# Relation Boosting for Unbiased PSG with Vision-Language Model

IW-FCV2025 Submission No.10

**Abstract.** Panoptic Scene Graph Generation (PSG) is a recently introduced task in image scene understanding, which aims to segment images and extract subject-object-relation triplets to construct a comprehensive scene graph. Current one-stage methods for PSG face challenges in predicting low-frequency relationships due to semantic ambiguity. Additionally, the prediction mechanisms inherent to Transformer contribute to instability in learning for segmentation tasks. In this study, we introduce a novel PSG approach, **R**elation **B**oosting (RB) with plug-in Relation Boosting Module (RBM) and Relation Boosting Training strategy (RBT) to improve PSG models effectively. Specifically, our RBM extracts high-level features from vision-language foundation models and reinforces the decoding process from visual and linguistic perspectives. RBT strategy leverages ground-truth label and mask information to guide the decoding and significantly reduce training time. Our results demonstrate the effectiveness of incorporating vision-language information and mitigating the semantic ambiguity issue in the one-stage PSG model.

**Keywords:** Panoptic Scene Graph Generation · Vison Language Model.

## 1 Introduction

Panoptic Scene Graph Generation (PSG) extends traditional Scene Graph Generation (SGG) by incorporating panoptic segmentation to capture a richer and more detailed representation of images. In PSG, both foreground and background object classes are included, providing a comprehensive depiction of scenes. PSG serves as a bridge between vision and language, enabling various downstream applications such as visual question answering, image captioning, and visual reasoning.

Based on current PSG methods and available datasets [19], two primary challenges need to be addressed. First is semantic ambiguity, we categorize predicate classes in the dataset samples into general(e.g. on, beside) and specific(e.g. throwing, looking at). Similar to the Visual Genome (VG) dataset [4], annotators tend to favor general classes over specific ones. This preference leads to inconsistent predicate annotations for the same object pairs in single-label annotation tasks, making the model prone to favoring general classes during prediction. For instance, as shown in Figure 1, these are the top three relationships predicted by biased model [19] for person and frisbee. Due to the significantly larger annotation frequency of the general class <beside> compared to the other two predicates, biased models often predict this relationship—even when it is




**Fig. 1.** Semantic ambiguity of the same object pair. The text marked in red is ground-truth

not contextually accurate. Second is training instability, current one-stage PSG models are predominantly based on Transformer modules, such as DETR [1] and Mask2Former [3]. These models rely on queries to predict both the mask and class of targets, where the predicted masks are constructed from scratch through dot-product query embeddings and feature maps. Since the masks are reconstructed from scratch at each layer, the masks predicted by queries across different layers will lead to instability during training. Corresponding to the issues mentioned above, the contributions of this paper are as follows:

- We address the semantic ambiguity problem in scene graphs and introduce Relation Boosting Modul (RBM), which enhances high-level features by leveraging visual knowledge from a vision-language model [5] and incorporates language knowledge with visual information to guide triplet learning during training.
- We propose Relation Boosting Training (RBT), a novel training strategy that addresses query instability during the training process. By leveraging ground-truth data as a supervisory signal, RBT guides the query learning process more effectively. Our approach significantly reduces training time while preserving model accuracy.

## 2 Related Work

### 2.1 Panoptic Scene Graph Generation

Panoptic Scene Graph Generation (PSG) [19] has emerged as a high-level task to advance scene understanding by integrating object relationships and panoptic segmentation. PSG builds upon the foundation of Scene Graph Generation (SGG), which typically represents objects with bounding boxes. In contrast, PSG leverages panoptic segmentation to provide a more detailed and holistic

representation of scenes. The segmentation targets include object instances and stuff classes—amorphous regions with richness in the generated scene graphs.

**Traditional approaches** Yang *et al.* [19] utilize some two-stage [14, 17, 4, 10] and one-stage [19] baseline methods. Since the baseline methods were simple extensions of SGG and object detection approaches, they were not well-suited for the PSG task. Building upon prior work, CATQ [18] introduces a specialized pipeline for PSG, which employs three dedicated queries to decode individual elements of the triplet target. Within this framework, the Subject-Object Attention Guide (SOAG) module enhances relation learning by incorporating spatial object information as prior knowledge. Then, the subject-object-relation information from three kinds of queries is aggregated into a unified contextual representation. Due to its well-reasoned design and high efficacy, we adopted CATQ as the base architecture.

**Unbiased approaches** The current PSG methods for debiasing include data augmentation and loss weight redistribution strategies. HiLo [23] addresses the long-tail problem by specializing in different network branches for learning high-frequency and low-frequency relations. It effectively extracts high and low-frequency features from two stream data sets, improving the tail relationship accuracy without compromising the head relationship accuracy. Data augmentation methods have shown excellent results. Nevertheless, they require extensive and complex manual preprocessing and postprocessing, compromising their generalizability. PairNet [16] introduces a novel framework that uses a pair proposal network to filter object pairs. Additionally, it employs SeeSaw [15] as a relational mechanism to dynamically adjust the gradients of positive and negative samples used in relation classification. Moreover, the unbiased loss or reallocation method negatively impacts the prediction of common high-frequency relationships and the precision of recall.

### 2.2 Vision-Language Models for triplet detection

Recent advances in large multimodal models [13, 6, 5] have transformed diverse modality inputs into a unified representational space, facilitating more comprehensive information extraction and understanding. This advancements are also shown in tasks such as Scene Graph Generation (SGG) [7, 20] and Human-Object Interaction (HOI) [8, 11, 12], where vision-language models have demonstrated notable effectiveness. Despite these advances, current approaches only partially utilize multimodal features and do not fully exploit the rich knowledge encoded within large language models to enhance triplet detection capabilities. Previous methods [8, 20, 7] primarily rely on triplet class labels as prompts to extract linguistic features. However, these approaches are still hard to deal with the semantic ambiguity among triplets. Without proper visual context, models tend to predict general relation terms rather than specific classes. Consequently, recent methods typically offer only limited improvements.

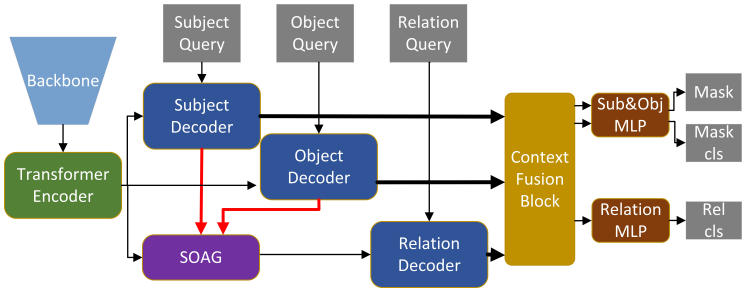


**Fig. 2.** CATQ framework [18]

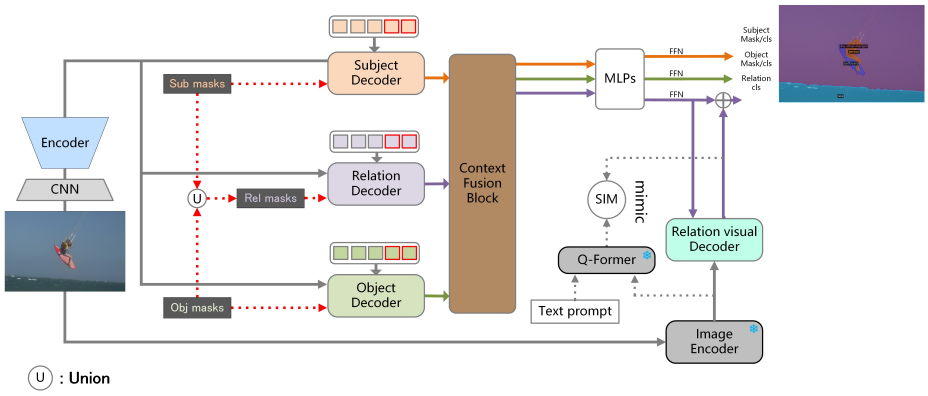


**Fig. 3.** The pipeline of our proposed method. The dashed lines indicate components that are reused during inference.

## 3 Approach

### 3.1 Baseline Model

Figure 2 shows our base model CATQ. Context Fusion Block is a multi-information fusion module designed to facilitate information interaction for the triplet query. For more details. We propose the RB model as illustrated in Figure 3. Although the Subject-Object Attention Guide (SOAG) module enhances the efficiency and accuracy of visual information processing during relation decoding, our analysis reveals that it inadvertently limits the contextual richness of relation queries. Given the constraints of limited image resources in our work, we determined that removing the SOAG module would optimize information utilization.

### 3.2 Plug-in Relation Boosting Module

The semantic ambiguity challenges stem from annotation inconsistencies in the PSG dataset and insufficient data volume. To address these limitations, RBM enhances triplet representations by extracting high-level vision-language features from large vision-language models. This approach guides the baseline model toward a more nuanced understanding of complex semantic relationships.

**Visual Knowledge.** BLIP2 [5] proposes a novel framework that integrates image features, extracted by a dedicated visual encoder, with text embeddings processed through a large language model, enabling advanced vision-language understanding. We adopt the prior knowledge from BLIP2 as the core component of our framework.

First, the image encoder converts the image $x$ into a visual feature map $x^{'} \in \mathbb{R}^{H\times W\times 3}$. Then, high-level visual features $\mathbf{V}_q \in \mathbb{R}^{N_q\times D_q}$ are extracted from the image encoder of BLIP2, where $N_q$ denotes the number of queries in the Q-former and $D_q$ is the output dimension of the Q-former. The image encoder, trained on a more diverse and extensive dataset than traditional visual detectors, demonstrates enhanced capabilities for comprehensive image understanding. In our case containing object pair location and relationship information are employed to align the higher-level semantics of the vision-language model with spatial and complex semantic relationships between object pairs. Additionally, we design a Relation Visual Decoder $\psi$, which transports relationship information to the feature tokens from the vision-language model. This process is defined as shown in Eq.1.

$$\mathbf{Q}_b^R = \Psi(\mathbf{V}_q, F(\mathbf{Q}^R)) \tag{1}$$

To align the dimensions from the Transformer to BLIP2, we employ a linear layer $F$. The query $Q^R$ represents the relationship query from the base model, $Q_b^R$ is enriched with visual relational information.

Regarding the design of $\psi$, we use a different decoder architecture that diverges from Mask2former [3] approach. Instead, we first perform self-attention computation with a 6-layer architecture. This design choice stems from our intention to maintain the focus on relation prompts. Computing cross-attention first would create an over-reliance on key quality, potentially leading to attention dispersion across non-relevant spatial regions.

**Text Guidance.** Language models enable the effective transfer of linguistic knowledge to visual architectures, addressing a critical limitation in relationship understanding. Only using visual prompts is insufficient for learning nuanced relationship definitions due to their inherent ambiguity. Language models enhance semantic comprehension of relationships by providing rich linguistic context. This enhancement is particularly valuable for datasets where semantic ambiguity has traditionally impaired model performance, as the integration of language model knowledge significantly improves relationship classification accuracy.

However, exclusive reliance on linguistic knowledge presents challenges, as relationship definitions can vary significantly across different visual contexts, involving diverse object pairs with distinct spatial configurations. For instance, the relationships between a person and a chair can manifest in multiple forms, such as "sitting on" or "standing by" depending on the spatial context. To address this complexity, our approach leverages multi-modal features from vision-language models, enabling more contextually appropriate integration of linguistic knowledge.

The detailed procedure is outlined as follows: the realtion features $\mathbf{Q}_b^R$ is aligned with the dimension $D_q$ of BLIP2 [5] through a linear layer. The ground-truth triplet text $T_{tri}$ and image $x$ for each batch are input into the Q-former of BLIP2, producing visual query features $\mathbf{Q}_\phi \in \mathbb{R}^{N_q \times D_q}$ enriched with linguistic knowledge. Next, the cosine similarity matrix is computed between the context triplet features $\mathbf{Q}_b^R$ and $\mathbf{Q}_\phi$, from which matched features are obtained based on the optimal matching index. Finally, the $L_1$ loss between the prediction and the target is calculated.

This process is defined as shown in Eq.2,3

$$\mathbf{Q}_\phi = \text{Q-Former}(x, T_{tri}) \tag{2}$$

$$\mathcal{L}_{match} = L_1(\mathbf{Q}_{\phi_j}, \mathbf{Q}_b^R), \qquad j = \arg\max \mathrm{S_C}(\mathbf{Q}_b^R, \mathbf{Q}_\phi) \tag{3}$$

### 3.3 Relation Boosting Training Stratety

Mask2Former [3] has emerged as the predominant detector for universal segmentation tasks, with widespread adoption across current PSG models [23, 18, 16]. However, Mask2Former exhibits performance limitations due to query prediction instability, where predictions for identical queries can vary substantially across different layers, resulting in inefficient training. To address this limitation, we propose the Relation Boosting Training (RBT) strategy, inspired by [22]. Our approach comprises two principal components: label-lead training and mask-lead training. The pipeline of RBT is illustrated by Figure 4.

**Mask-lead training.** Leveraging the decoder architecture of Mask2Former [3], the object ground-truth masks are used as cross-attention masks in the input. For the attention mask in relation decoding, we apply the union of the subject and object masks as the relation cross-attention mask as Eq.4.

$$\text{Mask}_{rel} = \text{Mask}_{sub} \cup \text{Mask}_{obj} \tag{4}$$

**Label-lead training.** We randomly flip the class labels of the Ground-Truth (GT) and encode them as the noised query. Unlike MPformer [22], we do not use all the labels from the dataset as possible candidates. Instead, we sample from the triplet ground-truth within each image.

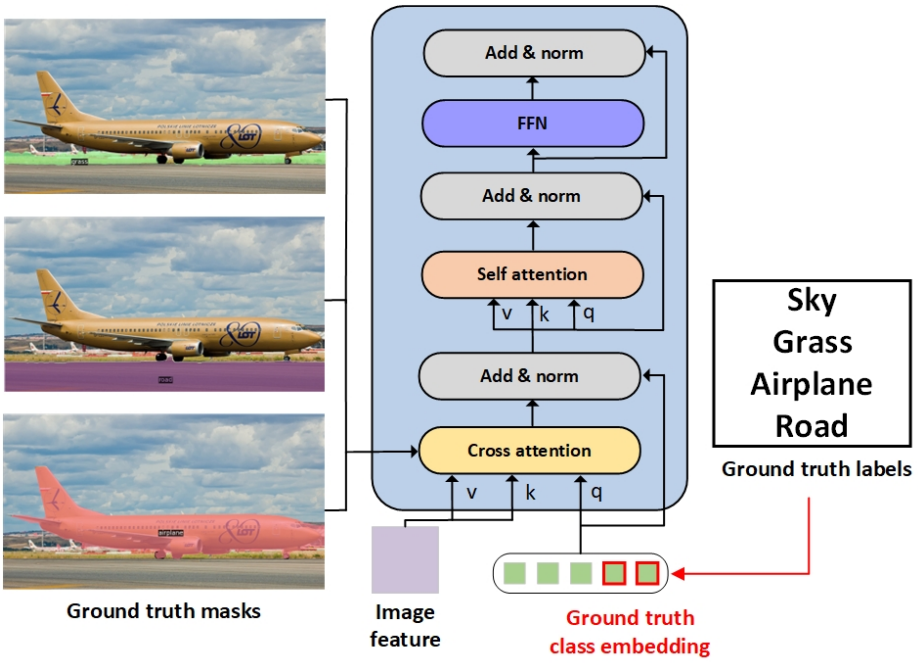


**Fig. 4.** An illustration of Relation Boosting Training Stratety

### 3.4 Training and Inference

Training loss is thus:

$$\mathcal{L} = \mathcal{L}_{seg} + \mathcal{L}_{cls} + \lambda \mathcal{L}_{match}$$

$\mathcal{L}_{seg}$ and $\mathcal{L}_{cls}$ represent the segmentation and classification losses, respectively. $\mathcal{L}_{match}$ refers to the text-guided loss, and $\lambda$ is the hyperparameter for the text-guided loss. Each triplet is transformed into a sentence for processing by the Q-former. For example, "subject, relation, object" is converted to "The <subject> is <relation> the <object>." The inference does not involve text guidance and relies solely on visual knowledge.

## 4 Experiment

We show experimental analysis and do ablation studies on OpenPSG [19] with CATQ [18] as the baseline model.

### 4.1 Dataset

The OpenPSG dataset comprises 46,572 training images and 2,177 test images, totaling 48,749 images with panoptic segmentation and scene graph annotations. The object categories include 133 classes (80 foreground and 53 background), while the relationship categories consist of 56 classes.

**Table 1.** Performance of our method and other state-of-the-art models

| Method | Backbone | *R@20* | *R@50* | *R@100* | *mR@20* | *mR@50* | *mR@100* | *PQ* |
|---|---|---|---|---|---|---|---|---|
| **Two-stage** | | | | | | | | |
| IMP [17] | ResNet-50 | 16.5 | 18.2 | 18.6 | 6.5 | 7.1 | 7.2 | 40.2 |
| MOTIFS [21] | ResNet-50 | 20.0 | 21.7 | 22.0 | 9.1 | 9.6 | 9.7 | 40.2 |
| VCTree [14] | ResNet-50 | 20.6 | 22.1 | 22.5 | 9.7 | 10.2 | 10.2 | 40.2 |
| GPSNet [10] | ResNet-50 | 17.8 | 19.6 | 20.1 | 7.0 | 7.5 | 7.7 | 40.2 |
| **One-stage** | | | | | | | | |
| PSGTR [19] | ResNet-50 | 28.4 | 34.4 | 36.3 | 16.6 | 20.8 | 22.1 | 13.9 |
| PSGFormer [19] | ResNet-50 | 18.0 | 19.6 | 20.1 | 14.8 | 17.0 | 17.6 | 36.8 |
| PairNet [16] | ResNet-50 | 29.6 | 35.6 | 39.6 | **24.7** | 28.5 | 30.6 | - |
| HiLo [23] | ResNet-50 | <u>34.1</u> | **40.7** | **43.0** | 23.7 | <u>30.3</u> | **33.1** | 41.6 |
| CATQ [18] | ResNet-50 | **34.8** | 39.7 | 40.3 | 20.9 | 24.9 | 25.2 | 35.9 |
| CATQ-RB(ours) | ResNet-50 | **34.8** | <u>40.4</u> | <u>41.2</u> | <u>23.8</u> | **30.8** | <u>32.1</u> | 35.9 |

### 4.2 Tasks and Metrics

Given an input image, the model simultaneously predicts object segmentation and pairwise relationships between instances. The threshold for a correct mask is set to an IoU of 0.5, and a correct match is defined when all three elements <subject, relation, object> are correctly classified. Recall@K [14] and mRecall@K [2] are used as evaluation metrics for scene graph generation, while PQ is employed as the evaluation metric for segmentation. mR@K is the primary metric of interest for our study, it provides a more equitable evaluation of model performance, as it treats each predicate category equally and does not disproportionately prioritize "general" predicates based on the number of samples.

### 4.3 Implementation Details

The image encoder of BLIP2 employs ViT-L/14. The dimension of the Transformer is set to $D = 256$, with $N^q = 100$ queries. We utilize the AdamW optimizer with a learning rate of 1e-4 (1e-5 for backbone). The image encoder and three decoders are initialized from Mask2former, which is pre-trained on COCO[9]. The setup includes 8 NVIDIA A6000 GPUs with a batch size of 8 (1 image per GPU). We train our model for 20 epochs, which is two-thirds of CATQ.

### 4.4 Comparison to the state-of-the-art

Table 1 presents the performance of our method in comparison to state-of-the-art approaches on the PSG dataset. Compared to CATQ [18], our method demonstrates improvements across all metrics, with particularly substantial gains of 13.9%, 23.7%, and 27.4% in the mR@K metric. Based on it, our model achieves convergence in 20 epochs, representing a significant reduction in training time compared to CATQ's requirement of 30 epochs. Using the same backbone and without introducing additional training data or post-processing steps, CATQ-RB demonstrates superior performance compared to the current state-of-the-art

HiLo [23], achieving gains of 0.1 and 0.5 points in mR@20 and mR@50 metrics, respectively.

### 4.5 Ablation Studies

**Table 2.** Ablation comparison of Visual prompt module.

| Method | *mRecall@20* | *mRecall@50* | *mRecall@100* |
|---|---|---|---|
| Relation | **22.5** | **30.0** | **31.0** |
| Object pair (1) | 20.9 | 26.52 | 28.44 |
| Object pair (2) | 20.0 | 24.1 | 25.8 |

**Table 3.** Ablation comparison of Text Guidance module.

| Method | *mRecall@20* | *mRecall@50* | *mRecall@100* |
|---|---|---|---|
| visual&text | **22.5** | **30.0** | **31.0** |
| text | 20.4 | 24.1 | 24.4 |
| w/o text | 21.7 | 26.8 | 27.0 |

**Table 4.** Ablation comparison of the RBT.

| *RBM* | *RBT* | *mR@20* | *mR@50* | *mR@100* |
|---|---|---|---|---|
| ✓ | ✓ | 23.8 | 30.8 | 32.1 |
| ✓ | - | 22.5 | 30.0 | 31.0 |
| - | - | 20.9 | 24.9 | 25.2 |

**Table 5.** Ablation comparison of Text Guidance loss hyperparameter.

| $\lambda$ | *mRecall@20* | *mRecall@50* | *mRecall@100* |
|---|---|---|---|
| 1 | 20.4 | 24.1 | 24.4 |
| 10 | 20.7 | 27.2 | 28.1 |
| 15 | 22.5 | **30.0** | **31.0** |
| 20 | **23.4** | 29.0 | 29.5 |

**Visual Prompt.** Our approach utilizes contextually-enriched relationship queries as prompts for the Q-former architecture. Additionally, we investigated an alternative configuration where only object pair information is presented to the image encoder. As shown in Table 2, there are two prompting methods. Object Pair (1) adds context information to the object pair query and divides by two, as represented in $(\mathbf{Q}^{S}_{context} + \mathbf{Q}^{O}_{context})/2$. Object Pair (2) applies a concatenation operation to the object pair query with context information, as described in $concat(\mathbf{Q}^{S}_{context}, \mathbf{Q}^{O}_{context})$. Ultimately, the results demonstrate that the vision-language model can also comprehend relationship information.

**Text Guidance.** Table 3 presents a comparative analysis of model performance using multi-modal features (visual & text) versus text-only features. Our

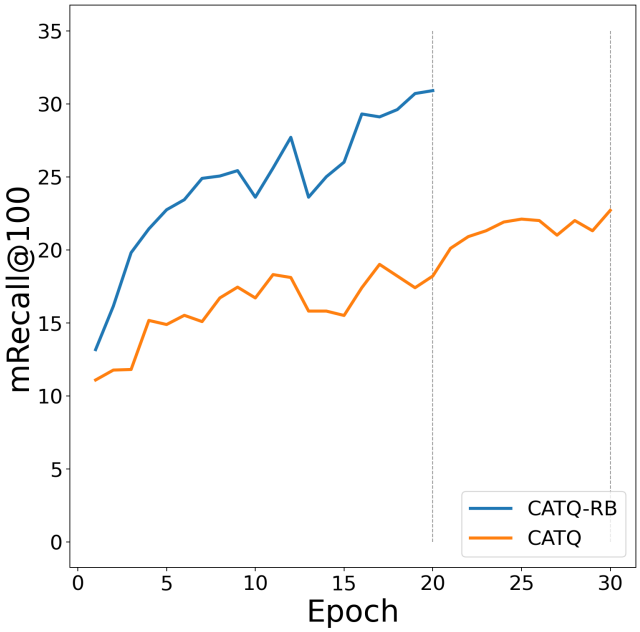


**Fig. 5.** Comparison of accuracy during traning

results demonstrate that the integration of visual information enhances the utilization of linguistic features, yielding superior accuracy compared to the baseline CATQ model. Notably, the exclusive use of textual information fails to improve performance and degrades model effectiveness. Furthermore, while utilizing solely visual features (w/o text) shows marginally lower accuracy than the multi-modal approach, it still outperforms the text-only configuration. Through the adjustment of text guidance hyperparameters (Table 5), we found that appropriately increasing the constraints on the language information can improve the model's precision. However, either over-constraining or under-constraining the language can introduce negative biases in the training process.

We extended our empirical evaluation to include CLIP [13], a vision-language model trained using contrastive image-text pairs. While BLIP2 exhibits inherent advantages for PSG tasks due to its pre-training on VQA datasets, our experiments with CLIP still show substantial improvements in accuracy (Table 6). These results demonstrate the robustness of our RBM approach across diverse vision-language architectures.

We compare the performance of adding only the RBM to CATQ as presented in Table 4. We observe improvements of 7.7%, 20.5%, and 23.0% in the mR@K metrics, respectively. This indicates that the RBM effectively leverages the information from the vision-language model and enhances the semantic representation of relationships.

**Table 6.** Results of CATQ-RB equipped with different Vision-language models

| Method | mRecall@20 | mRecall@50 | mRecall@100 |
|---|---|---|---|
| CATQ-RB(w/ CLIP) | **22.5**(+1.6) | 26.4(+1.5) | 27.2(+2.0) |
| CATQ-RB(w/ BLIP2) | **22.5**(+1.6) | **30.0**(+5.1) | **31.0**(+5.8) |

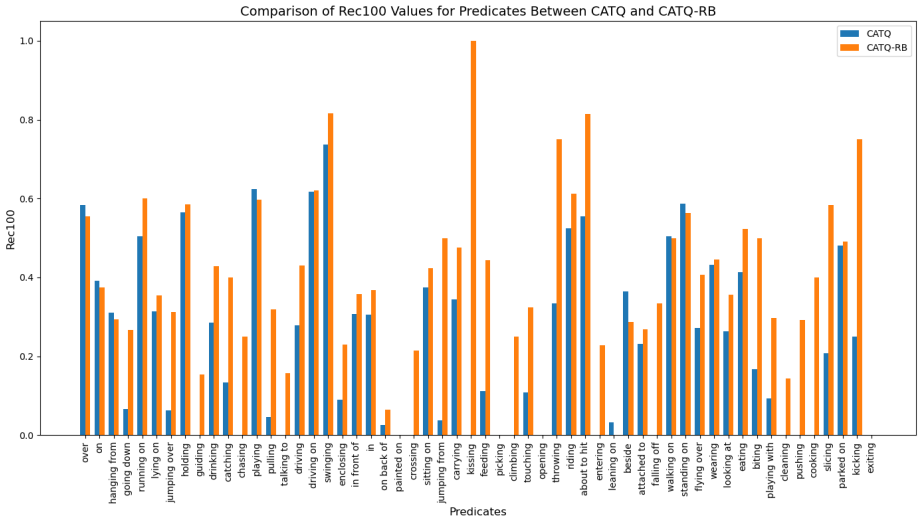


**Fig. 6.** Compairson of Rec100 for Predicates Between CATQ and CATQ-RB

**RBT.** The addition of the RBT strategy is primarily aimed at increasing the training efficiency of the Transformer model. As shown in Figure 5, RBT significantly reduces training time compared to the original model CATQ (reduction from 30 to 20 epochs). Moreover, the RBT strategy does not compromise model accuracy, instead, it results in a slight improvement after the strategy is added, as demonstrated in Table 4.

**Semantic ambiguity.** The primary challenge addressed in this study concerns the ambiguity of relational semantics. To demonstrate the improvements brought by the RB model, we conducted a comparative analysis of recall performance between CATQ and CATQ-RB across all relationship terms. The result is illustrated in Figure 6. Our findings reveal that RB demonstrates notable improvements in recognizing semantically rich but specific words (e.g., throwing, biting, kicking). Moreover, RB successfully comprehends and identifies relational terms that showed zero precision in CATQ. However, we observed a slight decrease in precision for general terms (e.g., on, over, beside). This decline effectively demonstrates that the model bias caused by annotation gaps has been successfully corrected.

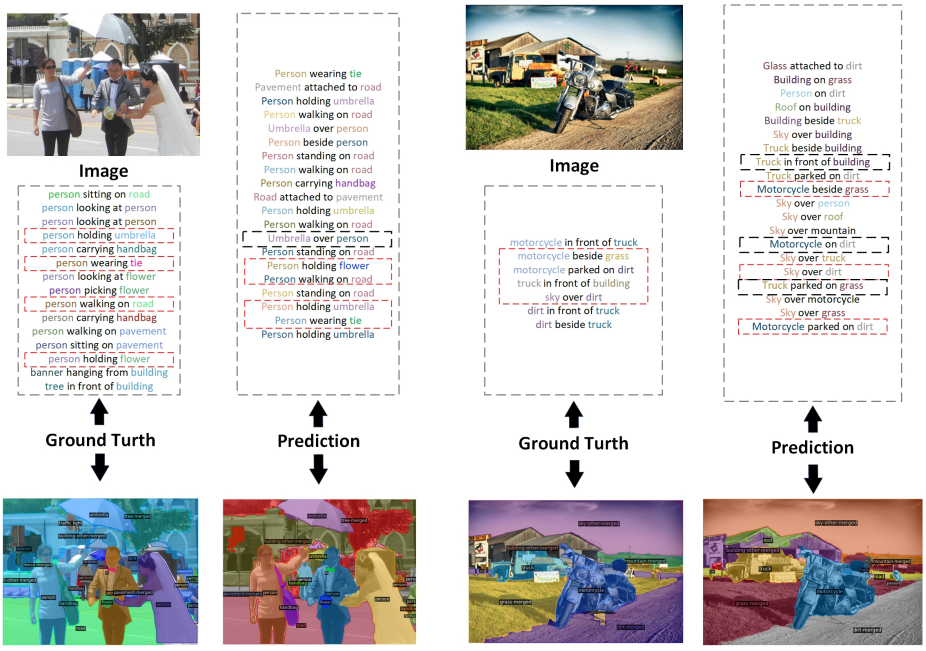


**Fig. 7.** Visualization of panoptic segmentations and the top 20 predicted triplets compared with the ground-truth of CATQ-RB. The red dashed box represents the correctly predicted triples. In the presented example on the left, within a complex scene, specific objects like the traffic light, banner, and tent are not detected. Conversely, in the comparatively simpler scene demonstrated on the right, our approach showcases the ability to identify instances not included in the ground-truth annotations, such as the roof and road.

## 5 Conclusion

In this work, we propose the plug-in RBM and the RBT training strategy. RBM leverages high-level information from vision-language models to guide the learning of existing PSG models, helping to mitigate biased relationship predictions caused by limited data. The RBT training strategy addresses query instability in transformer-based models by introducing ground-truth, guiding the predictions of each layer's queries. We evaluate the effectiveness of our RB approach on the CATQ model. Experimental results demonstrate that our method achieves significant improvements on the PSG dataset and alleviates the semantic ambiguity problem in relationships. In the future, we aim to extend the RB method and apply it to other transformer-based methods.

## References

1. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European conference on computer

vision. pp. 213–229. Springer (2020)
2. Chen, T., Yu, W., Chen, R., Lin, L.: Knowledge-embedded routing network for scene graph generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6163–6171 (2019)
3. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1290–1299 (2022)
4. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., et al.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. International Journal of Computer Vision **123**, 32–73 (2017)
5. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: International Conference on Machine Learning. pp. 19730–19742. PMLR (2023)
6. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International Conference on Machine Learning. pp. 12888–12900. PMLR (2022)
7. Li, R., Zhang, S., Lin, D., Chen, K., He, X.: From pixels to graphs: Open-vocabulary scene graph generation with vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 28076–28086 (2024)
8. Liao, Y., Zhang, A., Lu, M., Wang, Y., Li, X., Liu, S.: Gen-vlkt: Simplify association and enhance interaction understanding for hoi detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20123–20132 (2022)
9. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. pp. 740–755. Springer (2014)
10. Lin, X., Ding, C., Zeng, J., Tao, D.: Gps-net: Graph property sensing network for scene graph generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3746–3753 (2020)
11. Mao, Y., Deng, J., Zhou, W., Li, L., Fang, Y., Li, H.: Clip4hoi: towards adapting clip for practical zero-shot hoi detection. Advances in Neural Information Processing Systems **36**, 45895–45906 (2023)
12. Ning, S., Qiu, L., Liu, Y., He, X.: Hoiclip: Efficient knowledge transfer for hoi detection with vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23507–23517 (2023)
13. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning. pp. 8748–8763. PMLR (2021)
14. Tang, K., Zhang, H., Wu, B., Luo, W., Liu, W.: Learning to compose dynamic tree structures for visual contexts. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6619–6628 (2019)
15. Wang, J., Zhang, W., Zang, Y., Cao, Y., Pang, J., Gong, T., Chen, K., Liu, Z., Loy, C.C., Lin, D.: Seesaw loss for long-tailed instance segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9695–9704 (2021)

16. Wang, J., Wen, Z., Li, X., Guo, Z., Yang, J., Liu, Z.: Pair then relation: Pair-net for panoptic scene graph generation. IEEE Transactions on Pattern Analysis and Machine Intelligence (2024)
17. Xu, D., Zhu, Y., Choy, C.B., Fei-Fei, L.: Scene graph generation by iterative message passing. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5410–5419 (2017)
18. Xu, J., Chen, J., Yanai, K.: Contextual associated triplet queries for panoptic scene graph generation. In: Proceedings of the 5th ACM International Conference on Multimedia in Asia. pp. 1–5 (2023)
19. Yang, J., Ang, Y.Z., Guo, Z., Zhou, K., Zhang, W., Liu, Z.: Panoptic scene graph generation. In: European Conference on Computer Vision. pp. 178–196. Springer (2022)
20. Yu, Q., Li, J., Wu, Y., Tang, S., Ji, W., Zhuang, Y.: Visually-prompted language model for fine-grained scene graph generation in an open world. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 21560–21571 (2023)
21. Zellers, R., Yatskar, M., Thomson, S., Choi, Y.: Neural motifs: Scene graph parsing with global context. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5831–5840 (2018)
22. Zhang, H., Li, F., Xu, H., Huang, S., Liu, S., Ni, L.M., Zhang, L.: Mp-former: Mask-piloted transformer for image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18074–18083 (2023)
23. Zhou, Z., Shi, M., Caesar, H.: Hilo: Exploiting high low frequency relations for unbiased panoptic scene graph generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 21637–21648 (2023)