This ICCV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Visually-Prompted Language Model for Fine-Grained Scene Graph Generation in an Open World

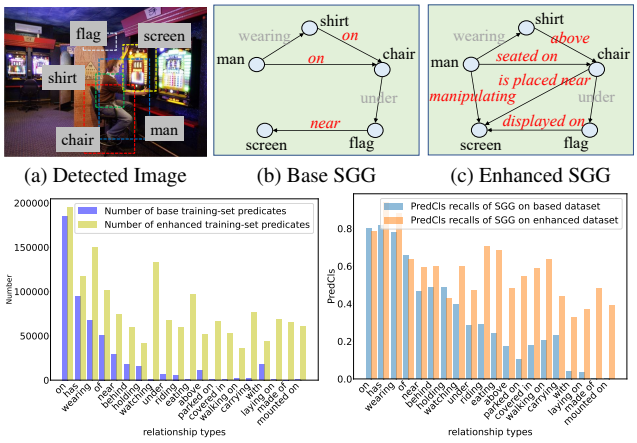
Qifan Yu[1] Juncheng Li[1†] Yu Wu[2] Siliang Tang[1] Wei Ji[3] Yueting Zhuang[1]
[1]Zhejiang University, [2]Wuhan University, [3]National University of Singapore
{yuqifan, junchengli, siliang, yzhuang}@zju.edu.cn
yu.wu-3@student.uts.edu.au, jiwei@nus.edu.sg

Abstract

*Scene **G**raph **G**eneration (SGG) aims to extract <subject, predicate, object> relationships in images for vision understanding. Although recent works have made steady progress on SGG, they still suffer long-tail distribution issues that tail-predicates are more costly to train and hard to distinguish due to a small amount of annotated data compared to frequent predicates. Existing re-balancing strategies try to handle it via prior rules but are still confined to pre-defined conditions, which are not scalable for various models and datasets. In this paper, we propose a **C**ross-mod**a**l predi**Ca**te b**o**osting (**CaCao**) framework, where a visually-prompted language model is learned to generate diverse fine-grained predicates in a low-resource way. The proposed CaCao can be applied in a plug-and-play fashion and automatically strengthen existing SGG to tackle the long-tailed problem. Based on that, we further introduce a novel **E**ntangled cross-modal **p**rompt approach for open-world pred**i**cate s**c**ene graph generation (**Epic**), where models can generalize to unseen predicates in a zero-shot manner. Comprehensive experiments on three benchmark datasets show that CaCao consistently boosts the performance of multiple scene graph generation models in a model-agnostic way. Moreover, our Epic achieves competitive performance on open-world predicate prediction. The data and code for this paper are publicly available.*[1]

1. Introduction

Scene graph generation (SGG) aims to detect visual relationships in real-world images, which consist of the subject, predicate, and object (*i.e.,* **subject**: *flag*, **predicate**: *displayed on*, **object**: *screen* in Figure 1 (a)). Since scene graphs bridge the gap between raw pixels and high-level visual semantics, SGG has been widely used in a variety of visual scene analysis and understanding tasks [3, 29, 31, 30], such as visual question answering [20, 15], image captioning [63, 54], and 3D scene understanding [14, 61].



(a) Detected Image (b) Base SGG (c) Enhanced SGG

(d) Long-tail predicate distribution (e) PredCls of different predicates

Figure 1. **Illustration of handling long-tail distribution problem by cross-modal predicate boosting in Visual Genome.** (b) and (c) show scene graphs enhanced by visual knowledge generating more informative predicates in long-tail distribution. (d) indicates the imbalance of predicates due to the long-tailed distribution in the training set. (e) For prediction of scene graph relationships (PredCls), our CaCao framework can obtain consistent improvement on both head predicates and tail predicates.

Recently, various methods [4, 59, 58, 48, 55, 53, 7] have been proposed to improve the SGG performance, but still tend to predict frequent but uninformative predicates due to the long-tailed distribution of predicates in SGG datasets [52, 33, 56]. In a way, those approaches degenerate into a trivial solution, which undermines the application of SGG. As shown in Figure 1 (d), in the Visual Genome [52], the top 20% of predicate categories account for almost 90% of samples, while other tail fine-grained predicates lack sufficient training data. Accordingly, the PredCls recalls of SGG models on those tail predicates are remarkably lower than head predicates, as demonstrated in Figure 1 (e).

†Corresponding Authors.
[1]https://github.com/Yuqifan1117/CaCao

Prior works have been proposed in recent years to alleviate the bias caused by the long-tail distribution based on causal rules [47, 32], reweighting [48, 51, 56] and resampling strategy [2, 53, 33] gradually. Nevertheless, these methods still require careful tuning of additional hyper-parameters, such as sampling frequency and category weight. They are sensitive to different architectures and data distributions, which are not flexible for real-world situations. Another alternative way is to increase the number of tail predicates in training. IETrans [60] uses internal relation correlation to enhance the existing dataset. However, these methods rely on the prior distribution of source data and only work in specific pre-defined conditions. Such a manner based on hand-designed rules covers only limited categories, which is time-consuming and unscalable.

In this paper, we propose a **C**ross-mod**a**l predi**Ca**te b**o**osting (**CaCao**) framework, which leverages the extensive knowledge from the pre-trained language models to enrich the tail predicates of scene graphs in a low-cost and easily scalable way. Our fundamental intuition is that language models gain extensive knowledge about informative relationships from massive text corpus during general sentence pre-training (*i.e. Large silver airplane **parked outside** an airport with a pilot **sitting in** it that has **come back from** a mission, while the pilot gets some rest.*) [44, 46]. While the pre-trained language models contain diverse relational knowledge, it is non-trivial to elicit this knowledge from them to scene graph generation. First, there is a significant modality gap in migrating extensive linguistic knowledge into scene graph predicate prediction since such large-scale language models are 'blind' to visual regions. An alternative way is to use vision-language pre-training (VLP) models. However, VLP models are mainly trained by image-text contrastive learning, lacking the delicate language ability to generate fine-grained predicate category words. Second, a predicate type might correspond to many different linguistic expressions (*e.g.*, he "walks through" / "is passing through" / "passed by" a street may correspond to the same predicate). Without considering such semantic co-reference phenomenon, the adapted language model for predicate generation can easily collapse to monotonic predictions.

To address the above challenges, we first introduce a novel cross-modal prompt tuning approach, which enables the language model to subtly capture visual context and predict informative predicates as masked language modeling, called the visually-prompted language model. As for semantic co-reference, we further present an adaptive semantic cluster loss for prompt tuning, which models the semantic structures of diverse predicate expressions and adaptively adjusts the distribution to inhibit excessive enhancement of specific predicates during boosting process, thus rendering a diverse and balanced distribution. Moreover, we introduce a fine-grained predicate-boosting strategy to extend the existing dataset with the informative predicates generated by our visually prompted language model. From the comprehensive view of Figure 1 (e), our CaCao can greatly improve the SOTA models' performance in a plug-and-play way, where **PredCls** of most predicates are consistently increased by 30% in the purple bar than the blue.

From a more general perspective, our CaCao can not only effectively alleviate the long-tail distribution problem even in large-scale SGG but also generalize to open-world predicates by leveraging the generalizability of human language. Inspired by the impressive zero-shot performance of vision-language pre-training models [42, 26, 22], which utilize the generalizability of human language for zero-shot transfer, we replace the traditional fixed predicate classification layer with category-name embedding and use the diverse predicates generated by our CaCao to learn general and transferable predicate embeddings. Specifically, we propose a novel **E**ntangled cross-modal **p**rompt approach for open-world pred**i**cate s**c**ene graph generation (**Epic**), where the entangled cross-modal prompt alternately tinkers with the predicate representation, making the scene graph model aware of the abstract interactive semantics.

Surprisingly, without using any ground-truth annotations and only with the informative relations generated by our CaCao framework, our Epic achieves competitive performance on the open-world predicate learning problem.

Our main contributions are summarized as follows:

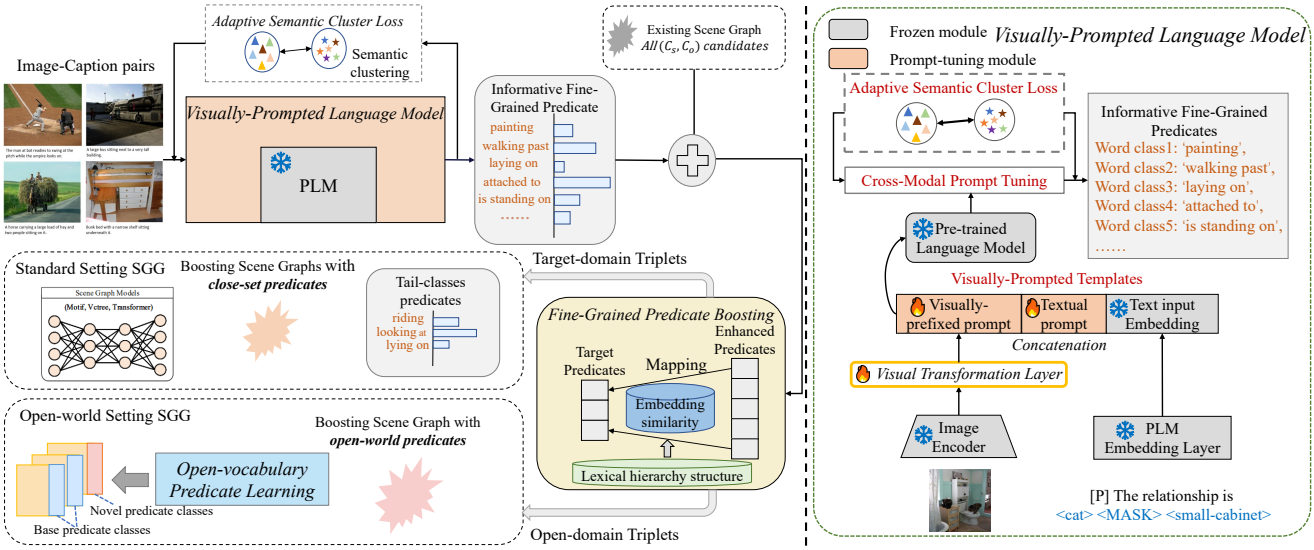
- We propose a novel **C**ross-mod**a**l predi**Ca**te b**o**osting (**CaCao**) framework, where a visually-prompted language model is learned to enrich the existing dataset with fine-grained predicates in a low-resource and scalable way.

- Our CaCao can be applied to SOTA models in a plug-and-play fashion. Experiments over three datasets show steady improvement in standard SGG tasks, demonstrating a promising direction to automatically boosting data by large-scale pre-trained language models rather than time-consuming manual annotation.

- In addition, we introduce **E**ntangled cross-modal **p**rompt approach for open-world pred**i**cate s**c**ene graph generation (**Epic**) to explore the expansibility of CaCao for unseen predicates, and validate its effectiveness with comprehensive experiments.

2. Related Work

Scene Graph Generation. Current scene graph generation is still far from practical since it suffers from long-tail distribution of predicates [59, 48, 4]. Recently, resampling [2, 33] and reweighting [51, 56] and causal rule-based



(a) The overall framework of CaCao

(b) Visually-Prompted Language Model

Figure 2. Illustration of our proposed Cross-Modal Predicate Boosting framework. **Visually-Prompted Language Model** is designed to exploit linguistic knowledge from pre-trained language models and migrate it into scene understanding via visual cues. The right subfigure shows the detail of the visually-prompted language model. **Fine-Grained Predicate Boosting** uses informative fine-grained predicates to boost the existing scene graph dataset for standard SGG and open-world predicate SGG in a model-agnostic way.

methods have been proposed to alleviate the biased prediction in the training stage. On the other hand, some approaches aim to balance long-tailed distribution classification following specific class distribution [62, 7]. Since the predicates in scene graphs are highly relevant to the context, the direct enhancement methods based on class distribution are inapplicable for the balanced scene graph generation. Hence, [55, 60] utilize visually relevant relationships from external knowledge bases to address the long-tail predicate problem. However, previous approaches require additional hyper-parameters or hand-designed enhancement rules limited to pre-defined scene graphs. In this work, we propose a predicate-boosting framework that can flexibly enhance SGG datasets with diverse fine-grained predicates.

Language Model Prompting. Recently, researchers find that large-scale pre-training models contain rich knowledge and exhibit remarkable generalization capabilities for various downstream tasks [11, 28, 57, 26], thereby achieving comparable performance with only little parameter-tuning [35, 41, 25, 19]. We are also immensely motivated by recent PET work [44], even though it primarily focuses on a semi-supervised situation with many unlabeled instances. FROZEN [50] and BLIP-2 [27] first explore few-shot learning in the multi-modal setting with frozen language models since vision and language can be attended by a unified attention map [22]. However, these naive prompting methods fail to align complex predicate semantics (*i.e.*, ambiguity and co-reference issues) due to their coarse-grained training paradigm. We differ from prior works by introducing the first LM with adaptive semantic cluster loss that can distinguish complex predicate semantics from a linguistic perspective, thus efficiently aligning fine-grained visual cues in scene graphs.

Zero-shot Scene Graph Generation. Current zero-shot SGG methods mainly focus on the generalization of relation combination [38, 47, 21, 13] or only roughly generalize to new categories based on category name similarity [12]. However, they fail to effectively handle the intricate and unseen predicates encountered in real-world scenarios. He *et al.* [17] first introduce open-vocabulary scene graph generation and attempt to predict unseen objects through representation-encoding. But it still cannot transfer to other SGG tasks well because of the enormous cost of dense-caption pre-training. Here we introduce a novel entangled cross-modal prompt to explore the extensibility of CaCao in open-world predicate scene graph generation without costly pre-training.

3. Cross-Modal Predicate Boosting

As illustrated in Figure 2, our **C**ross-mod**a**l predi**Ca**te b**o**osting (**CaCao**) framework mainly consists of three components: 1) First, the *visually-prompted language model* thoroughly exploits linguistic knowledge from pre-trained language models and migrates it into fine-grained predicates generation. 2) Then, *adaptive semantic cluster loss* is proposed to address the semantic co-reference problem in the visually-prompted language model by diverse predicate expression modeling and adaptive adjustment for predicate enhancement. 3) Finally, *fine-grained predicate boosting* uses these enhanced predicates to alleviate the long-tailed

problem of SGG in a model-agnostic way. Furthermore, CaCao can provide various predicates for Epic to achieve open-world SGG. We will elaborate on Epic in Section 4.

3.1. Preliminaries

Scene Graph Generation. In SGG, we try to locate all objects in the image and predict predicates between them to construct scene graphs. Concretely, given an image I, a scene graph $\mathcal{G} = (\mathcal{O}, \mathcal{R})$ corresponding to I has a set of objects $\mathcal{O} = (o_i)_{i=1}^{N_o}$, bounding boxes $\mathcal{B} = (b_i \in \mathbb{R}^4)$and a set of relationship relationships $\mathcal{R} = (s_i, p_i, o_i)_{i=1}^{N_r}, s_i, o_i \in \mathcal{O}$ with different predicate labels $p_i \in \mathcal{P}$, where N_o and N_r are the number of all objects and relationships, respectively.

3.2. Visually-Prompted Language Model

Although several weakly-supervised approaches improve visual relation modeling via specific knowledge bases [58, 45, 55, 60], they require hand-designed rules and have limited generalization ability. As a result, these methods can only enhance specific predicates and cannot flexibly improve tail predicate prediction in various setups. Thus, we attempt to utilize the linguistic knowledge of pre-trained language models to boost fine-grained predicates in a low-resource way and make language models aware of scenes through visual prompts, as shown in the *visually-prompted language model* module of CaCao in Figure 2 (a).

Visually-Prompted Templates. Due to the modality gap between linguistic knowledge and visual content, language models cannot directly perceive the visual relationships in the scene graph. To better utilize visual semantics, we propose the visually-prompted template containing both visual and textual information, which is designed as **X**="[*visually-prefixed prompts*] [P] [SUB][MASK][OBJ]", where [*visually-prefixed prompts*] is an image-conditioned token generated by a transformation layer h_θ from specific visual features and [P] indicates learnable textual prompt for efficient text prompt engineering. During training, we feed our visually-prompted templates into frozen language models to predict correct predicates at the masked position and only update the textual prompt [P] together with the parameters θ in the visual projection layer h_θ.

Cross-Modal Prompt Tuning aims to predict correct fine-grained predicates at the masked position based on cross-modal contexts from **X** by optimizing visually-prompted templates. We randomly collect 80k image-caption pairs from the web (*i.e.*, CC3M, COCO caption), which contain nearly 2k categories of predicates but with much noise of simple predicates. We further design heuristic rules (*e.g.*, corpus co-occurrence frequency) to filter out uninformative (*on, near*) and infrequent (*kneeling by*) predicates **automatically** instead of handling them **manually**. We finally obtain 585 categories of predicates, nearly covering most of the common situations in the real world. During training, we use a softmax classifier to predict the predicate tokens. Formally, we define $\phi(y_i)$ as a K-dimension one-hot label to represent each predicate category Y_i (suppose there are K predicate categories in total). Given the probability distribution $\psi(y_i|X_i)$ at the masked position for each input X_i and the corresponding predicate label $\phi(y_i)$, we can optimize visually-prompted templates as well as the predicate classifier by the Cross-Entropy Loss as follow:

$$\mathcal{L} = -\sum_{i=1}^{N_p} \phi(y_i) log(\psi(y_i|X_i)) \,, \tag{1}$$

where N_p represents the number of predicates for prompt tuning. Note that we only update the parameters of the visual-linguistic projection layer, as shown in Figure 2 (b).

3.3. Adaptive Semantic Cluster Loss

Although visually-prompted templates partially alleviate semantic ambiguity through instance-conditioned hints, it still suffers from semantic co-reference among predicates, where the same predicate semantic might have multiple linguistic expressions shown in Appendix C. Thus, we further design an adaptive semantic clustering loss (ASCL) to refine diverse predicate semantic expressions through synonym clustering structures and context-aware labels. Additionally, it adaptively suppresses excessively boosted categories based on the distribution of predicates, thus facilitating more various predicate distributions in CaCao.

Specifically, we first represent predicates as the average of the BERT [8] embedding vectors of its associated triples due to the strong dependency between triplets in complex scenes. We then cluster these predicates using K-means and initialize the number of centroids based on the similarity threshold between each predicate. During training, we employ semantic-synonym labels to reduce the penalty for predicates in the same cluster to prevent highly correlated predicates from over-suppressing. The objective is then adjusted by context-aware label and semantic-synonym label as follows:

$$\begin{aligned} -\min \sum_{i=1}^{N_p} \mathbb{E}_{\boldsymbol{\epsilon}} \big[\underbrace{\phi(y_i)}_{\text{context-aware label}} + \\ \underbrace{\sum_{j \in C_i} \frac{\epsilon_{i,j}}{|C_i|} \phi(y_j)}_{\text{semantic-synonym label}} \big] log(\psi(y_i|X_i)) \,, \end{aligned} \tag{2}$$

where $\epsilon_{i,j}$ is the correlation coefficient between the predicate y_i and other related predicates y_j in its same cluster C_i. $|C_i|$ represents the number of predicate categories in it.

Furthermore, we observe that assigned predicate augmentation fails to adequately accommodate the dynamic distribution of predicates, leading to the excessive boosting of some specific predicates that destroys diversity. To

address this issue, we set the adaptively re-weighting factor to dynamically adjust the boosting ratio of each predicate based on its proportion during training. We then adjust weights for each category in ASCL as follows:

$$\psi(y_i|X_i) = \frac{e^{z_i}}{\sum_{j=1}^{K} \omega_{ij} e^{z_j}}, \ \omega_{ij} = \delta \frac{z_j}{z_i} \cdot \frac{n_j}{n_i}, \quad (3)$$

where $\{z_i\}_{i=1}^{K}$ and $\{n_i\}_{i=1}^{K}$ represent the predicted logit and the initial number of each predicate category Y_i, respectively. ω_{ij} denotes the adaptively re-weighting factor concerning dynamic distribution between the target boosted predicate of index i and other predicates of index j. δ is a hyper-parameter representing prediction margins. When boosting one predicate enough, we will restrain its enhancement by reducing ω_{ij}, guaranteeing the distribution of generated predicates to be balanced and diverse.

$$\frac{\partial \mathcal{L}_i}{\partial z_j} = \frac{(z_j + 1) \cdot \frac{\delta n_j}{z_i n_i} e^{z_j}}{\sum_{k=1}^{K} \omega_{ik} e^{z_k}} + \underbrace{\left[\frac{\epsilon_{i,j} e^{z_j}}{\sum_{k=1}^{K} \omega_{jk} e^{z_k}} - \epsilon_{i,j} \right]}_{\leq 0, j \in C_i}, \quad (4)$$

Eq. 4 shows the negative gradient for predicate category Y_j. The penalty imposed on negative category Y_j is dynamically adjusted as z_j changes. Furthermore, if the negative category Y_j is related to the positive predicate Y_i, *i.e.*, $j \in C_i$, we will reduce its punishment to encourage the diversity of CaCao. Finally, it results in an adaptively boosting process to promote predicate diversity in CaCao.

3.4. Fine-Grained Predicate Boosting

Although we obtain abundant fine-grained predicates from CaCao, it is not straightforward to directly boost them into the target scene graph due to category inconsistencies with the predicates in the target scene graph. To address this limitation, we propose a fine-grained predicate boosting stage to effectively map open-world predicates to target categories, guaranteeing the smooth alignment of the enhanced predicates with the target scene graph.

Specifically, we establish a simple hierarchy structure of target predicate categories based on lexical analysis and map fine-grained predicates to the target category at each level by cosine similarity of triplet-level embedding. We then select the least frequent category from the mapped candidate target predicates as the final predicate. Note that we only boost unlabeled object pairs that overlap in the scene graph to preserve the original semantics. We will explore more complex structures in the future.

Given the existing scene graph dataset with $|\mathcal{N}|$ labeled samples, our CaCao can generate extra training data $\mathcal{D}$ automatically in a low-resource way and flexibly extend the current dataset by fine-grained predicate boosting. Finally,

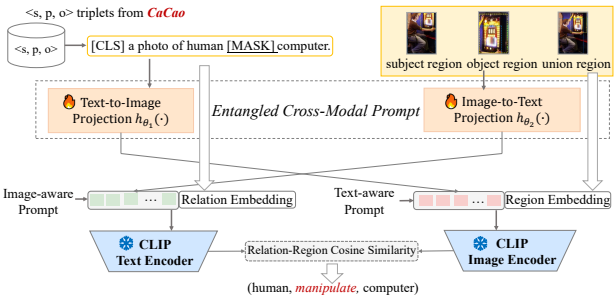


Figure 3. Illustration of our proposed **E**ntangled cross-modal **p**rompt approach for open-world pred**i**cate s**c**ene graph generation (**Epic**). It guides the model to learn the unified embedding of predicates by two complementary prompts in an associative way.

we retrain the refined SGG models with enhanced data $\hat{\mathcal{N}} = (\mathcal{N}, \mathcal{D})$ for a more balanced prediction. We then formulate the learning problem as follows:

$$\min_{\theta} \frac{1}{|\hat{\mathcal{N}}|} \sum_{i=1}^{|\hat{\mathcal{N}}|} L(N_i; \theta), \quad (5)$$

where $L(N_i; \theta)$ denotes the loss function of the learning procedure during the standard scene graph generation.

4. Open-World Predicate SGG

Since CaCao can generate fine-grained predicates, it can provide extra unseen data for open-world generalization. However, open-world predicate SGG has two extra challenges: (a) understanding multi-level semantics of images and triplets; (b) aligning novel predicate semantics into visual and textual contexts. To this end, we propose a novel **E**ntangled cross-modal **p**rompt approach for open-world pred**i**cate s**c**ene graph generation (**Epic**). With the help of Epic, we can fully exploit the potential of CaCao and extend it into open-world predicate SGG, as shown in Figure 3.

Open-World Predicate SGG Backbone. A straightforward way to predict unseen classes in an open world is replacing the fixed classifier with unified embeddings [42]. Let x be the region embedding generated by the visual encoder and $\{p_i\}_{i=1}^{K}$ be a set of relation embeddings produced by the text encoder, p_* is the embedding of the correct predicate. The loss for open-world predicate SGG is then,

$$\mathcal{L}_{open} = -\log \frac{exp(sim(x, p_*)/\tau)}{\sum_{i=1}^{K} \exp(sim(x, p_i)/\tau)}, \quad (6)$$

where p_* is the matched relation embedding and $sim(\cdot, \cdot)$ denotes the cosine similarity. τ is a temperature parameter.

Entangled Cross-Modal Prompt. Moreover, we notice that predicate semantics and image regions are closely related to visual and textual contexts. For example, "man on chair" and "shirt on chair" represent different semantics even though they correspond to the same image region in Figure 1 (a); "man" and "horse" may correspond

to different predicates "riding" or "holding" in different visual contexts. Inspired by the remarkable performance of prompts [36, 34, 64], we introduce entangled cross-modal prompts for text encoder and image encoder to alleviate the above problems. Let $h_{\theta_1}(\cdot)$, $h_{\theta_2}(\cdot)$ represent the the text-to-image and image-to-text projection, respectively in Figure 3. The predicate probability is then computed as:

$$P\left(p^*|x\right) = \frac{\exp\left(sim\left(f_x\left(p_*\right), t_{p_*}\left(x\right)\right)/\tau\right)}{\sum_{i=1}^{K}\exp\left(sim\left(f_x\left(p_i\right), t_{p_i}\left(x\right)\right)/\tau\right)}, \quad (7)$$

where $f_x\left(p_i\right)$ is conditional region embedding based on text-aware prompt $h_{\theta_1}(p_i)$ and $t_{p_i}\left(x\right)$ is conditional relation embedding based on image-aware prompts $h_{\theta_2}(x)$. During training, we only update the projection parameters $(\theta_1;\theta_2)$ to preserve the pre-trained language-vision model's capability for open-world predicate generalization.

5. Experiment

5.1. Dataset and Evaluation Settings

Datasets. We evaluate our proposed method for scene graph generation on the popular VG-50 benchmark similar to previous works [24, 52, 47, 48, 59], which consists of 50 predicate classes and 150 object classes. Furthermore, we explore more challenging datasets (*i.e.* GQA-200 [18, 23], VG-1800 [60]) where predicates are more diverse to validate CaCao's generalization ability in large-scale scenarios.

Data Split. For the standard SGG setting, we adopt a widely used data split following previous works [47, 59, 23] and expand to large-scale SGG datasets. we divide the dataset into 70% training set, 30% testing set, and additional 5k images for parameter tuning. For the open-world predicate SGG setting, we first establish the related dependencies from Chen *et al* [5]. We then randomly select 70% classes from each predicate level and assign them into the base set for training and the rest 30% classes that contain rare predicates (*e.g.*, painted on, flying in) into the novel set for evaluation similar to other zero-shot tasks [1, 21, 17]. To avoid disclosure of the unseen predicates, we remove all relations that contain novel predicates in the training set.

Evaluation and Metrics. Following recent works [60, 39], we evaluate our model on three widely used SGG tasks: PredCls, SGCls, and SGDet. Since the Recall@K of all predicates could be easily affected by biased distribution, it cannot precisely evaluate models' performance on long-tail distribution SGG. Thus, we use Mean Recall@K (**mR@K**) to evaluate the performance of SGG models on the whole category set. We further introduce a detailed metric **Tail-R@K** (Recall@K among tail 50% predicates) to better assess those tail predicates, as these predicates typically provide more information for image understanding. Besides, we use Recall@K of base predicates, novel predicates, and mean Recall@K of total predicates to evaluate the generalization ability of our method on open-world predicate SGG.

5.2. Implementation Details

Visually-Prompted Language Model. We use ViT [10] as the image encoder and set a transformer layer with the 768 embedding size to obtain visually-prefixed prompts. We set the length of visual prompts as 50 and set 10 learnable tokens as textual prompts [*P*] for textual alignment. We use BERT [8] as the language model to predict target predicates and train the model for 15 epochs with batch size of 32. We use AdamW [37] to optimize the model and set the basic learning rate as 3e-4 with a weight delay of 0.0004. Furthermore, the prediction margin δ is set as 9.0.

Object Detector. Following previous works[47, 48, 33], we use a pre-trained Faster R-CNN [43] with ResNet-101-FPN [16] as our backbone and train it on VG-50 dataset with SGD as the optimizer. We then fix the parameters of the object detector during standard SGG training.

Scene Graph Generation. We follow almost the framework of the SOTA unbiased SGG method [39], the only difference is that we integrate the enhanced triplets derived from CaCao into SGG training, thereby directing more attention towards tail predicates without any extra costs. Following [47], SGG models are trained with Cross-Entropy Loss and SGD optimizer by initial learning rate as 1e-3, and batch size as 16. Besides, we train SGG models with 16000 batch iterations for all sub-tasks. For the GQA-200 and VG-1800 datasets, we adjust the training batch iterations to 80000 for further training in large-scale SGG.

For open-world predicate SGG, we use CLIP [42] as the backbone to obtain region embedding and predicate embedding. The text-to-vision and vision-to-text projections are two-layer structures (Linear-ReLU-Linear) with the model dimension $d = 512$ to get the conditional prompts. We set the length of the vision-aware prompt to 4 and the length of the text-aware prompt to 2. Then we use the InfoNCE [40] loss and set the temperature as 0.9 with the batch size of 4 to learn the representation of predicate categories.

5.3. Comparison with State of the Arts

We report the results of our CaCao and other general SGG models for the VG-50 benchmark shown in Table 1. Based on the observation of experimental results, we have summarized the following conclusions:

Our CaCao framework can be flexibly equipped to different baseline models. We incorporate our CaCao into three backbone models for evaluation, including Motif [59], VCTree [48], and Transformer [47]. Despite the model diversity, our CaCao can consistently improve all baseline models' mR@K performance for all tasks that Motif+CaCao (38.9% *v.s.* 16.2%), VCTree+CaCao (40.8% *v.s.* 16.1%) and Transformer+CaCao (43.7% *v.s.* 17.6%) for

	Model Type	Methods	Predicate Classification		Scene Graph Classification		Scene Graph Detection	
			Tail-R@20/50/100 ↑	mR@20/50/100 ↑	Tail-R@20/50/100 ↑	mR@20/50/100 ↑	Tail-R@20/50/100 ↑	mR@20/50/100 ↑
	Specific	BGNN [33]	-/-	- / 30.4 / 32.9	-/-	- / 14.3 / 16.5	-/-	- / 10.7 / 12.6
		PCPL [53]	-/-	- / 35.2 / 37.8	-/-	- / 18.6 / 19.6	-/-	- / 9.5 / 11.7
		SVRP [17]	-/-	- / 24.3 / 25.3	-/-	- / 12.5 / 15.3	-/-	- / 10.5 / 12.8
		DT2-ACBS [7]	-/-	27.4 / 35.9 / 39.7	-/-	18.7 / 24.8 / 27.5	-/-	16.7 / 22.0 / 24.4
	One-stage	SSRCNN [49]	-/-	-/-	-/-	-/-	10.4 / 16.3 / 19.1	13.7 / 18.6 / 22.5
		+CaCao (ours)	-/-	-/-	-/-	-/-	**13.6 / 18.0 / 21.2**	**14.1 / 18.7 / 23.1**
Model-Agnostic strategy	Resample	Motif [59]	10.2 / 13.3 / 14.4	12.1 / 15.2 / 16.2	5.8 / 6.8 / 7.3	7.2 / 8.7 / 9.3	4.8 / 6.0 / 7.3	5.1 / 6.5 / 7.8
		+Resample [2]	-/-	14.7 / 18.5 / 20.0	-/-	9.1 / 11.0 / 11.8	-/-	5.9 / 8.2 / 9.7
	Reweight	+Reweight [51]	16.7 / 26.3 / 31.0	18.8 / 28.1 / 33.7	8.9 / 11.8 / 14.1	10.7 / 15.6 / 18.3	8.6 / 12.1 / 14.6	7.2 / 10.5 / 13.2
		+FGPL [39]	26.7 / 33.3 / 35.7	24.3 / 33.0 / 37.5	16.8 / 19.1 / 19.9	17.1 / 21.3 / 22.5	12.4 / 16.5 / 19.3	11.1 / 15.4 / 18.2
	Causal Rule	+TDE [47]	-/-	18.5 / 25.5 / 29.1	-/-	9.8 / 13.1 / 14.9	-/-	5.8 / 8.2 / 9.8
	Data Enhancement	+Only Caption Relations	16.7 / 20.5 / 21.8	15.2 / 19.8 / 21.2	8.1 / 9.6 / 10.1	8.0 / 9.8 / 10.5	5.3 / 7.7 / 9.4	6.0 / 8.2 / 10.0
		+VisualDS [55]	11.3 / 14.5 / 16.3	13.1 / 16.1 / 17.5	5.9 / 7.0 / 8.3	7.6 / 9.3 / 9.9	5.1 / 6.8 / 7.8	5.4 / 7.0 / 8.3
		+DLFE [6]	-/-	22.1 / 26.9 / 28.8	-/-	12.8 / 15.2 / 15.9	-/-	8.6 / 11.7 / 13.8
		+IETrans [60]	27.3 / 31.3 / 33.2	30.2 / 35.8 / 39.1	13.5 / 15.5 / 16.1	18.2 / 21.5 / 22.8	9.2 / 12.3 / 14.3	12.0 / 15.5 / 18.0
		+CaCao (ours)	**31.4 / 36.1 / 37.6**	**30.9 / 37.1 / 38.9**	**17.3 / 19.7 / 20.5**	**20.4 / 23.3 / 24.4**	**13.9 / 18.4 / 21.6**	**12.6 / 17.1 / 20.0**
	Reweight	VCTree [48]	9.9 / 13.0 / 14.0	11.7 / 14.9 / 16.1	6.2 / 7.4 / 7.9	9.1 / 11.3 / 12.0	4.3 / 6.1 / 7.2	5.2 / 7.1 / 8.3
		+Reweight [51]	23.9 / 30.7 / 33.7	19.4 / 29.6 / 35.3	12.2 / 14.9 / 16.1	13.7 / 19.9 / 23.5	8.4 / 12.2 / 14.7	7.0 / 10.5 / 13.1
	Causal Rule	+FGPL [39]	32.2 / 36.8 / 38.2	30.8 / 37.5 / 40.2	23.5 / 26.5 / 27.5	21.9 / 26.2 / 27.6	13.5 / 17.4 / 20.4	**11.9** / 16.2 / 19.1
		+TDE [47]	-/-	18.4 / 25.4 / 28.7	-/-	8.9 / 12.2 / 14.0	-/-	6.9 / 9.3 / 11.1
	Data Enhancement	+Only Caption Relations	16.2 / 20.3 / 21.7	14.7 / 19.3 / 20.9	8.0 / 9.8 / 10.4	8.2 / 10.1 / 10.8	6.0 / 8.0 / 9.7	5.5 / 7.8 / 9.5
		+DLFE [6]	-/-	20.8 / 25.3 / 27.1	-/-	15.8 / 18.9 / 20.0	-/-	8.6 / 11.7 / 13.8
		+IETrans [60]	27.3 / 31.6 / 33.0	31.7 / 37.0 / 39.7	11.6 / 13.6 / 14.3	18.2 / 19.9 / 21.8	9.0 / 11.8 / 13.7	9.8 / 12.0 / 14.9
		+CaCao (ours)	**33.1 / 37.5 / 38.9**	**33.8 / 39.0 / 40.8**	**23.8 / 27.2 / 28.2**	**23.8 / 27.5 / 28.7**	**14.6 / 19.4 / 22.6**	11.8 / **16.4 / 19.1**
	Reweight	Transformer [47]	10.8 / 13.5 / 14.6	12.4 / 16.3 / 17.6	8.8 / 10.3 / 11.8	8.7 / 10.1 / 10.7	5.3 / 7.3 / 8.8	5.8 / 8.1 / 9.6
		+Reweight [51]	19.9 / 26.0 / 28.4	19.5 / 28.6 / 34.4	9.5 / 12.6 / 13.4	11.9 / 17.2 / 20.7	7.0 / 10.3 / 12.4	8.1 / 11.5 / 14.9
		+FGPL [39]	26.6 / 33.6 / 36.0	27.5 / 36.4 / 40.3	17.0 / 19.9 / 20.1	19.2 / 22.6 / 24.0	13.1 / 17.0 / 19.8	13.2 / 17.4 / 20.3
	Data Enhancement	+Only Caption Relations	16.1 / 19.4 / 20.8	15.0 / 19.3 / 20.9	8.3 / 9.9 / 10.5	8.6 / 10.6 / 11.2	6.4 / 8.9 / 10.6	6.0 / 8.4 / 10.4
		+IETrans [60]	27.5 / 32.0 / 33.7	29.1 / 35.0 / 38.0	14.1 / 16.2 / 16.7	17.9 / 20.8 / 22.3	11.6 / 14.9 / 17.6	11.7 / 15.0 / 18.1
		+CaCao (ours)	**31.7 / 35.7 / 37.0**	**36.2 / 41.7 / 43.7**	**19.0 / 22.2 / 23.3**	**21.1 / 24.0 / 25.0**	**14.1 / 18.7 / 21.9**	**13.5 / 18.3 / 22.1**

Table 1. Performance (%) of our method **CaCao** and other baselines with different model types on the VG-50 dataset.

	Model	PredCls mR@50/100	SGCls mR@50/100	SGDet mR@50/100
GQA-200	Motif [59]	16.4 / 17.1	8.2 / 8.6	6.4 / 7.7
	Motif + GCL[9]	36.7 / 38.1	17.3 / 18.1	16.8 / 18.8
	Motif + CaCao (ours)	**37.5 / 40.5**	**19.6 / 21.9**	**17.8 / 19.6**
	Transformer[47]	17.5 / 18.7	8.5 / 9.0	6.6 / 7.8
	Transformer + GCL[9]	**35.6** / 36.7	17.8 / 18.3	16.6 / 18.1
	Transformer + CaCao (ours)	34.8 / **36.9**	**19.3 / 20.1**	**18.8 / 19.1**
VG-1800	BGNN [33]	1.3 / 2.4	0.8 / 1.4	0.5 / 0.9
	Motif [59]	1.7 / 2.6	0.9 / 1.9	0.6 / 1.1
	Motif + IETrans [60]	5.1 / 8.4	3.6 / 5.2	3.1 / 4.3
	Motif + CaCao (ours)	**10.0 / 10.8**	**4.6 / 6.3**	**4.1 / 6.2**

Table 2. Comparisons with our CaCao and other baseline methods on large-scale SGG datasets.

	Methods	Datasets	Predicate Classification		
			base R@50/100	novel R@50/100	total mR@50/100
	Backbone w/o Epic [42]	VG	17.6 / 21.1	6.4 / 8.7	8.5 / 9.7
		CaCao	17.4 / 20.4	7.2 / 9.2	8.1 / 10.4
		VG+CaCao	17.5 / 20.9	11.2 / 15.8	13.6 / 17.7
Ablations	**Epic**	VG	22.6 / 27.2	7.4 / 9.7	10.3 / 12.6
		CaCao	23.1 / 30.8	9.7 / 12.1	14.2 / 18.2
		VG+CaCao	**28.3 / 31.1**	**13.9 / 18.3**	**16.5 / 21.8**
	w/o text-aware prompt	VG+CaCao	16.8 / 23.1	12.5 / 13.9	13.1 / 15.4
	w/o vision-aware prompt	VG+CaCao	18.5 / 24.9	10.1 / 12.7	11.2 / 14.1

Table 3. Performance (%) of our **Epic** and the backbone without Epic for open-world settings on different datasets. VG denotes the VG-50 dataset with the open-world split, VG+CaCao represents the enhanced dataset with our CaCao framework and CaCao means only use CaCao's predicates for unsupervised settings.

PredCls. Also, we obtain similar performance improvements for SGCls and SGDet. Besides, we compare the ablation methods which directly extract raw relation triplets from captions (*i.e.* only Caption Relations in Tab. 1). Notably, our method Transformer+CaCao significantly surpasses the ablation method by 23.8% in mR@20 of PredCls, demonstrating that the gain power of CaCao is mainly derived from linguistic knowledge in PLM instead of extra collected data. Conversely, the triplets directly extracted from image captions are incomplete that only describe general semantics or partial visual relationships.

Compared with other model-agnostic methods, our CaCao outperforms all of them in both Tail-R@K and mR@K. Specifically, CaCao exceeds the SOTA of data enhancement models IETrans [60] for all three backbones with consistent improvements as 3.9%, 5.8%, and 3.6% on Tail-R@20 for PredCls and 0.7%, 2.1%, and 7.1% on mR@20 for Predcls. It shows that our CaCao can generate high-quality informative predicates to mitigate the long-tail distribution problem, which is conducive to fine-grained scene graph generation. It is worth noting that even when compared to SOTA methods of different model types, such as FGPL [39], Motif+CaCao, VCTree+CaCao, and Transformer+CaCao still achieve significant improvements by 6.6%, 3.0%, and 8.7% on mR@20 for PredCls. Besides, our CaCao can also integrate with one-stage methods (*e.g.*, SSRCNN [49]) and achieve better performance.

Our method can distinguish fine-grained predicates and achieve a large margin of improvements on these predicate predictions. Notably, our modal-agnostic approach can also achieve competitive performance compared with strong specific baselines (*e.g.*, 43.7% *v.s.* 39.7% on mR@100 for PredCls), demonstrating the superiority of our proposed model. For an intuitive illustration of CaCao's discriminatory power among hard-to-distinguish predicates, we visualize the PredCls results of fine-grained predicates

as shown in Figure 4. We observe that Transformer+CaCao obtains overall improvement on most predicates. One possible reason is that CaCao has been exposed to various informative predicates, strengthening its discriminatory power against fine-grained predicates. Qualitatively, we further visualize the prediction results of our Transformer+CaCao compared with its baseline model Transformer [47], shown in Figure 5. In the case of Transformer+CaCao, we observe a substantial improvement in the predicted ratio for the correct predicate '*flying in*' (8% $\rightarrow$ 40%). This result demonstrates the capability of Transformer+CaCao to effectively distinguish fine-grained predicates, as opposed to roughly predicting head predicates (*i.e.*, on, in).

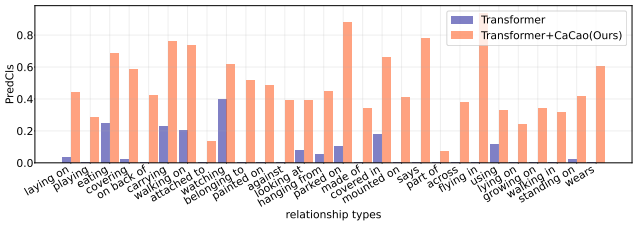


Figure 4. Diverse fine-grained predicates performance comparison between base Transformer [47] and our enhanced Transformer+CaCao on the VG-50 dataset.

5.4. Generalization to Large-Scale SGG

Table 2 summarizes the results of our CaCao and other baselines on large-scale datasets. Overall, our method can successfully generalize to more challenging datasets. Notably, simply resampling (*i.e.* BGNN [33]) can not work well in such exacerbated scenarios, where much more predicates have less than 10 samples. In contrast, our CaCao utilizes abundant corpus knowledge to balance diverse tail predicates and surpasses other baselines for almost large-scale SGG tasks. For quantitative comparison, our CaCao can obtain consistent improvement as 8.2%, 4.4%, and 5.1% on mR@100 for PredCls, SGCls, and SGDet in VG-1800 and largely enhance the unbiased predictions in GQA-200 (*e.g.*, 11.9% improvement with Motif [59] and 11.7% improvement with Transformer [47] on SGDet mR@100).

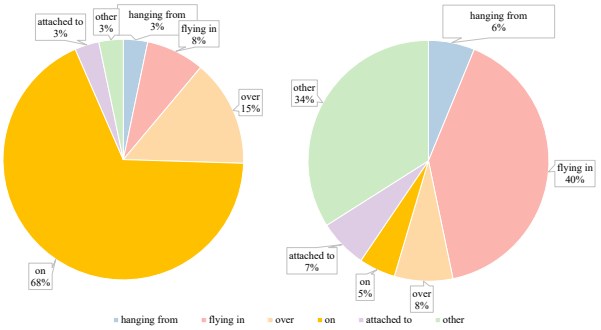


Figure 5. The prediction distribution of CaCao on the fine-grained predicate '*flying in*'. The left pie chart shows the distribution by Transformer [47] and the right pie chart shows the prediction distribution of various predicates by Transformer+CaCao.

5.5. Expansibility to Open-World Predicate SGG

Inspired by the abundant fine-grained predicates produced by CaCao, we also validate our CaCao with Epic on the open-world setting for the base, novel, and total PredCls tasks to show its expansibility to open-world predicate scene graph generation. Since current SGG models cannot solve this challenging task, we verify the performance of CaCao and Epic by comparing them with the naive backbone and present the comparison results fully in Table 3.

Empirically, CaCao can bring out more informative predicates for better generalization. With the help of diverse predicates from CaCao, our Epic obtained a significant improvement of 9.6% on novel R@100 for PredCls, verifying its effectiveness for challengeable open-world predicate SGG. The CaCao and Epic not only improve the novel categories but also greatly improve the base categories on PredCls (28.3% / 31.1% v.s. 17.6% / 21.1%), indicating that the entangled cross-modal prompt can provide general benefit to the representation learning of predicates, rather than merely an additional hint to the unseen predicates.

Surprisingly, even only using rich predicates generated by CaCao, Epic still performs better than VG alone. It indicates that predicates generated by CaCao are diverse and contain richer information to help SGG models learn predicate semantics for predicate generalization.

	Methods	Predicate Prediction Accuracy			
		ASCL	TPT	VPT	A@1/10 ↑
1	Backbone	✗	✗	✗	0.08 / 0.21
2	w/o ASCL	✗	✓	✓	0.38 / 0.74
3	w/o TPT	✓	✗	✓	0.47 / 0.80
4	w/o VPT	✓	✓	✗	0.25 / 0.68
8	**CaCao**	✓	✓	✓	**0.74 / 0.92**

Table 4. Ablation study on each module of our proposed CaCao with predicate labels prediction accuracy (**A@1/10**) metrics.

5.6. In-depth Analysis

Visually-Prompted Language model. To deeply investigate our CaCao, we further study the ablation variants of different modules in Table 4. Specifically, we train the following ablation models. 1) w/o ASCL: we remove the Adaptive Semantic Cluster Loss (ASCL). 2) w/o TPT: we remove the Textual Prompt (TPT) in prompt templates. 3) w/o VPT: we remove the visually-prefixed Prompt (VPT). We use Acc@1/10 as metric in Table 4 because it assesses the prediction accuracy of predicates from CaCao equally.

The results of Row 2 indicate that adaptive semantic cluster learning is crucial for diverse fine-grained predicate prediction. Also, the results of Row 3 validate the importance of learnable prompts on textual semantic understanding. Furthermore, Row 4 suggests that the main perfor-

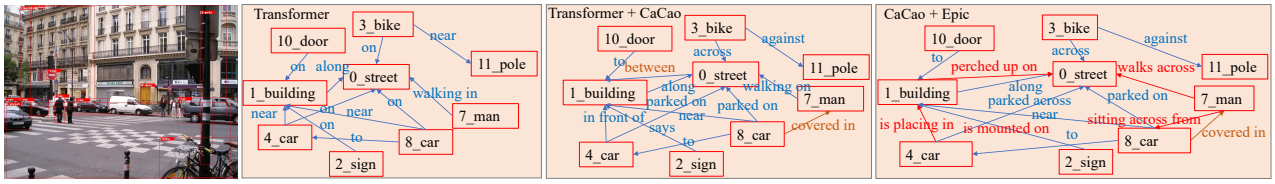


Figure 6. Visualization of base Transformer model [47], Transformer equipped with our CaCao framework for predicate enhancement and our Epic equipped with CaCao framework for open-world predicate SGG.

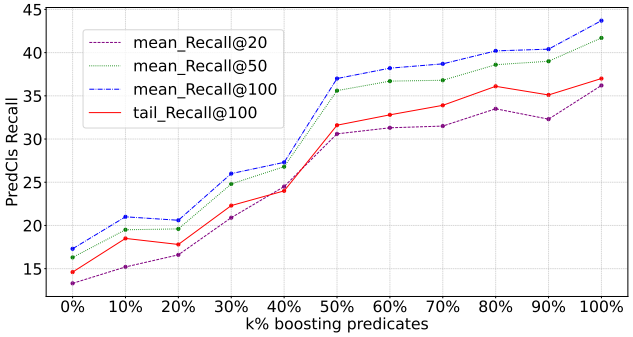


Figure 7. The influence of different proportions k% of boosted predicates for different mR@K and tail-R@100 (**red line**).

mance gain comes from these visual semantics contained in images (0.25 / 0.68 → 0.74 / 0.92).

Influence of k% Boosting Predicates. As shown in Figure 7, with the increase of k% boosting predicates, the mean recall and the tail recall gradually increased in the form of overall growth. The phenomenon indicates that predicates enhanced by CaCao are all informative and consistently bring enhancements to the existing SGG models.

Adaptive Semantic Cluster Learning. Since the quality of clustering is critical for the adaptive prompt tuning in CaCao, we further explore the effect of predicate clustering under different similarity threshold initialization on the fine-grained predicates generation in Table 5. Here we use A@1 as the ablation metric to clearly show the performance of predicates generation. Our observations reveal that excessively low or high similarity thresholds can lead to a decrease in predicate prediction accuracy. The possible reason is that too low similarity aggregate nearly all predicates into the same cluster and too high similarity regards each predicate individually may lead to incorrect clusters. Thus, we set the appropriate threshold as 0.7 for ASCL and obtain the optimal performance of 0.74 A@1 in CaCao.

Similarity threshold	w/o ASCL	0.1	0.3	0.5	0.7	0.9
A@1	0.38	0.39	0.57	0.63	**0.74**	0.48

Table 5. The influence of different predicate similarity threshold for cross-modal prompt tuning in CaCao.

Entangled Cross-Modal Prompts. We explore the effectiveness of the text-aware prompt and the vision-aware prompt in Epic, shown in the last two lines of Table 3. We gradually removed these entangled prompts and observed a significant decrease in performance for both base and novel classes without either prompt from another modality. These findings suggest mutual hints between the two modalities are necessary to extract associated linguistic semantics and image features for open-world predicate learning.

Human Evaluation. A key element of effective SGG boosting is to obtain high-quality data. Thus, we conduct a human evaluation for automatic labels from CaCao and find the radio of reasonable fine-grained predicates is **73%**. Please refer to Appendix D for more details.

Visualization Results. In Figure 6, we visualize the enhancement SGG benefits from CaCao compared with the base scene graph and further present open-world predicate SGG visualization results by CaCao+Epic, intuitively illustrating the effectiveness of our proposed CaCao and Epic. The examples (blue labels) in Figure 6 clearly show that the Transformer+CaCao successfully generates more fine-grained predicates than the Transformer, such as "car-parked on-street" instead of "car-on-street". In addition, we find that with the help of CaCao and Epic, our model can predict additional predicates (orange labels) and even predicates of unseen categories (red labels), such as "building-between-street" and "man-walk across-street".

6. Conclusions

In this work, we propose an automatic boosting framework CaCao that exploits linguistic knowledge from pre-trained language models to enrich existing datasets in a low-resource way. We tackle the long-tail issue of SGG with the help of abundant informative predicates from CaCao and generalize to open-world predicate learning with the entangled cross-modal prompt design based on VL models. Our extensive experiments on three datasets illustrate the significant improvement of our CaCao on fine-grained scene graph generation and open-world generalization capability.

Acknowledgment. This work has been supported in part by the Zhejiang NSF (LR21F020004), the National Key R&D Program of China (2022ZD0160101), the NSFC (No. 62272411), Alibaba-Zhejiang University Joint Research Institute of Frontier Technologies, and Ant Group. We thank all the reviewers for their valuable comments.

References

[1] Ankan Bansal, Karan Sikka, Gaurav Sharma, Rama Chellappa, and Ajay Divakaran. Zero-shot object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 384–400, 2018. 6

[2] Evgeny Burnaev, Pavel Erofeev, and Artem Papanov. Influence of resampling on accuracy of imbalanced classification. In *Eighth international conference on machine vision (ICMV 2015)*, volume 9875, pages 423–427. SPIE, 2015. 2, 7

[3] Xiaojun Chang, Pengzhen Ren, Pengfei Xu, Zhihui Li, Xiaojiang Chen, and Alex Hauptmann. A comprehensive survey of scene graphs: Generation and application. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):1–26, 2021. 1

[4] Tianshui Chen, Weihao Yu, Riquan Chen, and Liang Lin. Knowledge-embedded routing network for scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6163–6171, 2019. 1, 2

[5] Vincent S Chen, Paroma Varma, Ranjay Krishna, Michael Bernstein, Christopher Re, and Li Fei-Fei. Scene graph prediction with limited labels. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2580–2590, 2019. 6

[6] Meng-Jiun Chiou, Henghui Ding, Hanshu Yan, Changhu Wang, Roger Zimmermann, and Jiashi Feng. Recovering the unbiased scene graphs from the biased ones. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1581–1590, 2021. 7

[7] Alakh Desai, Tz-Ying Wu, Subarna Tripathi, and Nuno Vasconcelos. Learning of visual relations: The devil is in the tails. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15404–15413, 2021. 1, 3, 7

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 4, 6

[9] Xingning Dong, Tian Gan, Xuemeng Song, Jianlong Wu, Yuan Cheng, and Liqiang Nie. Stacked hybrid-attention and group collaborative learning for unbiased scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19427–19436, 2022. 7

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 6

[11] Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. Measuring and improving consistency in pretrained language models. *Transactions of the Association for Computational Linguistics*, 9:1012–1031, 2021. 3

[12] Nikolaos Gkanatsios, Vassilis Pitsikalis, and Petros Maragos. From saturation to zero-shot visual relationship detection using local context. In *BMVC*, 2020. 3

[13] Arushi Goel, Basura Fernando, Frank Keller, and Hakan Bilen. Not all relations are equal: Mining informative labels for scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15596–15606, 2022. 3

[14] Nishad Gothoskar, Marco Cusumano-Towner, Ben Zinberg, Matin Ghavamizadeh, Falk Pollok, Austin Garrett, Josh Tenenbaum, Dan Gutfreund, and Vikash Mansinghka. 3dp3: 3d scene perception via probabilistic programming. *Advances in Neural Information Processing Systems*, 34:9600–9612, 2021. 1

[15] Xinzhe Han, Shuhui Wang, Chi Su, Qingming Huang, and Qi Tian. Greedy gradient ensemble for robust visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1584–1593, 2021. 1

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6

[17] Tao He, Lianli Gao, Jingkuan Song, and Yuan-Fang Li. Towards open-vocabulary scene graph generation with prompt-based finetuning. *arXiv preprint arXiv:2208.08165*, 2022. 3, 6, 7

[18] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019. 6

[19] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021. 3

[20] Huaizu Jiang, Ishan Misra, Marcus Rohrbach, Erik Learned-Miller, and Xinlei Chen. In defense of grid features for visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10267–10276, 2020. 1

[21] Xuan Kan, Hejie Cui, and Carl Yang. Zero-shot scene graph relation prediction through commonsense knowledge integration. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 466–482. Springer, 2021. 3, 6

[22] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR, 2021. 2, 3

[23] Boris Knyazev, Harm de Vries, Cătălina Cangea, Graham W Taylor, Aaron Courville, and Eugene Belilovsky. Generative compositional augmentations for scene graph prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15827–15837, 2021. 6

[24] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalan-

tidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017. 6

[25] Juncheng Li, Minghe Gao, Longhui Wei, Siliang Tang, Wenqiao Zhang, Mengze Li, Wei Ji, Qi Tian, Tat-Seng Chua, and Yueting Zhuang. Gradient-regulated meta-prompt learning for generalizable vision-language models. 2023. 3

[26] Juncheng Li, Xin He, Longhui Wei, Long Qian, Linchao Zhu, Lingxi Xie, Yueting Zhuang, Qi Tian, and Siliang Tang. Fine-grained semantically aligned vision-language pre-training. *Advances in neural information processing systems*, 35:7290–7303, 2022. 2, 3

[27] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 3

[28] Juncheng Li, Kaihang Pan, Zhiqi Ge, Minghe Gao, Hanwang Zhang, Wei Ji, Wenqiao Zhang, Tat-Seng Chua, Siliang Tang, and Yueting Zhuang. Empowering vision-language models to follow interleaved vision-language instructions, 2023. 3

[29] Juncheng Li, Siliang Tang, Linchao Zhu, Haochen Shi, Xuanwen Huang, Fei Wu, Yi Yang, and Yueting Zhuang. Adaptive hierarchical graph reasoning with semantic coherence for video-and-language inference. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1867–1877, 2021. 1

[30] Juncheng Li, Siliang Tang, Linchao Zhu, Wenqiao Zhang, Yi Yang, Tat-Seng Chua, and Fei Wu. Variational cross-graph reasoning and adaptive structured semantics learning for compositional temporal grounding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 1

[31] Juncheng Li, Junlin Xie, Long Qian, Linchao Zhu, Siliang Tang, Fei Wu, Yi Yang, Yueting Zhuang, and Xin Eric Wang. Compositional temporal grounding with structured variational cross-graph correspondence learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3032–3041, 2022. 1

[32] Lin Li, Long Chen, Yifeng Huang, Zhimeng Zhang, Songyang Zhang, and Jun Xiao. The devil is in the labels: Noisy label correction for robust scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18869–18878, 2022. 2

[33] Rongjie Li, Songyang Zhang, Bo Wan, and Xuming He. Bipartite graph network with adaptive message passing for unbiased scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11109–11119, 2021. 1, 2, 6, 7, 8

[34] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597, 2021. 6

[35] Sheng Liang, Mengjie Zhao, and Hinrich Schütze. Modular and parameter-efficient multimodal fusion with prompting. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2976–2985, 2022. 3

[36] Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 61–68, 2022. 6

[37] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[38] Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. Visual relationship detection with language priors. In *European conference on computer vision*, pages 852–869. Springer, 2016. 3

[39] Xinyu Lyu, Lianli Gao, Yuyu Guo, Zhou Zhao, Hao Huang, Heng Tao Shen, and Jingkuan Song. Fine-grained predicates learning for scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19467–19475, 2022. 6, 7

[40] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 6

[41] Bosheng Qin, Haoji Hu, and Yueting Zhuang. Deep residual weight-sharing attention network with low-rank attention for visual question answering. *IEEE Transactions on Multimedia*, 2022. 3

[42] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 2, 5, 6, 7

[43] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 6

[44] Timo Schick and Hinrich Schütze. It's not just size that matters: Small language models are also few-shot learners. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2339–2352, 2021. 2, 3

[45] Jing Shi, Yiwu Zhong, Ning Xu, Yin Li, and Chenliang Xu. A simple baseline for weakly-supervised scene graph generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16393–16402, 2021. 4

[46] Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. *arXiv preprint arXiv:2010.15980*, 2020. 2

[47] Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. Unbiased scene graph generation from biased training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3716–3725, 2020. 2, 3, 6, 7, 8, 9

[48] Kaihua Tang, Hanwang Zhang, Baoyuan Wu, Wenhan Luo, and Wei Liu. Learning to compose dynamic tree structures

for visual contexts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6619–6628, 2019. 1, 2, 6, 7

[49] Yao Teng and Limin Wang. Structured sparse r-cnn for direct scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19437–19446, 2022. 7

[50] Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems*, 34:200–212, 2021. 3

[51] Jiaqi Wang, Wenwei Zhang, Yuhang Zang, Yuhang Cao, Jiangmiao Pang, Tao Gong, Kai Chen, Ziwei Liu, Chen Change Loy, and Dahua Lin. Seesaw loss for long-tailed instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9695–9704, 2021. 2, 7

[52] Danfei Xu, Yuke Zhu, Christopher B Choy, and Li Fei-Fei. Scene graph generation by iterative message passing. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5410–5419, 2017. 1, 6

[53] Shaotian Yan, Chen Shen, Zhongming Jin, Jianqiang Huang, Rongxin Jiang, Yaowu Chen, and Xian-Sheng Hua. Pcpl: Predicate-correlation perception learning for unbiased scene graph generation. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 265–273, 2020. 1, 2, 7

[54] Xuewen Yang, Yingru Liu, and Xin Wang. Reformer: The relational transformer for image captioning. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 5398–5406, 2022. 1

[55] Yuan Yao, Ao Zhang, Xu Han, Mengdi Li, Cornelius Weber, Zhiyuan Liu, Stefan Wermter, and Maosong Sun. Visual distant supervision for scene graph generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15816–15826, 2021. 1, 3, 4, 7

[56] Jing Yu, Yuan Chai, Yujing Wang, Yue Hu, and Qi Wu. Cogtree: Cognition tree loss for unbiased scene graph generation. *arXiv preprint arXiv:2009.07526*, 2020. 1, 2

[57] Qifan Yu, Juncheng Li, Wentao Ye, Siliang Tang, and Yueting Zhuang. Interactive data synthesis for systematic vision adaptation via llms-aigcs collaboration, 2023. 3

[58] Alireza Zareian, Svebor Karaman, and Shih-Fu Chang. Weakly supervised visual semantic parsing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3736–3745, 2020. 1, 4

[59] Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. Neural motifs: Scene graph parsing with global context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5831–5840, 2018. 1, 2, 6, 7, 8

[60] Ao Zhang, Yuan Yao, Qianyu Chen, Wei Ji, Zhiyuan Liu, Maosong Sun, and Tat-Seng Chua. Fine-grained scene graph generation with data transfer. *arXiv preprint arXiv:2203.11654*, 2022. 2, 3, 4, 6, 7

[61] Chaoyi Zhang, Jianhui Yu, Yang Song, and Weidong Cai. Exploiting edge-oriented reasoning for 3d point-based scene graph analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9705–9715, 2021. 1

[62] Songyang Zhang, Zeming Li, Shipeng Yan, Xuming He, and Jian Sun. Distribution alignment: A unified framework for long-tail visual recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2361–2370, 2021. 3

[63] Wenqiao Zhang, Haochen Shi, Siliang Tang, Jun Xiao, Qiang Yu, and Yueting Zhuang. Consensus graph representation learning for better grounded image captioning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 3394–3402, 2021. 1

[64] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. 6