In: *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pp. 10934–10946. IEEE, Paris, France (2023). <https://doi.org/10.1109/ICCV51070.2023.01007> . <https://doi.org/10.1109/ICCV51070.2023.01007>
- Liu, R., Wang, W., Yang, Y.: Vision-language navigation with energy-based policy. In: Globersons, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J.M., Zhang, C. (eds.) *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024* (2024). http://papers.nips.cc/paper_files/paper/2024/hash/c3ec89aed795f9deeff3f1390c3bd882-Abstract-Conference.html
- Li, X., Xiao, J., Chen, G., Feng, Y., Yang, Y., Liu, A., Chen, L.: Decomposed prototype learning for few-shot scene graph generation. *ACM Trans. Multim. Comput. Commun. Appl.* **21**(1), 30–13024 (2025) <https://doi.org/10.1145/3700877>
- Li, L., Xiao, J., Shi, H., Zhang, H., Yang, Y., Liu, W., Chen, L.: NICEST: noisy label correction and training for robust scene graph generation. *IEEE Trans. Pattern Anal. Mach. Intell.* **46**(10), 6873–6888 (2024) <https://doi.org/10.1109/TPAMI.2024.3387349>
- Li, R., Zhang, S., He, X.: SGTR+: end-to-end scene graph generation with transformer. *IEEE Trans. Pattern Anal. Mach. Intell.* **46**(4), 2191–2205 (2024) <https://doi.org/10.1109/TPAMI.2023.3332246>

- Li, R., Zhang, S., Wan, B., He, X.: Bipartite graph network with adaptive message passing for unbiased scene graph generation. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, Virtual, June 19-25, 2021, pp. 11109–11119. Computer Vision Foundation / IEEE, Virtual Event (2021). <https://doi.org/10.1109/CVPR46437.2021.01096> . <https://doi.org/10.1109/CVPR46437.2021.01096>
- Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: Moschitti, A., Pang, B., Daelemans, W. (eds.) Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A Meeting of SIGDAT, a Special Interest Group of The ACL, pp. 1532–1543. ACL, Doha, Qatar (2014). <https://doi.org/10.3115/V1/D14-1162> . <https://doi.org/10.3115/V1/D14-1162>
- Sudhakaran, G., Dharmi, D.S., Kersting, K., Roth, S.: Vision relation transformer for unbiased scene graph generation. In: IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023, pp. 21825–21836. IEEE, Paris, France (2023). <https://doi.org/10.1109/ICCV51070.2023.02000> . <https://doi.org/10.1109/ICCV51070.2023.02000>
- Sun, W., Lin, X., Shi, Y., Zhang, C., Wu, H., Zheng, S.: Sparsedrive: End-to-end autonomous driving via sparse scene representation. In: IEEE International Conference on Robotics and Automation, ICRA 2025, Atlanta, GA, USA, May 19-23, 2025, pp. 8795–8801. IEEE, Atlanta, GA, USA (2025). <https://doi.org/10.1109/ICRA55743.2025.11128800> . <https://doi.org/10.1109/ICRA55743.2025.11128800>
- Suhail, M., Mittal, A., Siddiquie, B., Broaddus, C., Eledath, J., Medioni, G.G., Sigal, L.: Energy-based learning for scene graph generation. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, Virtual, June 19-25, 2021, pp. 13936–13945. Computer Vision Foundation / IEEE, Virtual Event (2021). <https://doi.org/10.1109/CVPR46437.2021.01372> . <https://doi.org/10.1109/CVPR46437.2021.01372>
- Singh, K.P., Salvador, J., Weihs, L., Kembhavi, A.: Scene graph contrastive learning for embodied navigation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10884–10894 (2023)
- Team, Q.: Qwen3 technical report. CoRR **abs/2505.09388** (2025) <https://doi.org/10.48550/ARXIV.2505.09388> 2505.09388
- Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., Lample, G.: Llama: Open and efficient foundation language models. CoRR **abs/2302.13971** (2023) <https://doi.org/10.48550/ARXIV.2302.13971> 2302.13971
- Tang, K., Niu, Y., Huang, J., Shi, J., Zhang, H.: Unbiased scene graph generation from biased training. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern

- Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020, pp. 3713–3722. Computer Vision Foundation / IEEE, Seattle, WA, USA (2020). <https://doi.org/10.1109/CVPR42600.2020.00377> . <https://doi.org/10.1109/CVPR42600.2020.00377>
- Tang, K., Zhang, H., Wu, B., Luo, W., Liu, W.: Learning to compose dynamic tree structures for visual contexts. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019, pp. 6619–6628. Computer Vision Foundation / IEEE, Long Beach, CA, USA (2019). <https://doi.org/10.1109/CVPR.2019.00678> . <https://doi.org/10.1109/CVPR.2019.00678>
- Wu, W., Chang, T., Li, X., Yin, Q., Hu, Y.: Vision-language navigation: a survey and taxonomy. *Neural Comput. Appl.* **36**(7), 3291–3316 (2024) <https://doi.org/10.1007/S00521-023-09217-1>
- Wang, J., Zhang, C., Huang, J., Ren, B., Deng, Z.: Improving scene graph generation with superpixel-based interaction learning. In: Saddik, A.E., Mei, T., Cucchiara, R., Bertini, M., Vallejo, D.P.T., Atrey, P.K., Hossain, M.S. (eds.) *Proceedings of the 31st ACM International Conference on Multimedia, MM 2023, Ottawa, ON, Canada, 29 October 2023- 3 November 2023*, pp. 1809–1820. ACM, Ottawa, ON, Canada (2023). <https://doi.org/10.1145/3581783.3611889> . <https://doi.org/10.1145/3581783.3611889>
- Xie, F., Wu, Y., Liu, F.: Frequency-adaptive prototype learning for long-tail scene graph generation. *J. King Saud Univ. Comput. Inf. Sci.* (2026) <https://doi.org/10.1007/s44443-026-00546-4>
- Xu, D., Zhu, Y., Choy, C.B., Fei-Fei, L.: Scene graph generation by iterative message passing. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017, pp. 3097–3106. IEEE Computer Society, Honolulu, HI, USA (2017). <https://doi.org/10.1109/CVPR.2017.330> . <https://doi.org/10.1109/CVPR.2017.330>
- Yu, J., Chai, Y., Wang, Y., Hu, Y., Wu, Q.: Cogtree: Cognition tree loss for unbiased scene graph generation. In: Zhou, Z. (ed.) *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, pp. 1274–1280. ijcai.org, Virtual Event / Montreal, Canada (2021). <https://doi.org/10.24963/IJCAI.2021/176> . <https://doi.org/10.24963/IJCAI.2021/176>
- Yang, J., He, Y., Yang, J., Yang, L.T., Gao, Y., Dai, C.: Modality-aligned anchor learning based on multi-level fusion for accurate scene graph generation. *Inf. Fusion* **127**, 103755 (2026) <https://doi.org/10.1016/J.INFFUS.2025.103755>
- Yoon, K., Kim, K., Moon, J., Park, C.: Unbiased heterogeneous scene graph generation with relation-aware message passing neural network. In: Williams, B., Chen, Y., Neville, J. (eds.) *Thirty-Seventh AAAI Conference on Artificial Intelligence*,

- AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023, pp. 3285–3294. AAAI Press, Washington, DC, USA (2023). <https://doi.org/10.1609/AAAI.V37I3.25435> . <https://doi.org/10.1609/AAAI.V37I3.25435>
- Yuan, J., Zhu, S., Huang, S., Zhang, H., Xiao, Y., Li, Z., Wang, M.: Discriminative style learning for cross-domain image captioning. *IEEE Trans. Image Process.* **31**, 1723–1736 (2022) <https://doi.org/10.1109/TIP.2022.3145158>
- Yang, G., Zhang, J., Zhang, Y., Wu, B., Yang, Y.: Probabilistic modeling of semantic ambiguity for scene graph generation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12527–12536 (2021). <https://doi.org/10.1109/CVPR46437.2021.01234> . <https://doi.org/10.1109/CVPR46437.2021.01234>
- Zhang, Y., Hare, J.S., Prügel-Bennett, A.: Learning to count objects in natural images for visual question answering. In: *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, Vancouver, BC, Canada (2018). https://openreview.net/forum?id=B12Js_yRb
- Zheng, C., Lyu, X., Gao, L., Dai, B., Song, J.: Prototype-based embedding network for scene graph generation. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pp. 22783–22792. IEEE, Vancouver, BC, Canada (2023). <https://doi.org/10.1109/CVPR52729.2023.02182> . <https://doi.org/10.1109/CVPR52729.2023.02182>
- Zhang, J., Shih, K.J., Elgammal, A., Tao, A., Catanzaro, B.: Graphical contrastive losses for scene graph parsing. In: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 11535–11543. Computer Vision Foundation / IEEE, Long Beach, CA, USA (2019). <https://doi.org/10.1109/CVPR.2019.01180> . <https://doi.org/10.1109/CVPR.2019.01180>
- Zhang, R., Shang, Z., Wang, F., Yang, Z., Cao, S., Cen, Y., An, G.: Synergetic prototype learning network for unbiased scene graph generation. In: Cai, J., Kankanhalli, M.S., Prabhakaran, B., Boll, S., Subramanian, R., Zheng, L., Singh, V.K., César, P., Xie, L., Xu, D. (eds.) *Proceedings of the 32nd ACM International Conference on Multimedia, MM 2024, Melbourne, VIC, Australia, 28 October 2024 - 1 November 2024*, pp. 945–954. ACM, Melbourne, VIC, Australia (2024). <https://doi.org/10.1145/3664647.3680973> . <https://doi.org/10.1145/3664647.3680973>
- Zou, C., Wang, B., Hu, Y., Liu, J., Wu, Q., Zhao, Y., Li, B., Zhang, C., Zhang, C., Wei, Y., Sun, J.: End-to-end human object interaction detection with HOI transformer. In: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, Virtual, June 19-25, 2021*, pp. 11825–11834. Computer Vision Foundation / IEEE, Virtual Event (2021). <https://doi.org/10.1109/CVPR46437.2021.01165> . <https://doi.org/10.1109/CVPR46437.2021.01165>

doi.org/10.1109/CVPR46437.2021.01165

- Zemskova, T., Yudin, D.A.: 3dgraphllm: Combining semantic graphs and large language models for 3d scene understanding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 8885–8895 (2025). https://openaccess.thecvf.com/content/ICCV2025/papers/Zemskova_3DGraphLLM_Combining_Semantic_Graphs_and_Large_Language_Models_for_3D_ICCV_2025_paper.pdf
- Zhang, A., Yao, Y., Chen, Q., Ji, W., Liu, Z., Sun, M., Chua, T.: Fine-grained scene graph generation with data transfer. In: Avidan, S., Brostow, G.J., Cissé, M., Farinella, G.M., Hassner, T. (eds.) Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVII. Lecture Notes in Computer Science, pp. 409–424. Springer, Tel Aviv, Israel (2022). https://doi.org/10.1007/978-3-031-19812-0_24 . https://doi.org/10.1007/978-3-031-19812-0_24
- Zellers, R., Yatskar, M., Thomson, S., Choi, Y.: Neural motifs: Scene graph parsing with global context. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18–22, 2018, pp. 5831–5840. Computer Vision Foundation / IEEE Computer Society, Salt Lake City, UT, USA (2018). <https://doi.org/10.1109/CVPR.2018.00611> . <https://doi.org/10.1109/CVPR.2018.00611>
- Zeng, P., Zhang, H., Gao, L., Song, J., Shen, H.T.: Video question answering with prior knowledge and object-sensitive learning. IEEE Trans. Image Process. **31**, 5936–5948 (2022) <https://doi.org/10.1109/TIP.2022.3205212>
- Zeng, P., Zhang, H., Song, J., Gao, L.: S2 transformer for image captioning. In: Raedt, L.D. (ed.) Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23–29 July 2022, pp. 1608–1614. ijcai.org, Vienna, Austria (2022). <https://doi.org/10.24963/IJCAI.2022/224> . <https://doi.org/10.24963/IJCAI.2022/224>