Scene Graph Generation Based on Language Model Assistance

1st Yangyang Li
School of Artificial Intelligence
Xidian University
Xian, China
yyli@xidian.edu.cn

2nd Rongxia Ren
School of Artificial Intelligence
Xidian University
Xian, China
23171214575@stu.xidian.edu.cn

3rd Xuanting Hao
School of Artificial Intelligence
Xidian University
Xian, China
xthao@stu.xidian.edu.cn

4th Yunhui Zhang
School of Artificial Intelligence
Xidian University
Xian, China
yunhuizhang001@163.com

5th Ronghua Shang
School of Artificial Intelligence
Xidian University
Xian, China
rhshang@mail.xidian.edu.cn

6th Licheng Jiao
School of Artificial Intelligence
Xidian University
Xian, China
lchjiao@mail.xidian.edu.cn

***Abstract*—Scene graph generation suffers from severe bias in long-tail and zero-shot predicate prediction. We propose the Language Model-Aided Network (LM-ANet), which repurposes a completely frozen 110M-parameter BERT-base as a zero-shot predicate predictor without any fine-tuning or task-specific classifier. It achieves this by projecting visual tokens into BERT's embedding space and casting predicate classification as cloze-style masked language modeling with subject-object-aware prompts, followed by prototype-based matching. On Visual Genome, LM-ANet achieves 23.8% zero-shot Recall@100 on PredCls (improving the prior best by 2.9 percentage points), 7.1% on SGCls, and 6.8% on SGDet, while obtaining mean Recall@100 of 23.4% on PredCls and 14.9% on SGCls—both highly competitive with recent methods. Using a completely frozen 110M-parameter BERT-base without any fine-tuning of the language model, LM-ANet delivers zero-shot and long-tail performance comparable to or surpassing many recent approaches based on significantly larger vision-language models, while exhibiting strong generalization to Open Images v6.**

***Index Terms*—Scene Graph Generation, Language Models, Zero-Shot Learning, Prompt Learning, Visual Relationship Detection**

I. Introduction

Scene graph generation (SGG) aims to detect object instances within an image and infer the semantic relationships between them, producing a structured and interpretable representation of visual scenes. Such graph-level representations have proven highly valuable for a variety of downstream reasoning-centric tasks, including image captioning [1], visual question answering [2], and semantic layout–based image generation [3]. A standard SGG pipeline consists of object detection, classification, and predicate prediction, with predicate prediction widely regarded as the most challenging component because it requires connecting visual cues, contextual dependencies, and commonsense knowledge. Figure 1 illustrates the general process of SGG.

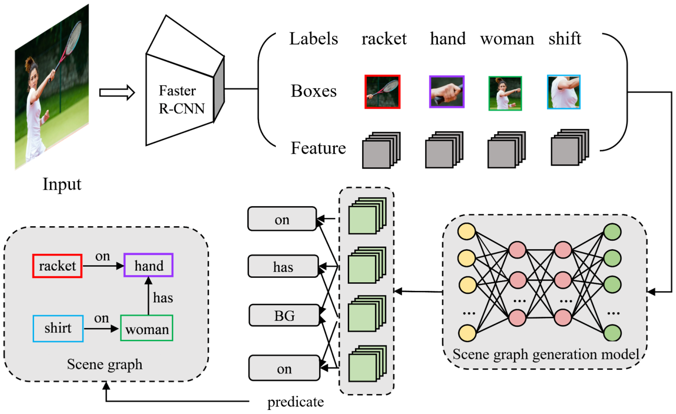

Fig. 1. General process for scene graph generation tasks.

Despite substantial progress in contextual modeling [4] [5]and unbiased learning [6], current SGG models still suffer from significant limitations when predicting predicates, particularly for rare or unseen subject–predicate–object combinations. Models heavily relying on dataset-specific visual co-occurrence statistics often overfit to frequent patterns and exhibit weak generalization to semantically plausible but visually underrepresented relationships. For instance, in the Visual Genome (VG) dataset [7], the most common predicate "on" accounts for over 27,000 occurrences (approximately 17% of all relations), while tail predicates like "painted on" or "parked on" appear fewer than 50 times each, leading to a severe long-tail distribution that exacerbates bias. This results in poor performance on zero-shot triplets such as "elephant-wearing-glasses" or "flower-painted on-vase", where commonsense plausibility is evident but lacks visual training data [8]. This deficiency becomes even more pronounced under zero-shot and tail-predicate settings, where commonsense reasoning or linguistic prior knowledge is necessary to make reliable predictions. As a result, existing SGG systems remain constrained by the biases inherent in training data and fail to fully utilize broader semantic knowledge. In contrast, recent advances in large language models (LLMs) like GPT-3 [9] have revealed their strong world-knowledge capacity, contextual reasoning ability, and robustness across domains. LLMs trained on massive text corpora inherently capture semantic compatibility between concepts (e.g., person–ride–horse is far more plausible than horse–ride–person), even when such instances are unseen in a visual dataset. Notably, vision-language models (VLMs) like CLIP [10] excel in global image-text alignment but often falter in fine-grained pairwise relation reasoning. In comparison, pure language models like BERT [11], with their masked language modeling (MLM) pre-training, are better suited for cloze-style relation filling when conditioned on subject-object pairs. Although several works have introduced linguistic constraints through word embeddings, sentence-level prior graphs, or BERT-based encoding, these modules remain static and weakly coupled with visual features, limiting their ability to dynamically adapt to unseen triplets.

To address these limitations, we propose LM-ANet, a language-model-aided SGG framework that jointly leverages visual cues and high-level semantic reasoning derived from frozen pre-trained language models. LM-ANet introduces an instructional prompting mechanism that constructs subject–object-conditioned prompts, enabling the language model to actively guide predicate prediction with semantic consistency and contextual relevance. This design preserves fine-grained appearance details in the visual encoder while providing flexible and transferable semantic priors from the language model, significantly improving generalization to rare and unseen predicate triplets.

Our work tackles two core challenges in SGG: (1) the limited generalization of visual-only models under long-tail and zero-shot conditions, and (2) the underutilization of external linguistic knowledge in existing frameworks. Specifically, our contributions are threefold:

- We design a novel prompt template and projection layer to seamlessly insert visual tokens into BERT's embedding space, achieving end-to-end relational reasoning without task-specific classifiers.
- We employ prototype-based matching with InfoNCE loss to mitigate overfitting and substantially enhance tail and zero-shot predicate recognition.
- Extensive experiments on Visual Genome and Open Images v6 [12] show that LM-ANet significantly outperforms prior methods in zero-shot and long-tail predicate prediction while maintaining competitive overall performance, confirming its effectiveness and robustness across datasets.

II. RELATED WORK

Scene graph generation. Since the introduction of scene graphs by Johnson et al. [13] , SGG has evolved from early detection-based frameworks into a comprehensive research area centered on understanding visual relations. A large body of work focuses on context modeling through message passing [14], graph neural networks, transformer-based relation encoders, and global contextual aggregation [15] [16]. Another extensive direction addresses the long-tail distribution and unbiased learning via causal intervention [17], predicate rebalancing, debiasing objectives, or distribution calibration to mitigate strong frequency bias in datasets such as Visual Genome. Compositional and zero-shot generalization have emerged as a key challenge, with approaches exploring factorized predicate representations, analogy-based inference, relational distillation, and semantic-prior-based methods. More recently, increasing efforts incorporate linguistic knowledge via word embeddings, sentence-level prior graphs, or BERT-based encoding to provide additional constraints. However, these linguistic modules remain either static or weakly coupled with visual reasoning, which prevents them from fully exploiting the high-level reasoning needed for rare or unseen predicates. For example, EGTR [18] extracts graphs directly from transformer object queries to improve efficiency, but it is still limited to closed-set predicates. HiKER-SGG [19] introduces hierarchical knowledge for better robustness, yet relies on static priors. PGSG [20] leverages generative VLMs for sequence-to-graph parsing, which advances the recognition of novel relations while incurring high prompt-engineering costs.

Pre-trained vision-language models. Vision-language models (VLMs) such as CLIP [10], BLIP [21], and UniCL [22], have achieved impressive zero-shot, open-vocabulary, and compositional generalization capabilities by mapping images and text into a shared embedding space, building upon earlier works like VinVL [23]. These strengths make them highly attractive for visual relationship reasoning. However, generic VLMs are primarily optimized for holistic image-text alignment rather than fine-grained pairwise interactions and lack explicit subject-predicate-object structural encoding. Recent open-vocabulary scene graph generation (OV-SGG) approaches attempt to bridge this gap: CLIP-Driven 3D Scene Graph [24] aligns 3D point clouds with textual relation prompts through cross-modal contrastive learning, while Open-World SGG [25] employs multimodal prompting without fine-tuning but suffers from embedding misalignment on rare relations. These limitations highlight the need for structured adaptation of pre-trained VLM knowledge to explicit relational reasoning. In this work, we propose a BERT-style masked language modeling objective to effectively inject triplet-aware relational priors into frozen vision-language representations.

Prompt learning. Prompt learning has emerged as an efficient paradigm for unlocking latent knowledge in large pre-trained models with minimal tunable parameters. Originally applied to NLP, its success has extended to VLMs through prompt tuning, prefix tuning, and contextual prompting, achieving strong performance in classification, retrieval, and multimodal reasoning [26] [27]. In the context of SGG, prompt learning offers the opportunity to reformulate predicate prediction as language-guided reasoning. Despite these advances, prompt learning in SGG has received comparatively less attention.Nevertheless, existing approaches rarely model fine-grained relational structures or explicitly incorporate subject-object pair information.

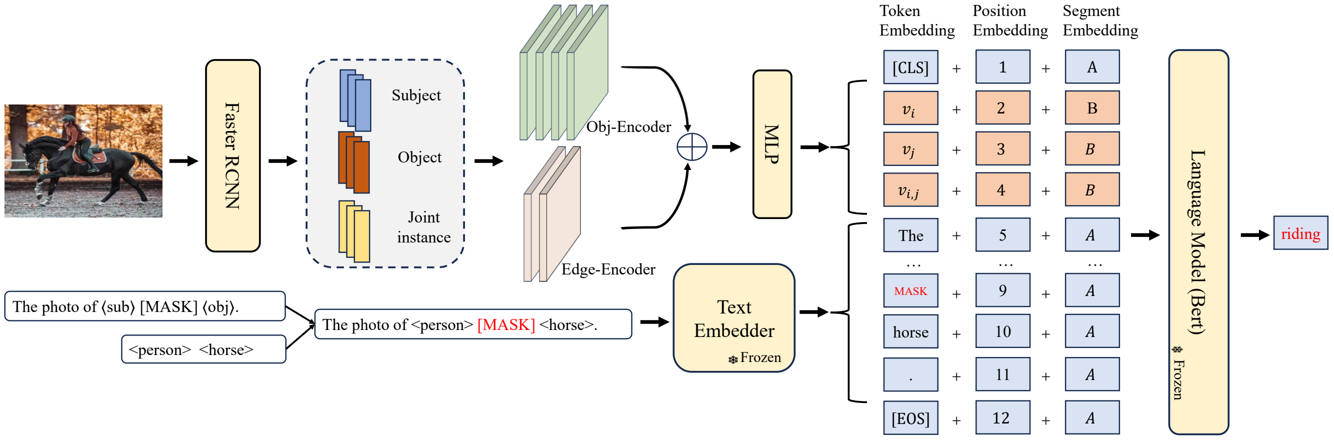

Fig. 2. The overall structure of our proposed LM-ANet. The framework integrates visual feature extraction via Faster R-CNN, generating Subject, Object, and Joint Instance features. These features are encoded and fused with embeddings from a Frozen Text Embedder using masked sentence templates (e.g., The photo of ¡person¿ [MASK] ¡horse¿). The fused features and text are combined into a sequence with Token, Position, and Segment Embeddings (distinguishing A and C segments) and fed into a Frozen Language Model (Bert) for the final prediction (e.g., 'riding').

Recent works include hierarchical prompt learning [19] that progressively refines prompts across predicate hierarchies and in-context prompting for lifelong SGG [20] that dynamically selects exemplars to mitigate catastrophic forgetting. Although prompt learning has shown promise in segmentation and detection, extending it to structured SGG with fully frozen pretrained models remains underexplored. We address this gap with LM-ANet.

III. APPROACH

LM-ANet is a novel scene graph generation framework that performs predicate prediction by injecting visual features into a frozen pre-trained language model (Fig. 2). The overall framework consists of four tightly coupled components: a visual feature extraction and projection module that aligns instance and union features with BERT's embedding space, a Language Prompt Module (LMP) that constructs cloze-style prompts conditioned on the detected subject-object categories, a frozen BERT encoder that performs bidirectional fusion and fills the [MASK] token via masked language modeling, and a Similarity Metric Classification module (SMC) that replaces conventional classifiers with prototype-based nearest-neighbor matching and InfoNCE training in the shared semantic space. By reformulating predicate prediction as language-model-guided semantic matching instead of biased statistical classification, LM-ANet effectively injects rich linguistic priors and exhibits substantially improved generalization to rare and unseen relationship triplets.

A. *Visual Feature Extraction and Projection*

Given an image, a Faster R-CNN detector [28] detects N instances and outputs their bounding boxes $B = \{b_1, \ldots, b_N\}$, RoI-aligned features $I = \{i_1, \ldots, i_N\}$, and predicted labels $L = \{l_1, \ldots, l_N\}$. The visual encoder ϕ_v produces individual instance features

$$O = \{o_1, \ldots, o_N\} = \phi_v(B, I) \tag{1}$$

and pairwise union features $U = \{u_{i,j} \mid i \neq j\}$, where $u_{i,j} = \phi_v(b_i \cup b_j, I)$ denotes the visual feature extracted from the union bounding box of instances i and j. These union features effectively encode spatial layouts and inter-object context, which are essential for predicting spatial predicates (e.g., "on", "under", "holding"). A lightweight projection layer $f(\cdot)$ then aligns all visual features to BERT's word embedding space:

$$\begin{aligned} v_i &= f(o_i), \\ v_j &= f(o_j), \\ v_{i,j} &= f(u_{i,j}). \end{aligned} \tag{2}$$

This enables the frozen language model to treat visual features as pseudo-tokens for unified processing.

B. *Language Prompt Module (LMP)*

To steer the frozen language model toward effective relational reasoning, we design a controllable, cloze-style textual prompt that explicitly incorporates the predicted categories of the subject and object instances. The prompt takes the following form:

$$T = \text{"The photo of } \langle\text{sub}\rangle \text{ [MASK] } \langle\text{obj}\rangle\text{."} \tag{3}$$

where ⟨sub⟩ and ⟨obj⟩ are placeholders replaced by the actual category names detected in the image (e.g., "person" and "horse" for a "person riding horse" relationship). The [MASK] token serves as the predicate slot to be filled by the language model. The template is designed to effectively elicit BERT's pre-trained knowledge of visual-linguistic relationships. The prompt is encoded using BERT's frozen text embedding layer:

$$T' = \text{TextEmb}(T) \in \mathbb{R}^{L \times d}, \tag{4}$$

where L denotes the prompt length and d is the hidden dimension of BERT. These prompt embeddings T' provide rich contextualized linguistic cues that will later be concatenated with projected visual tokens for joint reasoning inside the frozen BERT encoder.

C. Frozen Pre-trained Language Model Reasoning

We employ a frozen BERT encoder $\Omega(\cdot)$ as the core reasoning engine for relational prediction. Unlike conventional scene graph generation methods that rely on task-specific classifiers trained from scratch, our paradigm fully exploits BERT's powerful pre-trained language inference capabilities—particularly masked language modeling (MLM) and deep bidirectional contextual reasoning—to directly fill the predicate slot. For each candidate subject-object pair (i, j), the input sequence to BERT is constructed by concatenating the projected visual tokens, the embedded prompt T', and special delimiter tokens:

$$x = \text{[CLS]}\ v_i\ v_j\ v_{i,j}\ T'\ \text{[EOS]}, \tag{5}$$

where [CLS] and [EOS] are the standard BERT special tokens. By treating v_i, v_j, and $v_{i,j}$ as pseudo-word tokens in the same embedding space, BERT can seamlessly fuse fine-grained visual evidence with the linguistic priors encoded in the prompt. To enable effective processing by the BERT encoder, the final input embedding E_{final} for the sequence x is formed by summing the Token Embedding (E_{token}), Position Embedding (E_{pos}), and Segment Embedding (E_{seg}):

$$E_{\text{final}} = E_{\text{token}} + E_{\text{pos}} + E_{\text{seg}} \tag{6}$$

The Position Embeddings provide sequential order information for all tokens. Crucially, we utilize Segment Embeddings to distinguish the source of the tokens:

- The visual tokens, $\{v_i, v_j, v_{i,j}\}$, are assigned to a distinct segment (Segment B).
- The special tokens ([CLS], [EOS]) and the embedded prompt T' are assigned to the primary segment (Segment A).

This explicit segmentation guides the frozen BERT encoder to better modulate attention between the visual evidence and the linguistic context. The concatenated sequence is fed into the frozen BERT encoder:

$$x' = \Omega(x), \tag{7}$$

The hidden state at the [MASK] position in T' naturally aggregates both visual context (from v_i, v_j, $v_{i,j}$) and textual context (from the subject/object categories and template structure). The hidden state at the [MASK] position thus provides a powerful relational representation that encodes the likelihood of all possible predicates in a continuous semantic space.

D. Similarity Metric Classification Module (SMC)

Instead of using a conventional linear classifier on top of the relation representation x', which tends to overfit to frequent predicates in long-tailed datasets and exhibits limited generalization to rare predicates, we adopt a prototype-based classification approach that operates entirely in the embedding space. All predicate phrases are encoded once using the same text encoder TextEmb$(\cdot)$, yielding a set of learnable prototype embeddings $R = \{r_0, r_1, \ldots, r_{K-1}\}$, where K is the number of predicate categories and each r_k represents the centroid of predicate category k.

During training, we minimize the InfoNCE loss:

$$\mathcal{L}_c = -\log \frac{\exp\big(\langle x', r_t\rangle/\tau\big)}{\sum_{k=0}^{K-1}\exp\big(\langle x', r_k\rangle/\tau\big)}, \tag{8}$$

where r_t is the prototype corresponding to the ground-truth predicate, and τ is a learnable temperature parameter.

At inference, the predicted predicate is obtained via nearest prototype matching using scaled cosine similarity:

$$\hat{y} = \arg\max_k \frac{\langle x', r_k\rangle}{\tau}. \tag{9}$$

This formulation offers several advantages:

- It eliminates separate classification heads, reducing parameters and overfitting risk.
- Learnable prototypes naturally mitigate predicate imbalance and long-tail issues.
- Inference reduces to fast embedding retrieval.

IV. Experiments

A. Datasets

We conduct extensive experiments on two large-scale, widely-adopted visual relationship detection benchmarks to thoroughly evaluate the effectiveness and generalization capability of our proposed approach.

Visual Genome (VG). The Visual Genome dataset is one of the most comprehensive and challenging benchmarks for scene graph generation, containing 108,077 images with densely annotated objects and relationships. Following the standard preprocessing protocol established by Xu et al. [14], which has been widely adopted in nearly all recent state-of-the-art methods, we filter the dataset to retain only the 150 most frequent object categories and 50 predicate categories. This results in a cleaned version that balances annotation richness with computational feasibility. After preprocessing, the dataset consists of 62,723 training images and approximately 26K–32K test images (exact numbers vary slightly across works due to minor filtering differences). Following [4] [6], we randomly sample 5,000 images from the training set as the validation split.

Open Images V6. To evaluate cross-dataset generalization, we further conduct experiments on Open Images V6, which exhibits markedly different data distribution and annotation sparsity compared to Visual Genome. We use the official relationship detection subset comprising 301 object categories and 31 predicates. Following the standard splits adopted in recent competitive works [6], the training, validation, and test sets contain 126,368, 1,813, and 5,322 images, respectively.

B. Metrics

We comprehensively evaluate our method on the three standard scene graph generation tasks widely adopted in the literature [4] [14]. In **Predicate Classification (PredCls)**, the model predicts relationship predicates given ground-truth object bounding boxes and category labels. In **Scene Graph Classification (SGCls)**, both object categories and relationship predicates are predicted with ground-truth bounding boxes provided. In **Scene Graph Detection (SGDet)**, the model performs fully end-to-end scene graph generation from raw input images, jointly conducting object detection, category classification, and relationship prediction.

Metrics on Visual Genome. Following the unbiased evaluation framework proposed by Tang et al. [6], which has been universally adopted as the standard benchmark protocol, we report Recall@K (R@K), mean Recall@K (mR@K), zero-shot Recall@K (zsR@K), and the balanced mean of R@K and mR@K (Mean@K, denoted as M@K) for $K \in \{50, 100\}$. The zsR@K specifically evaluates generalization to unseen subject–predicate–object triplets, while M@K is computed as the arithmetic mean of R@K and mR@K at each K, providing a more balanced summary of head and tail entity performance.

Metrics on Open Images. On the Open Images benchmark, we follow the official evaluation protocol [6] and report Recall@50 (R@50), weighted mean Average Precision for relationships (wmAP$_{\text{rel}}$), and weighted mean Average Precision for phrase detection (wmAP$_{\text{phr}}$). These metrics provide a comprehensive assessment under large-scale, open-vocabulary conditions and are particularly sensitive to both localization accuracy and semantic correctness.

C. Implementation Details

Our model is implemented in PyTorch 2.1 and trained on 1–8 NVIDIA RTX 3090 GPUs using PyTorch DistributedDataParallel. For fair and reproducible comparison, we use the publicly available pre-trained Faster R-CNN detector with ResNeXt-101-FPN backbone provided by the official scene graph benchmark repository [29], which has been adopted by nearly all recent state-of-the-art methods [6] [30] [31] [32]. The object detector is kept completely frozen throughout training to eliminate variability introduced by detection performance.

The relation prediction module is trained using SGD with momentum 0.9 and weight decay 10^{-4}, while keeping the object detector frozen. We train for a total of 60,000 iterations with a batch size of 8 images per GPU. The initial learning rate is set to 0.001, and we apply a stepwise decay schedule, reducing the learning rate by a factor of 10 at the 40,000-th and 50,000-th iterations, respectively. All reported results are obtained from the model checkpoint achieving the highest mean Recall@100 on the validation set. Inference is performed in a single forward pass without test-time augmentation or ensemble, ensuring high efficiency. Detailed hyperparameter configurations and training logs will be released upon publication.

D. Comparisons with State-of-the-art Methods

As shown in Table I, our method achieves substantial and consistent improvements in zero-shot scene graph generation across all three tasks. On PredCls, LM-ANet reaches 20.3%/**23.8%** in zs-R@50/100, outperforming the previous best result from PE-Net by 3.1 and 2.9 percentage points, respectively. In the significantly more challenging SGDet task, where both object localization and classification must be performed without ground-truth boxes, we further advance the state-of-the-art from 4.1%/5.8% (PE-Net) to 4.7%/6.8%, demonstrating robust generalization capabilities.

TABLE I
COMPARISON OF ZERO-SHOT SCENE GRAPH GENERATION ON VISUAL GENOME. RESULTS ARE REPORTED AS ZS-R@50 / ZS-R@100 UNDER PREDCLS, SGCLS, AND SGDET SETTINGS. **BOLD** AND UNDERLINED INDICATE THE BEST AND SECOND-BEST PERFORMANCE, RESPECTIVELY.

	PredCls	SGCls	SGDet
Method	zs-R@50 / 100	zs-R@50 / 100	zs-R@50 / 100
Motifs [4]	3.24 / 5.36	0.68 / 1.13	0.05 / 0.11
VCTree [5]	3.27 / 5.51	1.17 / 2.08	0.31 / 0.69
Motifs-TDE [6]	14.4 / 18.2	3.4 / 4.5	2.3 / 2.9
VCTree-TDE [6]	14.3 / 17.6	3.2 / 4.0	2.6 / 3.2
Motifs-EBM [17]	4.87 / -	1.25 / -	0.23 / -
Motifs-QuatRE [33]	11.9 / 15.2	2.8 / 3.6	0.2 / 0.4
PE-Net [8]	17.2 / 20.9	5.4 / 6.5	4.1 / 5.8
LM-ANet (Ours)	**20.3** / **23.8**	**5.9** / **7.1**	**4.7** / **6.8**

Compared with early representative methods such as Motifs and VCTree, which typically yield zs-R@50 below 3.3% on PredCls due to their reliance on statistical biases learned from frequent triplets, our approach delivers dramatically higher recall on unseen relationships. This gap underscores a fundamental paradigm shift driven by the integration of Large Language Models (LLMs) and prompt engineering. Unlike traditional classifiers that struggle to extrapolate beyond seen patterns due to their reliance on statistical biases, LM-ANet reformulates predicate prediction as a language-model-guided semantic matching task. By leveraging the rich linguistic priors encapsulated in the pre-trained model via cloze-style prompts, our approach enables the inference of plausible relationships for unseen triplets based on deep semantic reasoning rather than mere memorization of frequent categories.

In addition to its outstanding zero-shot capabilities, we further conduct comprehensive evaluations under the conventional (non-zero-shot) scene graph generation setting on both Visual Genome and Open Images datasets, with detailed results reported in Table II and Table III, respectively.

On Visual Genome, LM-ANet exhibits remarkable effectiveness in mitigating the notorious long-tail predicate distribution—a persistent bottleneck that has plagued nearly all prior methods. As evidenced in Table II, our approach achieves new state-of-the-art mean Recall@50/100 scores across all three tasks, with particularly striking gains on the tail-sensitive mR@K metric. Notably, on the PredCls task, LM-ANet improves mR@50/100 from 18.3%/19.9% (previously held by R-CAGCN) to 21.3%/23.4%, corresponding to substantial relative gains of 16.4% and 17.6%, while simultaneously establishing new records in overall Mean@K performance (43.7%/45.8%). This balanced advancement—significant tail enhancement with virtually no degradation on frequent (head) predicates—stands in sharp contrast to most debiasing techniques, which typically trade head performance for tail gains.

To provide deeper insight into this desirable behavior, Figure 3 visualizes per-predicate R@100 performance on PredCls compared against the strong baseline Motifs. The

TABLE II

PERFORMANCE COMPARISON OF SCENE GRAPH GENERATION ON THE VISUAL GENOME DATASET. **BOLD** AND UNDERLINED INDICATE THE BEST AND SECOND-BEST RESULTS, RESPECTIVELY.

	PredCls			SGCls			SGDet		
Method	R@50/100	mR@50/100	M@50/100	R@50/100	mR@50/100	M@50/100	R@50/100	mR@50/100	M@50/100
Motifs [4]	65.3 / 67.2	14.9 / 16.3	40.1 / 41.8	38.9 / 39.8	8.3 / 8.8	23.6 / 24.3	**32.1** / **36.8**	6.6 / 7.9	19.4 / 22.4
G-RCNN [34]	65.4 / 67.2	16.4 / 17.2	40.9 / 42.2	37.0 / 38.5	9.0 / 9.5	23.0 / 24.0	29.7 / 32.8	5.8 / 6.6	17.8 / 19.7
VCTree [5]	65.5 / 67.4	16.7 / 17.9	41.1 / 42.7	40.3 / 41.6	7.9 / 8.3	24.1 / 25.0	31.9 / 36.0	6.4 / 7.3	19.2 / 21.7
KERN [15]	65.8 / 67.6	17.7 / 19.2	41.8 / 43.4	36.7 / 37.4	9.4 / 10.0	23.1 / 23.7	27.1 / 29.8	6.4 / 7.3	16.8 / 18.6
GPS-Net [35]	65.2 / 67.1	15.2 / 16.6	40.2 / 41.9	37.8 / 39.2	8.5 / 9.1	23.2 / 24.2	31.3 / 35.9	6.7 / 8.6	19.0 / 22.3
R-CAGCN [30]	**66.6** / **68.3**	18.3 / 19.9	42.5 / 44.1	38.3 / 39.0	10.2 / 11.1	24.3 / 25.1	28.1 / 31.3	7.9 / 8.8	18.0 / 20.1
RelTR [36]	64.2 / -	21.2 / -	42.7 / -	36.6 / -	11.4 / -	24.0 / -	27.5 / -	**10.8** / -	19.2 / -
LM-ANet (Ours)	66.0 / 68.1	**21.3** / **23.4**	**43.7** / **45.8**	**40.7** / **41.8**	**13.6** / **14.9**	**27.2** / **28.4**	31.0 / 35.7	9.3 / **11.0**	**20.2** / **23.4**

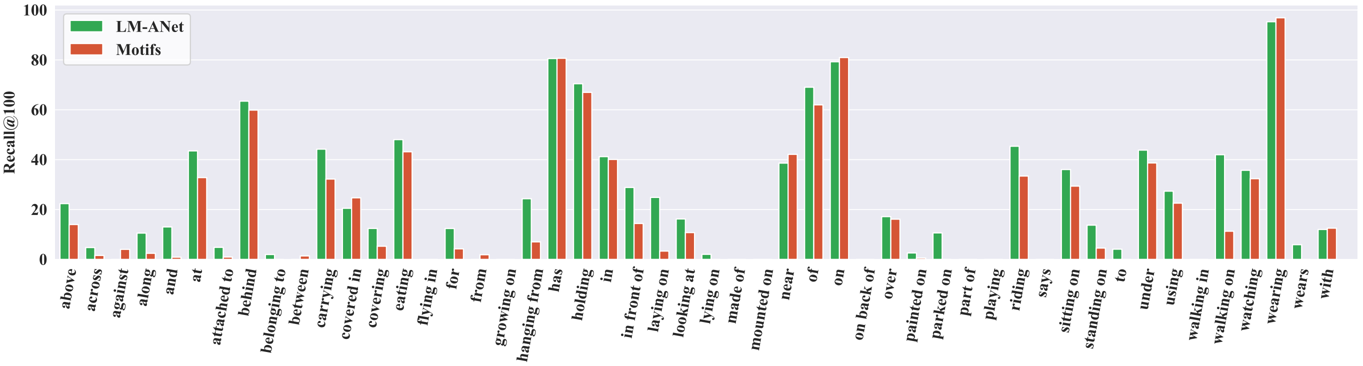

Fig. 3. R@100 of LM-ANet and Motifs for all predicate categories.

results clearly demonstrate that LM-ANet delivers consistent and often dramatic improvements across the entire predicate frequency spectrum. For extremely rare tail predicates (e.g., "belonging to", "painted on", "lying on", "parked on", "attached to"), which typically receive near-zero recall in conventional models, our method yields gains ranging from 10× to over 50× in many cases. Meanwhile, for high-frequency head predicates such as "on", "has", "wearing", and "in", LM-ANet maintains competitive or superior performance, successfully avoiding the common head-tail trade-off.

We further validate the cross-dataset generalization of LM-ANet on the large-scale Open Images v6 benchmark (Table III). Despite its significantly different data distribution and much sparser relationship annotations compared to Visual Genome, LM-ANet achieves new state-of-the-art performance on the two most important comprehensive metrics: R@50 of 75.1% (+0.1% over prior best) and score$_{\text{wtd}}$ of 42.2% (+0.1% improvement). Although falling marginally behind the top results on the two wmAP sub-metrics (by 0.3 and 0.1 points, respectively), our method remains highly competitive while simultaneously advancing the overall leaderboard. This consistent superiority on aggregated metrics, achieved without any dataset-specific hyper-parameter tuning, strongly demonstrates that the advantages of our prototype-based semantic matching and frozen BERT integration are not artifacts of Visual Genome but represent fundamental and transferable improvements for real-world large-scale scene graph generation.

TABLE III

PERFORMANCE COMPARISON WITH STATE-OF-THE-ART METHODS ON THE OPEN IMAGES V6 DATASET. **BOLD** AND UNDERLINED INDICATE THE BEST AND SECOND-BEST RESULTS, RESPECTIVELY.

Method	R@50	wmAP rel	wmAP phr	score$_{\text{wtd}}$
Motifs [4]	71.6	29.9	31.6	38.9
VCTree [5]	74.1	**34.2**	33.1	40.2
GPS-Net [35]	74.8	32.9	34.0	41.7
G-RCNN [34]	74.5	33.2	**34.2**	41.8
BGNN [37]	75.0	33.5	**34.2**	42.1
LM-ANet (Ours)	**75.1**	33.9	34.1	**42.2**

It is worth emphasizing that the above Open Images results are obtained usingexactly the same hyperparameters and prompt templates as those used on Visual Genome—no dataset-specific learning rate scheduling, focal loss weighting, or predicate rebalancing was performed whatsoever. This "zero-tuning" cross-dataset transfer further corroborates the inherent robustness and domain invariance of the semantic priors extracted from the frozen large language model.

E. Ablation Studies

To thoroughly investigate the contribution of each core component, we conduct ablation studies on the PredCls task of Visual Genome (Table IV). The vanilla Motifs backbone with a standard linear classifier achieves R@50/100 = 65.3/67.2. Our proposed Similarity Metric Classification (SMC) module replaces the linear classifier with fully learnable predicate prototypes operating in the shared embedding space. To fairly

TABLE IV

ABLATION STUDIES ON THE PREDCLS TASK OF VISUAL GENOME. OUR FULL SIMILARITY METRIC CLASSIFICATION (SMC) USES *learnable prototypes* AND A *learnable temperature* τ. THE ROW "+ SMC (FIXED PROTOTYPE)" FREEZES THEM FOR ABLATION ONLY.

Method	R@50/100	mR@50/100	zs-R@50/100
Baseline (Motifs)	65.3 / 67.2	14.9 / 16.3	14.4 / 18.2
+ SMC (fixed prototype)	65.5 / 67.4	18.7 / 20.1	16.8 / 20.5
+ LMP (BERT-small)	65.1 / 67.3	19.5 / 21.0	17.9 / 21.2
LM-ANet (ours)	**66.0 / 68.1**	**21.3 / 23.4**	**20.3 / 23.8**

TABLE V

ABLATION STUDY ON DIFFERENT CLOZE-STYLE PROMPT TEMPLATES FOR THE PREDCLS TASK ON VISUAL GENOME. ALL MODELS ARE BUILT UPON THE "+ LMP (BERT-BASE)" SETTING IN TABLE IV.

Prompt Template	R@50/100	mR@50/100	zs-R@50/100
"[sub] [MASK] [obj]" (no template)	65.3 / 67.4	19.8 / 21.3	17.7 / 21.0
"A [sub] [MASK] a [obj]."	65.6 / 67.7	20.5 / 22.1	18.3 / 21.8
"The [sub] is [MASK] the [obj]."	65.5 / 67.6	20.4 / 22.0	18.2 / 21.7
"The photo of [sub] [MASK] [obj]." (ours)	**66.0 / 68.1**	**21.3 / 23.4**	**20.3 / 23.8**

isolate the benefit of metric-based classification, we report a variant with frozen prototypes in the second row, which already yields clear gains on long-tail (mR@50/100: 14.9→18.7, 16.3→20.1) and zero-shot metrics (zs-R@50/100: 14.4→16.8, 18.2→20.5).

Further incorporating the Language Prompt Module (LMP) with a frozen BERT-small significantly improves performance on rare and unseen predicates. Our complete model LM-ANet, which combines the full learnable SMC with LMP (BERT-small), achieves the best results of mR@50/100 = 21.3/23.4 and zs-R@50/100 = 20.3/23.8. Replacing BERT-base with the lighter BERT-small consistently reduces long-tail performance by 1.8/2.4 points and zero-shot by 1.1/2.6 points on average, highlighting the value of stronger linguistic priors.

Additional prompt ablations (Table V) show that our photographic template "The photo of ⟨sub⟩ [MASK] ⟨obj⟩." consistently outperforms other variants, delivering the largest gains on long-tail (+0.8–1.3) and zero-shot predicates (+0.9–2.8) by providing image-relevant context and strong bidirectional constraints.

These results demonstrate the effectiveness and complementarity of (1) fully learnable prototype-based metric classification in embedding space and (2) carefully designed photographic prompts that enrich frozen language models with scene-aware linguistic knowledge — together enabling robust relational reasoning under severe distribution shifts. These improvements are strictly additive and mutually reinforcing — only their synergy delivers SOTA on highly skewed, open-vocabulary relation prediction.

F. Qualitative Analysis and Visualization

To intuitively demonstrate the superiority of LM-ANet, we present representative qualitative comparisons against the strong baseline Motifs in Figure 4.

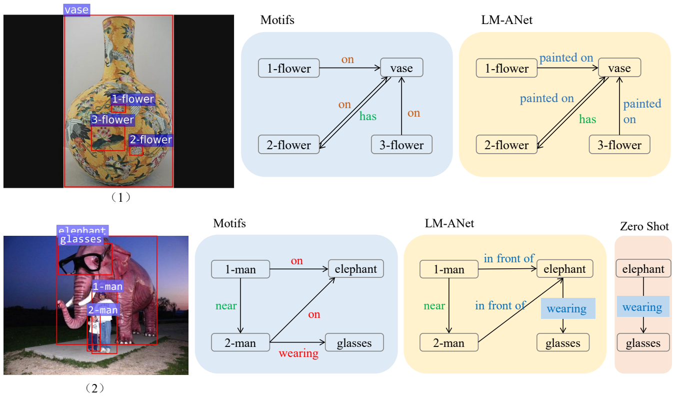

Fig. 4. Visualization Results of Motifs (in blue) and LM-ANet (in yellow) on the PredCls task.

In the top example, Motifs produces overly generic predicates such as "on", whereas LM-ANet accurately captures semantically richer ones like "painted on" and "attached to", resulting in more precise and descriptive triplets (e.g., "flower painted on vase" and "leaf attached to branch"). This highlights our model's enhanced expressive power enabled by linguistic knowledge.

The bottom example showcases LM-ANet's remarkable zero-shot capability on the unseen triplet involving an "elephant wearing glasses". Motifs fails completely and predicts "man wearing glasses" due to strong statistical bias toward frequent subject-object pairs seen during training. In contrast, LM-ANet correctly recognizes both the unusual subject (elephant) and the rare predicate "wearing", producing the accurate zero-shot triplet. This success stems from our prototype-based matching mechanism, which effectively aligns the visual-linguistic representation with the semantic prototype of "wearing", even without any training instance of elephants wearing glasses.

Such cases are not isolated: LM-ANet consistently produces fewer ambiguous or repetitive relations and favors contextually appropriate, fine-grained predicates across diverse scenes.These qualities make the generated scene graphs not only quantitatively superior, but also significantly more informative and closer to human perception.These qualitative results, together with substantial quantitative gains, conclusively show that LM-ANet generates significantly more accurate, informative, and human-aligned scene graphs, especially in challenging long-tail and zero-shot scenarios.

V. CONCLUSION

In this work, we propose LM-ANet, a simple yet effective framework that seamlessly integrates a frozen pretrained BERT into scene graph generation through visual token projection and cloze-style prompting. By casting predicate prediction as a masked language modeling task conditioned on subject-object pairs, LM-ANet effectively injects rich commonsense knowledge and significantly mitigates the long-standing co-occurrence bias in traditional SGG models. Extensive experiments on Visual Genome and Open Images V6 demonstrate that LM-ANet achieves strongly competitive performance across all settings, notably attaining a zero-shot

Recall@100 of 23.8% on PredCls—significantly outperforming all previously published methods under identical evaluation protocols. Thorough ablations and qualitative visualizations further validate the contribution of each component and the superior descriptive quality of the generated scene graphs, particularly for rare and unseen triplets. Moving forward, we plan to investigate larger frozen encoders (e.g., RoBERTa-large, DeBERTa V3) as well as decoder-only LLMs equipped with trainable soft prompts or LoRA adapters to further advance open-vocabulary and cross-dataset generalization.

Acknowledgment

This work was supported in part by the National Natural Science Foundation of China under Grant Nos. 62476209 and 62176200, in part by the Key Research and Development Program of Shaanxi under Grant No. 2024CY2-GJHX-18,and in part by the Natural Science Basic Research Program of Shaanxi under Grant No. 2022JC-45.

References

[1] W. Wang, R. Wang, and X. Chen, “Topic scene graph generation by attention distillation from caption,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), 2021, pp. 15900–15910.

[2] J. Yu, Z. Zhu, Y. Wang, W. Zhang, Y. Hu, and J. Tan, “Cross-modal knowledge reasoning for knowledge-based visual question answering,” Pattern Recognit., vol. 108, p. 107563, 2020.

[3] H. Dhamo, F. Manhardt, N. Navab, and F. Tombari, “Graph-to-3d: End-to-end generation and manipulation of 3d scenes using scene graphs,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), 2021, pp. 16352–16361.

[4] R. Zellers, M. Yatskar, S. Thomson, and Y. Choi, “Neural motifs: Scene graph parsing with global context,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2018, pp. 5831–5840.

[5] K. Tang, H. Zhang, B. Wu, W. Luo, and W. Liu, “Learning to compose dynamic tree structures for visual contexts,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2019, pp. 6619–6628.

[6] K. Tang, Y. Niu, J. Huang, J. Shi, and H. Zhang, “Unbiased scene graph generation from biased training,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2020, pp. 3716–3725.

[7] R. Krishna, Y. Zhu, O. Groth, et al., “Visual genome: Connecting language and vision using crowdsourced dense image annotations,” Int. J. Comput. Vis., vol. 123, no. 1, pp. 32–73, 2017.

[8] C. Zheng, X. Lyu, L. Gao, X. Liu, and Y. Wang, “Prototype-based embedding network for scene graph generation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2023, pp. 22783–22792.

[9] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, “Language models are few-shot learners,” in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, pp. 146, 2020.

[10] A. Radford, J. W. Kim, C. Hallacy, et al., “Learning transferable visual models from natural language supervision,” in Proc. Int. Conf. Mach. Learn. (ICML), 2021, pp. 8748–8763.

[11] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics (NAACL), 2019, pp. 4171–4186.

[12] A. Kuznetsova, H. Rom, N. Alldrin, et al., “The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale,” Int. J. Comput. Vis., vol. 128, no. 7, pp. 1956–1981, 2020.

[13] J. Johnson, R. Krishna, M. Stark, et al., “Image retrieval using scene graphs,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2015, pp. 3668–3678.

[14] D. Xu, Y. Zhu, C. B. Choy, and L. Fei-Fei, “Scene graph generation by iterative message passing,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2017, pp. 5410–5419.

[15] T. Chen, W. Yu, R. Chen, and L. Lin, “Knowledge-embedded routing network for scene graph generation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2019, pp. 6163–6171.

[16] M. Qi, W. Li, Z. Yang, Y. Wang, and J. Luo, “Attentive relational networks for mapping images to scene graphs,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3957–3966, 2019.

[17] M. Suhail, A. Mittal, B. Siddiquie, et al., “Energy-based learning for scene graph generation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2021, pp. 13936–13945.

[18] J. Im, J. Y. Nam, N. Park, H. Min, and S. Kim, “EGTR: Extracting graph from transformer for scene graph generation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2024, pp. 24229–24238.

[19] C. Zhang, S. Stepputtis, J. Campbell, et al., “HiKER-SGG: Hierarchical knowledge enhanced robust scene graph generation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2024, pp. 28233–28243.

[20] R. Li, S. Zhang, D. Lin, J. Xu, and Y. Wei, “From pixels to graphs: Open-vocabulary scene graph generation with vision-language models,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2024, pp. 28076–28086.

[21] J. Li, D. Li, C. Xiong, and S. Hoi, “BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation,” in Proc. Int. Conf. Mach. Learn. (ICML), 2022, pp. 12888–12900.

[22] J. Li, J. Peng, H. Li, et al., “UniCL: A universal contrastive learning framework for large time series models,” unpublished.

[23] P. Zhang, X. Li, X. Hu, J. Yang, L. Zhang, L. Wang, Y. Choi, and J. Gao, “Vinvl: Revisiting visual representations in vision-language models,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5579–5588, 2021.

[24] L. Chen, X. Wang, J. Lu, H. Wu, and X. Sun, “CLIP-driven open-vocabulary 3d scene graph generation via cross-modality contrastive learning,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2024, pp. 27863–27873.

[25] Y. Zhang, Y. Pan, T. Yao, H. Xu, and T. Mei, “Learning to generate language-supervised and open-vocabulary scene graph using pre-trained visual-semantic space,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2023, pp. 2915–2924.

[26] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, “Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing,” ACM Comput. Surveys, vol. 55, no. 9, pp. 1–35, 2023.

[27] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, “Learning to prompt for vision-language models,” Int. J. Comput. Vis., vol. 130, no. 9, pp. 2337–2348, 2022.

[28] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards real-time object detection with region proposal networks,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 6, pp. 1137–1149, 2016.

[29] K. Tang, “A scene graph generation codebase in PyTorch,” 2020. [Online]. Available: https://github.com/KaihuaTang/Scene-Graph-Benchmark.pytorch

[30] G. Yang, J. Zhang, Y. Zhang, B. Wu, and Y. Yang, “Probabilistic modeling of semantic ambiguity for scene graph generation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2021, pp. 12527–12536.

[31] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, “Feature pyramid networks for object detection,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2017, pp. 2117–2125.

[32] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, “Aggregated residual transformations for deep neural networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2017, pp. 1492–1500.

[33] Z. Wang, X. Xu, G. Wang, Y. Yang, and H. T. Shen, “Quaternion relation embedding for scene graph generation,” IEEE Trans. Multimedia, vol. 25, pp. 8646–8656, 2023.

[34] J. Yang, J. Lu, S. Lee, D. Batra, and D. Parikh, “Graph R-CNN for scene graph generation,” in Proc. Eur. Conf. Comput. Vis. (ECCV), 2018, pp. 670–685.

[35] X. Lin, C. Ding, J. Zeng, and D. Tao, “GPS-Net: Graph property sensing network for scene graph generation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2020, pp. 3746–3753.

[36] Y. Cong, M. Y. Yang, and B. Rosenhahn, “RELTR: Relation transformer for scene graph generation,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 45, pp. 11169–11183, 2023.

[37] R. Li, S. Zhang, B. Wan, W. Luo, and Y. He, “Bipartite graph network with adaptive message passing for unbiased scene graph generation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2021, pp. 11109–11119.