Predicate Debiasing in Vision-Language Models Integration for Scene Graph Generation Enhancement

Yuxuan Wang and Xiaoyuan Liu
Nanyang Technological University

Abstract

Scene Graph Generation (SGG) provides basic language representation of visual scenes, requiring models to grasp complex and diverse semantics between objects. This complexity and diversity in SGG leads to underrepresentation, where parts of triplet labels are rare or even unseen during training, resulting in imprecise predictions. To tackle this, we propose integrating the pretrained Vision-language Models to enhance representation. However, due to the gap between pretraining and SGG, direct inference of pretrained VLMs on SGG leads to severe bias, which stems from the imbalanced predicates distribution in the pretraining language set. To alleviate the bias, we introduce a novel **LM Estimation** to approximate the unattainable predicates distribution. Finally, we ensemble the debiased VLMs with SGG models to enhance the representation, where we design a **certainty-aware** indicator to score each sample and dynamically adjust the ensemble weights. Our training-free method effectively addresses the predicates bias in pretrained VLMs, enhances SGG's representation, and significantly improve the performance.

1 Introduction

Scene Graph Generation (SGG) is a fundamental vision-language task that has attracted much effort. It bridges natural languages with scene representations and serves various applications, from robotic contextual awareness to helping visually impaired people. The key challenge in SGG is to grasp complex semantics to understand inter-object relationships in a scene.

Existing researches in SGG focus primarily on refining model architectures that are trained from scratch with datasets like Visual Genome (Krishna et al., 2017) or Open Images (Kuznetsova et al., 2020). However, SGG tasks inherently face another challenge of underrepresentation. Due to the inherent complexities of SGG, there exists exponential variability of triplets combined by the *subject*, *object*, and *relation (predicate)*. It is extremely challenging for a training set to cover such diversity. As a result, a part of the test distribution is underrepresented in training, leading to poor prediction quality. In a severe case, some triplet labels that appear in the test set are unseen in training.

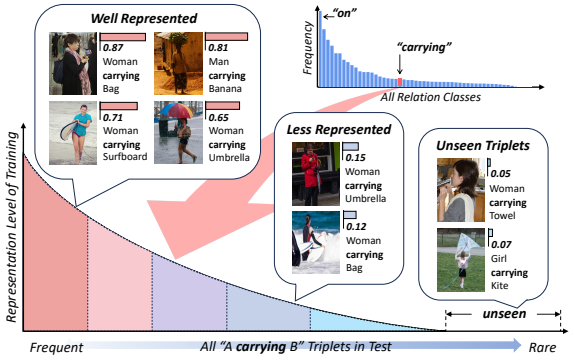

Figure 1: Illustration of the underrepresentation issue in Visual Genome. We highlight the relation class *"carrying"* from the *top-right* imbalanced class distribution. We present various samples with their training representation levels and confidence scores for the ground truth class, where lower scores indicate poorer prediction quality. We find that samples less represented by the training set tend to have lower-quality predictions.

In Figure 1, we highlight the relation class "*carrying*" from Visual Genome, showing samples and their confidence scores of the ground truth class from a baseline model's predictions. While well-represented samples score higher, the samples labeled with unseen triplets like "woman *carrying* towel" score fairly low. Furthermore, one "woman *carrying* umbrella" scores only 0.15 due to the umbrella being closed, while its counterpart with an open umbrella scores markedly higher (0.65). Although the triplet is seen in training set, the closed "umbrella" is still short of representation.

A straightforward solution to this issue is to

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 1627–1639
November 12-16, 2024

expand the model's knowledge by integrating advanced vision-language models (VLMs) pretrained on extensive datasets (Kim et al., 2021; Li et al., 2020, 2019; Qi et al., 2020; Yu et al., 2022; Radford et al., 2021), using their comprehensive knowledge to compensate for underrepresented samples. Employing the Masked Language Modeling (MLM) prompt format, such as "woman is [MASK] towel," allows for direct extraction of relation predictions from the fill-in answers provided by zero-shot VLMs, which fully preserve the pretraining knowledge. Nonetheless, this direct inference of zero-shot models on SGG introduces significant predicate bias due to disparities in data distribution and objectives between pretraining and SGG tasks.

This predicate bias originates from the imbalanced frequency of predicates in the pretraining language set, causing the VLMs to favor the predicates that are prevalent in the pretraining data. Unfortunately, existing debiasing methods rely on explicit training distribution, which is often unattainable for pretrained VLMs: (1) The pretraining data are often confidential. (2) Since the pretraining objectives are different with SGG, there is no direct label correspondence from pretraining to SGG.

To alleviate the predicate bias, we introduce a novel approach named **Lagrange-Multiplier Estimation** (LM Estimation) based on constrained optimization. Since there is no explicit distribution of relation labels in the pretraining data, LM Estimation seeks to estimate a surrogate distribution of SGG predicates within VLMs. Upon obtaining the estimated distribution, we proceed with predicates debiasing via post-hoc logits adjustment. Our LM Estimation, as demonstrated by comprehensive experiments, is proved to be exceedingly effective in mitigating the bias for zero-shot VLMs.

Finally, we ensemble the debiased VLMs with the SGG models to address their underrepresentation issue. We observe that some samples are better represented by the zero-shot VLM, while others align better with the SGG model. Therefore, we propose to dynamically ensemble the two models. For each sample, we employ a **certainty-aware** indicator to score its representation level in the pretrained VLM and the SGG model, which subsequently determines the ensemble weights. Our contributions can be summarized as follows:

- While existing methods primarily focuses on refining model architecture, we are among the pioneers in addressing the inherent underrepresentation issue in SGG using pretrained VLMs.
- Towards the predicates bias underlying in the pretraining language set, we propose our LM Estimation, a concise solution to estimate the unattainable words' distribution in pretraining.
- We introduce a plug-and-play method that dynamically ensemble the zero-shot VLMs. Needing no further training, it minimizes the computational and memory burdens. Our method effectively enhances the representation in SGG, resulting in significant performance improvement.

2 Related Work

Scene Graph Generation (SGG) is a fundamental task for understanding the relationships between objects in images. Various of innovations (Tang et al., 2019; Gu et al., 2019; Li et al., 2021; Lin et al., 2022a, 2020, 2022b; Zheng et al., 2023; Xu et al., 2017) have been made in supervised SGG from the Visual Genome benchmark (Krishna et al., 2017). A typical approach involves using a Faster R-CNN (Sun et al., 2018) to identify image regions as objects, followed by predicting their interrelations with a specialized network that considers their attributes and spatial context. Existing efforts (Li et al., 2021; Lin et al., 2022a,b; Zheng et al., 2023) mainly focus on enhancing this prediction network. For instance, (Lin et al., 2022b) introduced a regularized unrolling approach, and (Zheng et al., 2023) used a prototypical network for improved representation. These models specially tailored for SGG has achieved a superior performance.

Unbiased Learning in SGG has been a long-standing challenge. Started by (Tang et al., 2020), the debiasing methods (Dong et al., 2022; Li et al., 2021; Yan et al., 2020; Li et al., 2022b,a) seek to removing the relation label bias stemming from the imbalanced relation class distribution. These works have achieved more balanced performance across all relation classes. However, these methods rely on the interfere during training and are not feasible to the predicate bias in pre-trained VLMs.

Pre-trained Vision-Language models (VLMs) have been widely applied in diverse vision-language tasks (Su et al., 2019; Radford et al., 2021; Kim et al., 2021; Li et al., 2020) and have achieved substantial performance improvements with the vast knowledge base obtained during pre-training. Recently works start to adapt the comprehensive pre-trained knowledge in VLMs to relation recognition and scene graph generation (He et al., 2022; Gao et al., 2023; Li et al., 2023; Yu et al., 2023;

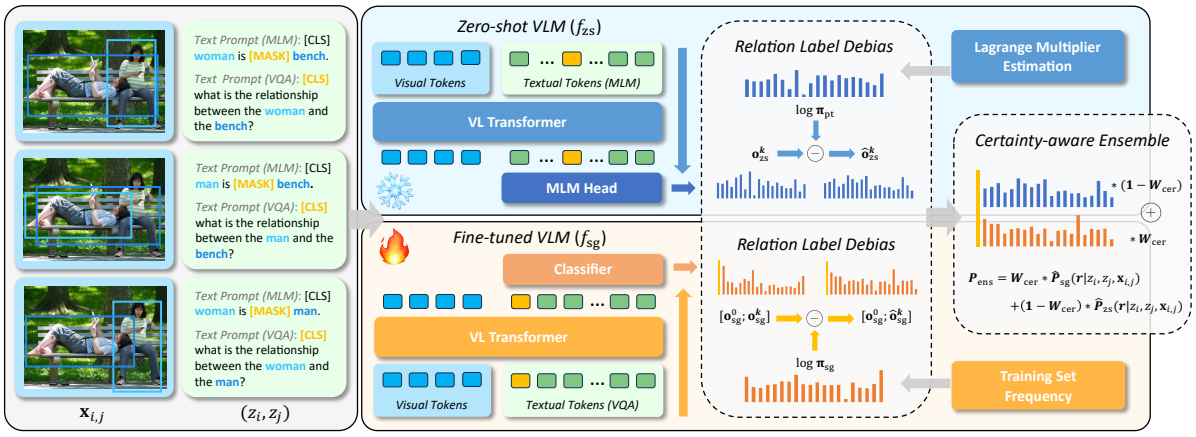

Figure 2: Illustration of our proposed architecture. *left*: the visual-language inputs processed from image regions $\mathbf{x}_{i,j}$ and object labels (z_i, z_j), either provided or predicted by Faster R-CNN detector. *middle*: the fixed zero-shot VLM f_{zs} and the trainable task-specific models f_{sg}, which we use a fine-tuned VLM as example. *right*: the relation label debias process and the certainty-aware ensemble.

Zhang et al., 2023; Zhao et al., 2023). Through prompt-tuning, (He et al., 2022) is the first employing VLMs to open-vocabulary scene graph generation. Then more approaches (Zhang et al., 2023; Yu et al., 2023; Gao et al., 2023) are designed towards this task. These works demonstrate the capability of VLMs on recognizing relation, inspiring us to utilize VLMs to improve the SGG representation.

3 Methodology

3.1 Setup

Given an image data $(\mathbf{x}, \mathcal{G})$ from a SGG dataset $\mathcal{D}_{\text{sg}}$, the image $\mathbf{x}$ is parsed into a scene graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$, where $\mathcal{V}$ is the object set and $\mathcal{E}$ is the relation set. Specifically, each object $\mathbf{v} \in \mathcal{V}$ consists of a corresponding bounding box $\mathbf{b}$ and a categorical label z either from annotation or predicted by a trained Faster R-CNN detector; each $\mathbf{e}_{i,j} \in \mathcal{E}$ denotes the relation for the subject-object pair $\mathbf{v}_i$ and $\mathbf{v}_j$, represented by a predicate label $y \in \mathcal{C}_e$. The predicate relation space $\mathcal{C}_e = \{0\} \cup \mathcal{C}_r$ includes one *background* class 0, indicating no relation, and K *non-background* relations $\mathcal{C}_r = [K]$. The objective is to learn a model f that, given the predicted objects z_i and z_j for each pair with their cropped image region $\mathbf{x}_{i,j} = \mathbf{x}(\mathbf{b}_i \cup \mathbf{b}_j)$, produces logits $\mathbf{o}$ for all relations $y \in \mathcal{C}_e$, *i.e.*, $\mathbf{o} = f(z_i, z_j, \mathbf{x}_{i,j})$.

3.2 Method Overview

As depicted in Figure 2, our framework f comprising two branches: a fixed zero-shot VLM f_{zs} and a task-specific SGG model f_{sg} trained on $\mathcal{D}_{\text{sg}}$. Here, we employ a SGG fine-tuned VLM as f_{sg}, where we forward the image region $\mathbf{x}_{i,j}$ to the visual encoder and use the prompt template "what is the relationship between the $\{z_i\}$ and the $\{z_j\}$?" as the text input. Then, a classifier head is added to the `[CLS]` token to generate logits $\mathbf{o}_{\text{sg}}$ of all relations $y \in \mathcal{C}_e$. Our experiments also adopt SGG models from recent works as f_{sg}.

Another zero-shot model, represented as f_{zs}, leverages pretrained knowledge to the SGG task without fine-tuning. By providing prompts to zero-shot VLMs in the form "$\{z_i\}$ is `[MASK]` $\{z_j\}$", one can derive the predicted logits $\mathbf{o}^k_{\text{zs}}$ of K relation categories from the fill-in answers. In SGG, the *background* class is defined when a relation is outside $\mathcal{C}_r = [K]$. Predicting the *background* relation is challenging for f_{zs}: In pretraining phase, the model has not been exposed to the specific definition of *background*. Therefore, we rely solely on f_{sg} to produce the logits of *background* class:

$$\begin{cases} \mathbf{o}^k_{\text{zs}} = f_{\text{zs}}(z_i, z_j, \mathbf{x}_{i,j}) \in \mathbb{R}^K \\ [\mathbf{o}^0_{\text{sg}}, \mathbf{o}^k_{\text{sg}}] = f_{\text{sg}}(z_i, z_j, \mathbf{x}_{i,j}) \in \mathbb{R}^{K+1}, \end{cases} \quad (1)$$

The two branches' prediction reflect the label distribution of their training sets, leading to potential predicates bias in output logits if the target distribution differs. To address this, we conduct predicate debiasing using our **Lagrange-Multiplier Estimation** (LM Estimation) method along with logits adjustment, generating the debiased logits $\hat{\mathbf{o}}^k_{\text{zs}}$ and $\hat{\mathbf{o}}^k_{\text{sg}}$. The details are demonstrated in Section 3.3.

To mitigate the underrepresentation issue, we ensemble the debiased two branch to yield the final improved prediction, where we employ a **certainty-**

aware indicator to dynamically adjust the ensemble weights, which is discussed in Section 3.4.

3.3 Predicate Debiasing

Problem Definition. For each subject-object pair that has a *non-background* relation, we denote its relation label as $r \in \mathcal{C}_r$. Given the logits $\mathbf{o}^k$ of K *non-background* relation classes, the conditional probability on the training set $\mathcal{D}_{\text{tr}}$ is computed by:

$$P_{\text{tr}}(r|z_i, z_j, \mathbf{x}_{i,j}) = \text{softmax}(\mathbf{o}^k)(r),\ r \in \mathcal{C}_r \quad (2)$$

In our task, the training set $\mathcal{D}_{\text{tr}}$ can be either the SGG dataset $\mathcal{D}_{\text{sg}}$ or the pretraining dataset $\mathcal{D}_{\text{pt}}$, on which the SGG model f_{sg} and the zero-shot model f_{zs} are respectively trained.

In the evaluation phase, our goal is to estimate the target test probability P_{ta} rather than P_{tr}. By Bayes' Rule, we have the following:

$$P(r|z_i, z_j, \mathbf{x}_{i,j}) \propto P(z_i, z_j, \mathbf{x}_{i,j}|r) \cdot P(r) \quad (3)$$

where $P \in \{P_{\text{tr}}, P_{\text{ta}}\}$. The relation-conditional probability term $P(z_i, z_j, \mathbf{x}_{i,j}|r)$ can be assumed as the same in training and testing. By changing variables and omitting the constant factor, we have:

$$\frac{P_{\text{tr}}(r|z_i, z_j, \mathbf{x}_{i,j})}{P_{\text{tr}}(r)} = \frac{P_{\text{ta}}(r|z_i, z_j, \mathbf{x}_{i,j})}{P_{\text{ta}}(r)} \quad (4)$$

In a case where training distribution $P_{\text{tr}}(r)$ not equals to the target distribution $P_{\text{ta}}(r)$, known as label shift, the misalignment results in the model's predicted probability $P_{\text{tr}}(r|z_i, z_j, \mathbf{x}_{i,j})$ not equals to the actual test probability, $P_{\text{ta}}(r|z_i, z_j, \mathbf{x}_{i,j})$.

In our framework in Figure 2, f_{zs} is trained on $\mathcal{D}_{\text{pt}}$ and f_{sg} on $\mathcal{D}_{\text{sg}}$, whose training label distributions $P_{\text{tr}}(r)$ are $\pi_{\text{pt}} \in \mathbb{R}^K$ and $\pi_{\text{sg}} \in \mathbb{R}^K$, respectively. The prevalent evaluation metric, Recall, is designed to assess performance when the test label distribution $P_{\text{ta}}(r)$ is the same as the **training** distribution π_{sg}. In contrast, the mean recall metric seeks to evaluate performance in a **uniform** test distribution where $P_{\text{ta}}(r) = 1/K$. The $P_{\text{tr}}(r)$ and $P_{\text{ta}}(r)$ in each case can be summarized as follow:

$$P_{\text{tr}}(r) = \begin{cases} \pi_{\text{sg}}, & \text{if } f_{\text{sg}} \\ \pi_{\text{pt}}, & \text{if } f_{\text{zs}} \end{cases}, P_{\text{ta}}(r) = \begin{cases} \pi_{\text{sg}}, & \text{training} \\ \frac{1}{K}, & \text{uniform} \end{cases} \quad (5)$$

From Equation 5, we observe that the inequality $P_{\text{ta}}(r) \neq P_{\text{tr}}(r)$ holds in the following scenarios:

- For the SGG model f_{sg} with $P_{\text{tr}}(r) = \pi_{\text{sg}}$, a label shift will be revealed when the test target is a uniform distribution evaluated by mean Recall. In this scenario, the target distribution $P_{\text{ta}}(r) = 1/K$ diverges from the imbalanced distribution π_{sg} in $\mathcal{D}_{\text{sg}}$ shown in *top right* of Figure 1.
- For the zero-shot VLM f_{zs} with $P_{\text{tr}}(r) = \pi_{\text{pt}}$, the $P_{\text{ta}}(r) \neq P_{\text{tr}}(r)$ holds in both **training** and **uniform** targets. Firstly, the label distribution π_{pt} in the pretraining set $\mathcal{D}_{\text{pt}}$ differs from π_{sg}, resulting in $P_{\text{tr}}(r) \neq \pi_{\text{sg}}$ under the training-aligned target. Secondly, the imbalanced predicates distribution in $\mathcal{D}_{\text{pt}}$ also leads to $P_{\text{tr}}(r) \neq 1/K$ under the uniform target distribution.

Post-hoc Logits Adjustments. The first case, where $P_{\text{tr}}(r) = \pi_{\text{sg}}$ but $P_{\text{ta}}(r) = 1/K$, is a long-existing issue with many effective approaches proposed in SGG. However, existing methods are not feasible in the second case for their debiasing in the training stage, while the pretraining stage of f_{zs} are not accessible. A feasible debiasing method for already-trained models is the post-hoc logit adjustment (Menon et al., 2020). Denoting the initial prediction logits as $\mathbf{o}^k$ and the debiased logits as $\hat{\mathbf{o}}^k$, one can recast Equation 4 into a logits form:

$$\hat{\mathbf{o}}^k(r) = \mathbf{o}^k(r) - \log P_{\text{tr}}(r) + \log P_{\text{ta}}(r) \quad (6)$$

It suggests that given the target label distribution, the unbiased logits $\hat{\mathbf{o}}^k(r)$ can be obtained through a post-hoc adjustment on the initial prediction logits $\mathbf{o}^k(r)$, following the terms' value in Equation 5. While π_{sg} can be obtained simply by counting the label frequencies in $\mathcal{D}_{\text{sg}}$, π_{pt} is the predicates distribution hidden in the pretraining stage.

Lagrange Multiplier Estimation. To estimate π_{pt}, we proposed a novel method based on constrained optimization. Our initial step involves collecting all samples that have *non-background* relation labels $r \in \mathcal{C}_r$ from the training or validation set of $\mathcal{D}_{\text{sg}}$. Leveraging the collected data, our optimization objective is to solve the optimal π_{pt} that minimizes the cross-entropy loss between the adjusted logits $\hat{\mathbf{o}}^k_{\text{zs}}$ (following Equation 5 and 6 using π_{pt}) and the ground truth relation labels r.

Since the data are collected from $\mathcal{D}_{\text{sg}}$, we designate the term $P_{\text{ta}}(r)$ to π_{sg} to offset the interference of its label distribution and ensure the solved $P_{\text{tr}}(r) = \pi_{\text{pt}}$. This approach allows us to estimate π_{pt} by solving a constrained optimization problem, where we set the constraints to ensure the solved

π_{pt} representing a valid probability distribution:

$$\pi_{\text{pt}} = \underset{\pi_{\text{pt}}}{\operatorname{argmin}} R_{ce}(\mathbf{o}^k - \log \pi_{\text{pt}} + \log \pi_{\text{sg}},\ r),$$
$$s.t.\ \pi_{\text{pt}}(r) \geq 0,\ \text{for}\ r \in \mathcal{C}_r,\ \sum_{r \in \mathcal{C}_r} \pi_{\text{pt}}(r) = 1 \quad (7)$$

where R_{ce} is the cross-entropy loss. Equation 7 can be solved using the Lagrange-Multiplier method:

$$\pi_{\text{pt}} = \underset{\pi_{\text{pt}}}{\operatorname{argmin}} \max_{\lambda_r \geq 0, v} R_{ce} - \sum_r \lambda_r \pi_{\text{pt}}(r) + v(1 - \sum_r \pi_{\text{pt}}(r)) \quad (8)$$

After obtaining π_{pt} and π_{sg}, we can then apply the post-hoc logits adjustments for predicates debiasing following Equation 5 and 6, which produces two sets of unbiased logits from the initial prediction of f_{zs} and f_{sg}, denoted as $\hat{\mathbf{o}}^k_{\text{zs}}$ and $\hat{\mathbf{o}}^k_{\text{sg}}$.

Upon mitigating the predicates bias inside f_{zs}, we can leverage the model to address the underrepresentation issue in f_{sg}. From the debiased logits $\hat{\mathbf{o}}^k_{\text{zs}}$ and $\hat{\mathbf{o}}^k_{\text{sg}}$, we compute the probabilities towards $r \in \mathcal{C}_r$, where we adopt a τ-calibration outlined in (Kumar et al., 2022) to avoid over-confidence:

$$\begin{cases} \hat{P}_{\text{zs}}(r|z_i, z_j, \mathbf{x}_{i,j}) = \text{softmax}(\hat{\mathbf{o}}^k_{\text{zs}}/\tau)_r \\ \hat{P}_{\text{sg}}(r|z_i, z_j, \mathbf{x}_{i,j}) = \text{softmax}(\hat{\mathbf{o}}^k_{\text{sg}}/\tau)_r \end{cases} \quad (9)$$

3.4 Certainty-aware Ensemble

Considering that each model may better represent different samples, we compute a dynamic confidence score inspired by (Hendrycks and Gimpel, 2016) for each sample as its certainty in the two models, which determines the proportional weight W_{cer} of the two models in ensemble:

$$\begin{cases} \text{conf} = \max_{r \in \mathcal{C}_r} P(r|z_i, z_j, \mathbf{x}_{i,j}), P \in \{\hat{P}_{\text{zs}}, \hat{P}_{\text{sg}}\} \\ W_{\text{cer}} \propto\ \text{sigmoid}(\text{conf}_{\text{sg}} - \text{conf}_{\text{zs}}) \end{cases} \quad (10)$$

The weights are then used to obtain the ensembled prediction on $\mathcal{C}_r$:

$$P_{\text{ens}}(r|z_i, z_j, \mathbf{x}_{i,j}) = W_{\text{cer}} * \hat{P}_{\text{sg}}(r|z_i, z_j, \mathbf{x}_{i,j}) + (1 - W_{\text{cer}}) * \hat{P}_{\text{zs}}(r|z_i, z_j, \mathbf{x}_{i,j}) \quad (11)$$

Since f_{zs} cannot predict the *background* relation, we rely solely on f_{sg} to compute the *background* probability. Denoting $\mathbf{o}_{\text{sg}} = [\mathbf{o}^0_{\text{sg}}, \mathbf{o}^k_{\text{sg}}]$ as the initial

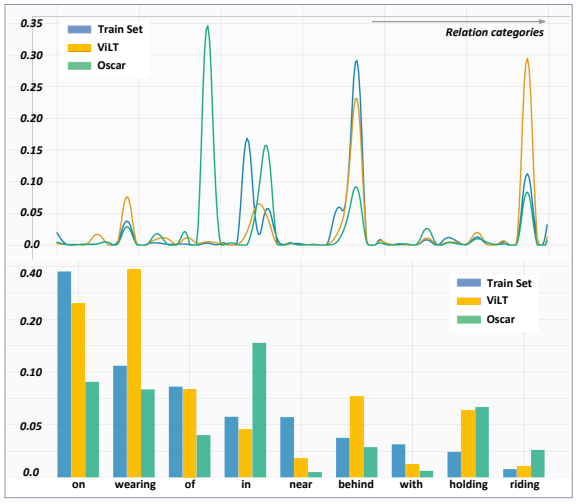

Figure 3: The relation label distributions on Visual Genome. The *upper* figure illustrates the distribution across all classes, while the *lower* one shows the probability distribution on some typical categories. *Train Set*: The class distribution π_{sg} in training set. *ViLT* and *Oscar*: The estimated distribution π_{pt} using LM Estimation in the two pre-training stages.

logits predicted by f_{sg} without debiasing (Equation 1), the *background* and *non-background* probability can be calculated by softmax function:

$$\begin{cases} P_{\text{sg}}(y \neq 0|z_i, z_j, \mathbf{x}_{i,j}) = 1 - \text{softmax}(\mathbf{o}_{\text{sg}})_0 \\ P_{\text{sg}}(y = 0|z_i, z_j, \mathbf{x}_{i,j}) = \text{softmax}(\mathbf{o}_{\text{sg}})_0 \end{cases} \quad (12)$$

Finally, the ensembled prediction on $\mathcal{C}_e$ is:

$$P_{\text{ens}}(y|z_i, z_j, \mathbf{x}_{i,j}) = [P_{\text{sg}}(y = 0|z_i, z_j, \mathbf{x}_{i,j}), P_{\text{sg}}(y \neq 0|z_i, z_j, \mathbf{x}_{i,j}) \cdot P_{\text{ens}}(r|z_i, z_j, \mathbf{x}_{i,j})] \quad (13)$$

which serves as the final representation-improved prediction of our proposed framework.

3.5 Summary

We integrate VLMs to mitigate the underrepresentation challenge inherent to SGG, where we propose the novel LM Estimation to approximate the unattainable pretraining distribution of predicates, π_{pt}, and conduct predicate debiasing for each model. Unlike previous SGG methods that are optimized for one target distribution per training, our method enables seamlessly adaptation between different targets without cost, outperforming existing SGG approaches under each target distribution.

4 Experiment

We conduct comprehensive experiments on SGG to assess our efficacy. In Section 4.2, we show

Models	Predicate Classification			Scene Graph Classification		
	mRecall@20	mRecall@50	mRecall@100	mRecall@20	mRecall@50	mRecall@100
VTransE(Zhang et al., 2017)	13.6	17.1	18.6	6.6	8.2	8.7
SG-CogTree(Yu et al., 2020)	22.9	28.4	31.0	13.0	15.7	16.7
BGNN(Li et al., 2021)	-	30.4	32.9	-	14.3	16.5
PCPL(Yan et al., 2020)	-	35.2	37.8	-	18.6	19.6
Motifs-Rwt(Zellers et al., 2018)	-	33.7	36.1	-	17.7	19.1
Motifs-GCL(Dong et al., 2022)	30.5	36.1	38.2	18.0	20.8	21.8
VCTree-TDE(Tang et al., 2020)	18.4	25.4	28.7	8.9	12.2	14.0
VCTree-GCL(Dong et al., 2022)	31.4	37.1	39.1	**19.5**	22.5	23.5
PENET-Rwt†(Zheng et al., 2023)	31.0	38.8	40.7	18.9	22.2	23.5
Oscar ft-la	30.4	38.4	41.3	17.9	22.6	23.8
Oscar ft-la + Ours	31.2(+0.8)	39.4(+1.0)	42.7(+1.4)	18.3(+0.4)	23.4(+0.8)	25.0(+1.2)
ViLT ft-la	31.2	40.5	44.5	17.4	22.5	24.3
ViLT ft-la + Ours	**32.3(+1.1)**	**42.3(+1.8)**	**46.5(+2.0)**	17.9(+0.5)	**23.5(+1.0)**	**25.5(+1.2)**
PENET-Rwt†	31.4	38.8	40.7	18.9	22.2	23.5
PENET-Rwt + Ours	31.8(+0.4)	39.9(+1.1)	42.3(+1.6)	19.2(+0.3)	23.0(+0.8)	24.5(+1.0)

Table 1: The mean Recall results on Visual Genome comparing with state-of-the-art models and debiasing methods. The results and performance gain applying our method is below the row of corresponding baseline. *ft*: The model is fine-tuned on Visual Genome. *la*: The prediction logits is debiased by logits adjustment with π_{sg}. †: Due to the absence of part of the results, we re-implement by ourselves.

our significant performance improvement through a comparative analysis with previous methods. Section 4.3 provides an illustrative analysis of the predicates distribution estimated by our LM Estimation. Subsequently, Section 4.4 offers an ablation study, analysing the contribution of individual components in our design to the overall performance.

4.1 Experiment Settings

Datasets. The Visual Genome (VG) dataset consists of 108,077 images with average annotations of 38 objects and 22 relationships per image. For Visual Genome, we adopted a split with 108,077 images focusing on the most common 150 object and 50 predicate categories, allocating 70% for training and 30% for testing, alongside a validation set of 5,000 images extracted from the training set.

Evaluation Protocol. For the Visual Genome dataset, we focus on two key sub-tasks: Predicate Classification (PredCls) and Scene Graph Classification (SGCls). We skip the Scene Graph Detection (SGDet) here and provide a discussion in supplementary, considering its substantial computational demands when employing VLMs and limited relevance to our method's core objectives. Our primary evaluation metrics are Recall@K and mean Recall@K (mRecall@K). Additionally, we propose another task of relation classification that calculates the top-1 predicate accuracy (Acc) for samples labeled with *non-background* relations, where we focus on the ability of model on predicting the relation given a pair of objects in the scene.

Baselines and Implementation. Here we utilize two prominent zero-shot vision-language models, ViLT (Kim et al., 2021) and Oscar (Li et al., 2020), as f_{zs}. For the task-specific branch f_{sg}, we employ three baseline models trained in SGG: (1) To explore the fine-tuning performance of VLMs on SGG, we fine-tune ViLT and Oscar using the PredCls training data and establish them as our first two baselines. (2) To show our methods' compatibility with existing SGG models, we undertake PENET (Zheng et al., 2023), a cutting-edge method with superior performance, as our third baseline. In our ensemble strategy, we explore three combinations: "fine-tuned ViLT + zero-shot ViLT", "fine-tuned Oscar + zero-shot Oscar", and "PENET + zero-shot ViLT", where each model is debiased by our methods. Following previous settings, an independently trained Faster R-CNN is attached to the front of each VLM model for object recognition. During pre-training, both ViLT and Oscar employ two main paradigms: Masked Language Modeling (MLM) and Visual Question Answering (VQA). In MLM, tokens in a sentence can be replaced by `[MASK]`, with the model predicting the original token using visual and language prompts. In VQA, the model, given a question and visual input, predicts an answer via an MLP classifier using the `[CLS]` token. For our task, we use MLM for the fixed branch f_{zs} with the prompt "z_i is `[MASK]` z_j." and VQA for fine-tuning f_{sg}, where we introduce a MLP with the query "`[CLS]` what is the relationship between the z_i and the z_j?", where the embedding of `[CLS]` token is forwarded to the

Models	Predicate Classification			Scene Graph Classification		
	Recall@20	Recall@50	Recall@100	Recall@20	Recall@50	Recall@100
KERN(Chen et al., 2019)	-	65.8	67.6	-	36.7	37.4
R-CAGCN(Yang et al., 2021)	60.2	66.6	68.3	35.4	38.3	39.0
GPS-Net(Lin et al., 2020)	60.7	66.9	68.8	36.1	39.2	40.1
VTransE(Zhang et al., 2017)	59.0	65.7	67.6	35.4	38.6	39.4
VCTree(Tang et al., 2019)	60.1	66.4	68.1	35.2	38.1	38.8
MOTIFS(Zellers et al., 2018)	59.5	66.0	67.9	35.8	39.1	39.9
SGGNLS(Zhong et al., 2021)	58.7	65.6	67.4	36.5	40.0	40.8
RU-Net(Lin et al., 2022b)	**61.9**	68.1	70.1	**38.2**	41.2	42.1
PENET†(Zheng et al., 2023)	61.7	**68.2**	**70.1**	37.9	**41.3**	**42.3**
Oscar ft	59.1	65.7	67.6	36.7	40.3	41.3
Oscar ft + Ours	60.5(+1.4)	67.4(+1.8)	69.3(+1.7)	37.3(+0.6)	41.4(+1.1)	42.3(+1.0)
ViLT ft	57.1	65.7	68.4	34.9	40.2	41.8
ViLT ft + Ours	58.0(+0.9)	66.7(+1.0)	69.8(+1.4)	35.3(+0.4)	41.2(+1.0)	**42.9(+1.1)**
PENET†	61.7	68.2	70.1	37.9	41.3	42.3
PENET + Ours	**62.0(+0.3)**	**69.0(+0.8)**	**71.1(+1.0)**	**38.1(+0.2)**	**41.8(+0.5)**	42.9(+0.6)

Table 2: The Recall results on Visual Genome dataset comparing with state-of-the-art models and debiasing methods. The results and performance gain applying our method is below the row of corresponding baseline. *ft*: The model is fine-tuned on Visual Genome. †: Due to the absence of part of the results, we re-implemented by ourselves.

MLP classification head.

4.2 Efficacy Analysis

To assess the efficacy of our method, in this section, we compare our method with recent studies through a detailed result analysis on Visual Genome. The Recall and mean Recall results are presented in Table 2, which showcases a performance comparison with a variety of cutting-edge models and debiasing methods. We ensure to compare against previous methods under their best-performance metric. For baseline models without debiasing strategies, we compare with their superior Recall metrics and exclude their lower mean Recall performances. Similarly, for the debiased SGG models, we only focus on their mean Recall outcomes.

Baseline Performance. Our analysis begins with the three f_{sg} baselines: fine-tuned ViLT, fine-tuned Oscar, and PENET. Specifically, for scenarios where the desired target is a uniform distribution assessed by mean Recall, we apply the post-hoc logits adjustment to the two fine-tuned baselines following Equations 5 and 6. For PENET, we implement a reweighting loss strategy (PENET-Rwt) following (Zheng et al., 2023) to train a debiased version tailored for the uniform target distribution, which achieved optimal performance.

Our main experiment results are presented in Table 1 and Table 2. As shown in Table 2, without task-specific designs, the two fine-tuned VLMs fall behind the SGG models on Recall and scored 67.6 and 68.4 on R@100, while PENET takes the lead. However, as shown in Table 1, when evaluated under the uniform target distribution and adjusted using simple post-hoc logits adjustment, the fine-tuned VLMs surpass all the cutting-edge debiased SGG models in mean Recall, achieving 41.3 and 44.5 of mR@100.

Our Improvements. Subsequently, we employ our certainty-aware ensemble to integrate debiased zero-shot VLMs f_{zs} into the f_{sg} baselines, where each f_{zs} is debiased by our LM Estimation. In Table 2, for each f_{sg} baseline, we observed a notable performance boost after applying our methods (+1.4 / + 2.0 / + 1.6 in mR@100 and +1.7 / +1.4 / + 1.0 in R@100). In both mRecall and Recall, our methods achieve the best performance (46.5 on mR@100 and 71.1 on R@100), while the improvement on mean Recall is particularly striking and surpasses the gains observed on Recall (+1.4/+2.0/+1.6 vs. +1.7/+1.4/+1.0). The results show that our methods achieve a significant improvement in each baseline, achieving the best performance compared to all existing methods.

Our results indicate the effectiveness of our methods, leading to a marked boost in performance. Moreover, the improvement in PENET baselines shows the adaptability of our method to existing SGG-specialized models. In addition, we observe that our representation improvements leads to a more significant gain in mean recall than in recall, suggesting the underrepresentation problem is more common in tail relation classes.

Models	All mAcc		All Acc		Unseen mAcc		Unseen Acc	
	Initial	Debiased	Initial	Debiased	Initial	Debiased	Initial	Debiased
ViLT-ft	46.53		68.92		14.98		17.72	
ViLT-zs	21.88	37.42	57.15	67.09	8.99	16.92	18.81	20.93
ViLT-ens	46.86	48.70	68.95	70.75	15.66	20.07	20.01	21.73
Ens. Gain	+0.33	+2.17	+0.03	+1.83	+0.68	+5.09	+2.29	+4.01
Oscar-ft	41.99		67.16		13.85		18.01	
Oscar-zs	17.18	33.96	45.78	57.31	6.68	16.01	19.11	20.05
Oscar-ens	42.02	44.28	67.77	69.03	14.83	19.56	20.97	22.08
Ens. Gain	+0.03	+3.29	+0.61	+1.87	+0.98	+5.71	+2.96	+4.07

Table 3: Top-1 accuracy and class-wise mean accuracy of relation classification on Visual Genome. *All*: The test results for all triplets with *non-background* relation labels. *Unseen*: The test results for triplets that are absent from the training set. *Initial*: The initial zero-shot VLMs without debiasing. *Debiased*: The zero-shot VLMs after debiasing using our **LM Estimation**. *ens*: Ensemble of the fine-tuned VLMs and *Initial* or *Debiased* zero-shot model. *Ens. Gain*: the performance gain of ensemble compared to the fine-tuned model.

4.3 Estimated Distribution Analysis

In Figure 3, we depict the predicate distributions of zero-shot ViLT and Oscar solved by LM Estimation, comparing them with the distribution in VG training set. The *upper* chart in Figure 3 depicts the distributions across all relations, where we find that all three distributions exhibit a significant imbalance. Furthermore, we extract the distribution of typical relations in the *lower* chart, where we see a substantial discrepancy among the three distributions. This variation affirms the two scenarios of $P_{\text{ta}}(r) \neq P_{\text{tr}}(r)$ discussed in Section 3.3, precluding the direct application of zero-shot VLMs without debiasing, indicating the necessity of our LM Estimation and subsequent debiasing method.

4.4 Ablation Study

In this section, we conduct an ablation study on Visual Genome dataset. Initially, we assess the effectiveness of our LM Estimation in addressing the predicates bias of zero-shot VLMs. Furthermore, we evaluate the capability of our method to enhance representation by focusing on the unseen triplets, which are entirely absent during training.

To precisely evaluate the performance in relation recognition and eliminate any influence from the *background* class, we require the model to perform relation classification exclusively on samples labeled with *non-background* relations. Subsequently, we calculate the top-1 accuracy (Acc) and class-wise mean accuracy (mAcc) as new metrics to accurately gauge the model's effectiveness in this context. Our findings are comprehensively detailed in Table 3, which details on two sample splits: one encompassing all triplets and the other exclusively focusing on unseen triplets. For each splits, we examine the performance of the two fine-tuned VLMs, f_{sg}, their initial and debiased zero-shot models, f_{zs}, and the ensemble of corresponding models.

Predicate Debiasing. In Section 3.3, we introduce our LM Estimation method for predicate debiasing. Here, we further evaluate the efficacy of our debiasing. We initially analysis on the relation classification accuracy of the zero-shot VLMs before and after debiasing. As presented in Table 3 (the *ViLT-zs* and *Oscar-zs* rows), without debiasing, the accuracies of initial predictions are lower either in all triplets or unseen triplets. However, after debiasing through LM Estimation, there is a notable enhancement in the zero-shot performance. For unseen triplets, the debiased zero-shot VLMs even surpass the performance of their fine-tuned counterparts, suggesting our method effectively addresses the predicate bias and smoothly adapts the pretraining knowledge to the SGG task.

Furthermore, from the ensemble performance in Table 3 (the *ViLT-ens* and *Oscar-ens* rows), we notice that ensembling the initial f_{zs} hardly improves the performance, only achieving a slight gain of +0.33/+0.03 on all triplets and +0.68/+2.29 on unseen triplets. In contrast, ensembling the debiased f_{zs} achieves a significantly more pronounced improvement, achieving +2.17/+1.83 gain on all triplets and +5.09/+4.01 on unseen triplets.

To keep consistent with previous settings, we present the Recall and mean Recall ablation results in Table 4. We observe a substantial improvement in both mean Recall and Recall when ensembling with our debiased zero-shot VLMs (the highlighted row in each group), while directly ensembling the initial zero-shot VLMs even harm to the performance (the *middle* row in each group). These results starkly underlines the necessity and efficacy of our LM Estimation in predicate debiasing.

Models	mR@20	mR@50	mR@100
ViLT-ft	31.2	40.5	44.5
ViLT-ens (Initial)	30.9(-0.3)	40.5(+0.0)	44.6(+0.1)
ViLT-ens (Debiased)	32.3(+0.9)	42.3(+1.8)	46.5(+2.0)
Oscar-ft	30.4	38.4	41.3
Oscar-ens (Initial)	30.3(-0.1)	38.5(+0.1)	41.6(+0.3)
Oscar-ens (Debiased)	31.2(+0.8)	39.4(+1.0)	42.7(+1.4)
Models	R@20	R@50	R@100
ViLT-ft	57.1	65.7	68.4
ViLT-ens (Initial)	56.9(-0.2)	65.7(+0.0)	68.8(+0.4)
ViLT-ens (Debiased)	58.0(+0.9)	66.7(+1.0)	69.8(+1.4)
Oscar-ft	59.1	65.7	67.6
Oscar-ens (Initial)	59.2(+0.1)	65.9(+0.2)	67.9(+0.3)
Oscar-ens (Debiased)	60.5(+1.4)	67.4(+1.7)	69.3(+1.7)

Table 4: The mean Recall and Recall ablation results on Visual Genome. *Initial*: The initial zero-shot VLMs without debiasing. *Debiased*: The zero-shot VLMs after predicates debiasing. *ens*: Ensemble of the fine-tuned VLMs and *Initial* or *Debiased* zero-shot model.

Representation Enhancement. To validate the enhancement of representation, we specifically examine the samples labeled with unseen triplets. These triplets are present in the test set but absent from the training set, which is the worst tail distribution in the underrepresentation issue.

Table 3 reveals that, across all triplets, the accuracies of both zero-shot VLMs (f_{zs}) fall short of their fine-tuned counterparts (f_{sg}). For example, the debiased zero-shot Oscar model achieves 33.96/57.31 of mAcc/Acc, which are lower than the fine-tuned Oscar (41.99/67.16). However, within the subset of unseen triplets, the debiased zero-shot f_{zs} outperforms the fine-tuned f_{sg}: The debiased zero-shot Oscar achieves 16.01/20.05 of mAcc/Acc, outperforming the fine-tuned model (13.85/18.01).

These findings substantiate our hypothesis that zero-shot models, with their pretraining knowledge fully preserved, are better at handling underrepresented samples compared to SGG-specific models. This advantage is particularly evident in the context of unseen triplets, where comprehensive pretraining knowledge of zero-shot models confers a significant performance benefit.

Moreover, we find that the gain of ensemble is significantly higher for unseen triplets (Debiased ViLT: +5.09/+4.01, Debiased Oscar: +5.71/4.07) than for all triplets (Debiased ViLT: +2.17/+1.83, Debiased Oscar: +3.29/1.87). This indicates that the underrepresented samples are improved much more than the well-represented samples, receiving higher gains than average. Considering the proportion of unseen triplets in all triplets, we infer the overall performance gain mainly comes from the improvement on unseen triplets. Since unseen triplets composing the worst case of underrepresentation, their performance gain can confirm our enhancement on representation.

5 Conclusion

In conclusion, our study has made significant strides in efficiently and effectively integrate pretrained VLMs to SGG. By introducing the novel **LM Estimation**, we effectively mitigate the predicate bias inside pre-trained VLMs, allowing their comprehensive knowledge to be employed in SGG. Besides, our **certainty-aware ensemble** strategy, which ensembles the zero-shot VLMs with SGG model, effectively addresses the underrepresentation issue and demonstrates a significant improvement in SGG performance. Our work contributes to the field of SGG, suggesting potential pathways for reducing language bias of pretraining and leverage them in more complex language tasks.

6 Limitation

Though our methods does not require any training, comparing with original f_{sg}, our ensemble framework still adds computational cost from f_{zs}'s inference. This inference can be costly in an extreme case that one scene has too many objects to predict their relations. Besides, even after we solve the word bias inside VLMs, the final ensemble performance relies highly on the pre-training quality, which requires the f_{zs} to be pre-trained on comprehensive data to improve SGG's representation. Another limitation arises from the forwarding pattern in VLM, where we adopt a pair-wise forwarding that taking a pair of objects along with their image region and text prompt. In this way, each possible object pair requires an entire forwarding of VLM. This process is rapid when the object is certainly detected. However, in the scenario of Scene Graph Detection, the large amounts of proposals can bring unavoidable time cost to our pipeline. We provide a more detailed discussion in appendix.

References

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

Xingning Dong, Tian Gan, Xuemeng Song, Jianlong Wu, Yuan Cheng, and Liqiang Nie. 2022. 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.

Kaifeng Gao, Long Chen, Hanwang Zhang, Jun Xiao, and Qianru Sun. 2023. Compositional prompt tuning with motion cues for open-vocabulary video relation detection. *arXiv preprint arXiv:2302.00268*.

Jiuxiang Gu, Handong Zhao, Zhe Lin, Sheng Li, Jianfei Cai, and Mingyang Ling. 2019. Scene graph generation with external knowledge and image reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1969–1978.

Tao He, Lianli Gao, Jingkuan Song, and Yuan-Fang Li. 2022. Towards open-vocabulary scene graph generation with prompt-based finetuning. In *European Conference on Computer Vision*, pages 56–73. Springer.

Dan Hendrycks and Kevin Gimpel. 2016. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations*.

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

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73.

Ananya Kumar, Tengyu Ma, Percy Liang, and Aditi Raghunathan. 2022. Calibrated ensembles can mitigate accuracy tradeoffs under distribution shift. In *Uncertainty in Artificial Intelligence*, pages 1041–1051. PMLR.

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. 2020. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision*, 128(7):1956–1981.

Lin Li, Long Chen, Yifeng Huang, Zhimeng Zhang, Songyang Zhang, and Jun Xiao. 2022a. 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.

Lin Li, Jun Xiao, Guikun Chen, Jian Shao, Yueting Zhuang, and Long Chen. 2023. Zero-shot visual relation detection via composite visual cues from large language models. *arXiv preprint arXiv:2305.12476*.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

Rongjie Li, Songyang Zhang, Bo Wan, and Xuming He. 2021. 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.

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

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*, pages 121–137. Springer.

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

Xin Lin, Changxing Ding, Yibing Zhan, Zijian Li, and Dacheng Tao. 2022a. Hl-net: Heterophily learning network for scene graph generation. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19476–19485.

Xin Lin, Changxing Ding, Jing Zhang, Yibing Zhan, and Dacheng Tao. 2022b. Ru-net: Regularized unrolling network for scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19457–19466.

Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. 2020. Long-tail learning via logit adjustment. *arXiv preprint arXiv:2007.07314*.

Di Qi, Lin Su, Jia Song, Edward Cui, Taroon Bharti, and Arun Sacheti. 2020. Imagebert: Cross-modal pre-training with large-scale weak-supervised image-text data. *arXiv preprint arXiv:2001.07966*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2019. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*.

Xudong Sun, Pengcheng Wu, and Steven CH Hoi. 2018. Face detection using deep learning: An improved faster rcnn approach. *Neurocomputing*, 299:42–50.

Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. 2020. Unbiased scene graph generation from biased training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3716–3725.

Kaihua Tang, Hanwang Zhang, Baoyuan Wu, Wenhan Luo, and Wei Liu. 2019. 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.

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

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

Gengcong Yang, Jingyi Zhang, Yong Zhang, Baoyuan Wu, and Yujiu Yang. 2021. Probabilistic modeling of semantic ambiguity for scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12527–12536.

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. 2022. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*.

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

Qifan Yu, Juncheng Li, Yu Wu, Siliang Tang, Wei Ji, and Yueting Zhuang. 2023. Visually-prompted language model for fine-grained scene graph generation in an open world. *arXiv preprint arXiv:2303.13233*.

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

Hanwang Zhang, Zawlin Kyaw, Shih-Fu Chang, and Tat-Seng Chua. 2017. Visual translation embedding network for visual relation detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5532–5540.

Yong Zhang, Yingwei Pan, Ting Yao, Rui Huang, Tao Mei, and Chang-Wen Chen. 2023. Learning to generate language-supervised and open-vocabulary scene graph using pre-trained visual-semantic space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2915–2924.

Long Zhao, Liangzhe Yuan, Boqing Gong, Yin Cui, Florian Schroff, Ming-Hsuan Yang, Hartwig Adam, and Ting Liu. 2023. Unified visual relationship detection with vision and language models. *arXiv preprint arXiv:2303.08998*.

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

Yiwu Zhong, Jing Shi, Jianwei Yang, Chenliang Xu, and Yin Li. 2021. Learning to generate scene graph from natural language supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1823–1834.

A More Theoretical Justifications

In the main paper, we introduce the post-hoc logits adjustment methods (Menon et al., 2020) for label debiasing, which is first proposed in long-tail classification. In the main paper, we skipped part of the derivation due to the limit of length. Here, we provide a detailed derivation for easier understanding.

Taking $(z_i, z_j, \mathbf{x}_{i,j})$ as input for a subject-object pair, the conditional probability for the relations is $P(r|z_i, z_j, \mathbf{x}_{i,j})$. From the Bayes' Rule, the conditional probability can be expressed as:

$$P(r|z_i, z_j, \mathbf{x}_{i,j}) = \frac{P(z_i, z_j, \mathbf{x}_{i,j}|r)P(r)}{P(z_i, z_j, \mathbf{x}_{i,j})} \quad (14)$$

We further denote the empirical probability fitted to the training set as P_{tr} and the target test probability as P_{ta}. We further rewrite Equation 14 with the two probabilities as:

$$P_{\text{tr}}(r|z_i, z_j, \mathbf{x}_{i,j}) = \frac{P_{\text{tr}}(z_i, z_j, \mathbf{x}_{i,j}|r)P_{\text{tr}}(r)}{P_{\text{tr}}(z_i, z_j, \mathbf{x}_{i,j})} \quad (15)$$

$$P_{\text{ta}}(r|z_i, z_j, \mathbf{x}_{i,j}) = \frac{P_{\text{ta}}(z_i, z_j, \mathbf{x}_{i,j}|r)P_{\text{ta}}(r)}{P_{\text{ta}}(z_i, z_j, \mathbf{x}_{i,j})} \quad (16)$$

Then let us look into each term. Firstly, the $P(z_i, z_j, \mathbf{x}_{i,j})$ is irrelavant with r and thus has no effect on the relation label bias. Therefore, the numerator term can be replaced by a constant C and omitted in further computation. Secondly, when focusing on the label bias, according to the prevalent **label-shift hypothesis** proposed in long-tail classification, one can assume $P(z_i, z_j, \mathbf{x}_{i,j}|r)$ to be the

same in the training and testing domains. Based on this equality, we connect the two probabilities by:

$$\frac{P_{\text{tr}}(r|z_i, z_j, \mathbf{x}_{i,j})}{P_{\text{tr}}(r)} \cdot C_{\text{tr}} = \frac{P_{\text{ta}}(r|z_i, z_j, \mathbf{x}_{i,j})}{P_{\text{ta}}(r)} \cdot C_{\text{te}} \tag{17}$$

Taking the logarithm form for both sides, we derive the final form of post-hoc logits adjustments (Menon et al., 2020):

$$\log P_{\text{ta}}(r|z_i, z_j, \mathbf{x}_{i,j}) = \log P_{\text{tr}}(r|z_i, z_j, \mathbf{x}_{i,j}) - \log P_{\text{tr}}(r) + \log P_{\text{ta}}(r) + \log \frac{C_{\text{tr}}}{C_{\text{te}}} \tag{18}$$

In our main paper, the last term of constant is omitted since the softmax function will naturally erase any constant term that irrelavant to r. Given the target distribution P_{ta}. From Equation 18, by taking softmax operation on both sides, we can derive:

$$P_{\text{ta}}(r|z_i, z_j, \mathbf{x}_{i,j}) = \text{softmax}(\log P_{\text{tr}}(r|z_i, z_j, \mathbf{x}_{i,j}) - \log P_{\text{tr}}(r) + \log P_{\text{ta}}(r)) \tag{19}$$

After adjusting using our strategy, the final predicted label is determined by an argmax operation:

$$r = \underset{r \in \mathcal{C}_r}{\text{argmax}}(\text{softmax}(\log P_{\text{tr}}(r|z_i, z_j, \mathbf{x}_{i,j}) - \log P_{\text{tr}}(r) + \log P_{\text{ta}}(r))) \tag{20}$$

Then from Equation 19, we can rewrite Equation 20 as:

$$r = \underset{r \in \mathcal{C}_r}{\text{argmax}}(P_{\text{ta}}(r|z_i, z_j, \mathbf{x}_{i,j})) \tag{21}$$

it is called a **Bayes optimal classifier**. According to the definition of Bayes optimal classifier, on average no other classifier using the same hypothesis and prior knowledge can outperform it. Thus, when considering only label bias, our strategy is not only effective, but also optimal among all adjustments.

B More Experiment Analysis

B.1 Scene Graph Detection

In our main paper, we skipped the SgDet sub-task, considering its substantial computational demands when employing VLMs and limited relevance to our method's core objectives. In this section, we provides a discussion and a brief corresponding experiments results.

Existing SGG models usually employs a Faster R-CNN (Sun et al., 2018) detector and fix the number of generated proposals to be 80 per image for a fair comparison. However, unlike the existing relation recognition networks that processes all pairs of proposals in an image simutaniously, the attention module in VLMs requires a one-by-one pair as input. In this case, inferencing one image requires 80×80 times of forwarding.

This huge inference cost make it less practical to compare with existing methods under the current prevalent settings. However, it does not suggest using VLMs in SGG is meaningless. We strongly believe that the main concern of SGG task is to correctly recognize the relation given a pairs of objects, instead of the object detection, given the fact that the detector could be trained separately while achieving the same good performance. And by equipping with more efficient and effective detectors, the performance in Scene Graph Detection and Scene Graph Classification should be closed to Predicate Classification.

B.2 Analysis on Tail Categories

In this section, we conducted an additional experiment to demonstrate the performance enhancement for tail relation classes. We divided the relation categories into three splits, *frequent*, *medium*, and *rare*, based on the frequency in the training set. Subsequently, we evaluated and reported the ensemble gain on mean Recall@100 for each split brought by our methods. We opted for mean Recall@100 as the metric due to its superior representation of rare relations and reduced susceptibility to background class interference. Across all three baselines, we observed a substantial improvement in performance for rare relation categories, which confirms our hypothesis that the underrepresentation issue is more severe in rare relation classes.

Ensemble Gain on mRecall@100.			
Models	frequent	medium	rare
ViLT ft-la + Ours	+0.12	+1.78	+4.13
Oscar ft-la + Ours	+0.04	+1.04	+3.15
PENET + Ours	+0.06	+1.27	+3.49

Table 5: The performance gain of mRecall@100 on PredCls sub-task achieved by our methods compared with each baseline, where the rare categories achieve significantly higher improvement.

C More Details of Implementation

This section shows more details of our implementation. In existing models designed for SGG, the object detector is attached in front of the relation recognition network and jointly trained with the objectives of SGG tasks. However, when fine-tuning

VLMs on SGG tasks, this paradigm could be time-consuming and less flexible, given the higher training cost of VLM comparing with existing models.

Therefore, we decide to take the Faster R-CNN detector out and train it separately without the main network. This implementation is proved to be effective when we take the detector out of PENET (Zheng et al., 2023) and train it separately with the PENET relation network. We observe that the independently trained detector achieved the same performance with that jointly trained with the PENET. Hence, all fine-tuned VLMs in this paper used a separately-trained Faster R-CNN detector. In the fine-tuning stage on Visual Genome, we employ two different paradigms for ViLT (Kim et al., 2021) and Oscar (Li et al., 2020) for a more general comparison. We freeze the ViLT backbone while training the MLP head for 50 epochs. In another way, we use an end-to-end fine-tuning for 70k steps on Oscar. We keep the fine-tuning cost comparable to the existing SGG models, which ensures its practical feasibility.

Why don't we debias on the triplets' distribution instead of the relation words distribution? In the paper, we declare the relation words bias caused by different frequency of relation labels. And the underrepresentation issue caused by different representation level of samples. One can infer that the representation level is largely effect by the frequency of triplets. In other words, the samples of frequent triplets are usually better represented in training compared with those samples of rare triplets. Therefore, one intuitive thinking is to debias directly on the triplets' distribution by substracting $\log P(z_i, z_j, r)$ instead of the relation words distribution $\log P(r)$. This thought is indeed the most throughly debiasing strategy. However, one need to consider that the conditional prior of $\log P(r|z_i, z_j)$ could largely help the prediction of relationship (Tang et al., 2020). For example, in natural world, the relation between a "man" and a "horse" is more likely to be "man *riding* horse" than "man *carrying* horse". Directly debiasing on the triplets' distribution would erase all these helpful conditional priors, resulting in a drastically drop in performance.

D Other Discussions

Question 1: Is our improvement from representation improvement or simply parameter increase from ensembled VLMs? Because of the predicates biases in pretraining data, integrating large pretrained models does not guarantee improvement. In Table 2 of the main paper, we showed that ensembling the original VLMs without debiasing cannot bring any improvements. Only by integrating the VLM debiased by our LM Estimation can enhancements be brought.

By integrating our debiased VLM, the underrepresentation issue is alleviated since underrepresented samples are improved much more than well-represented samples. In Table 2 in the main paper, we show that unseen triplets are improved higher than all triplets' average. Integrating our debiased VLMs indeed brings a slight overall improvement, but most are from addressing the representation improvement.

Question 2: Is it fair for us to use distinct P_{ta} to measure Recall and mRecall and compare with existing methods? Unlike previous methods in SGG, our framework accepts a user-specified target distributions P_{ta} as input. In SGG settings, measuring both Recall and mRecall is to evaluate under two distinct test distributions, as discussed in Section 3.3 of our main paper. For our method, using the same P_{ta} under these two distinct distributions will input a wrong distribution P_{ta} that is far from the actual target. This goes against our original intention.

Previous methods are measured by both metrics without any change because once trained, unless by time-costing re-training, they cannot be transferred from one target distribution P_{ta} to another P'_{ta}. However, our method achieves this transfer instantaneously by simply $+\log\left(P'_{\text{ta}}/P_{\text{ta}}\right)$ to the logits. So it is fair to compare with previous methods since our transfer adds no extra time cost.

Question 3: Is underrepresentation issue a specific characteristic problem for SGG? The problem of this inadequate sample representation is a typical and specific characteristics of SGG and is far more severe than that in other related fields, like long-tailed classification in Computer Vision. In SGG, a sample's representation includes two objects' attributes and their high-level relationship. Due to this unique complexity, it is extremely hard for SGG datasets to adequately represent all triplets combinations. For instance, there are 375k triplets combinations in Visual Genome (Krishna et al., 2017), much more than the label sets of any classification dataset in Computer Vision. This inevitably leads to the majority of triplets having only a few samples in training.