Neurosymbolic Scene Graph Generation in the Open-Vocabulary Setting

First Author , **Second Author** , **Third Author**
Affiliation
{first, second, thrid}@example.com

Abstract

Scene graph generation (SGG), the task of representing an image as a structured graph of objects and their pairwise relationships, has traditionally been dominated by closed-set models trained on fixed vocabularies such as VG150. We present NEUROSGG, a training-free neurosymbolic pipeline for open-vocabulary SGG that enforces a strict separation between semantic and geometric reasoning. A large language model (LLM) performs all semantic tasks, including proposing object labels and predicting pairwise predicates, while deterministic rule-based components handle geometric reasoning. LLM-generated object labels are used to prompt SAM 3 for concept-driven instance segmentation. The resulting masks are subsequently canonicalized through deterministic post-processing, and a Set-of-Marks visual representation is constructed to provide contextual grounding for relation prediction over segmented instances. We evaluate our approach on a preliminary subset of the VG150 benchmark and report approximately 53% node recall and up to 7.8% triplet recall across three open-weight LLMs. Our results establish an interpretable, training-free baseline for open-world scene graph generation that requires no task-specific training at any stage of the pipeline.

1 Introduction

Scene graphs provide a structured, symbolic abstraction of visual scenes, in which nodes correspond to objects and predicate-labeled edges encode semantic relationships between them. By explicitly modeling relational structure, scene graphs support a wide range of downstream vision-and-language tasks, including visual question answering [Hildebrandt *et al.*, 2020], image captioning [Yang *et al.*, 2019], and embodied navigation [Singh *et al.*, 2023], where reasoning about *who interacts with whom, and in what way*, is fundamental [Krishna *et al.*, 2017]. The dominant paradigm for SGG, established by Xu et al. [2017] and refined through methods such as Neural Motifs [Zellers *et al.*, 2018], VCTree [Tang *et al.*, 2019], and RelTR [Cong *et al.*, 2023], treats both object and predicate prediction as classification over a fixed vocabulary—typically the VG150 split of 150 object categories and 50 predicate types. While this closed-set assumption permits clean benchmarking, it is fundamentally at odds with open-world visual understanding: real scenes contain objects and relations that fall outside any predefined taxonomy, and the VG150 predicate set is heavily skewed toward generic spatial terms (*e.g.,* "on", "in", "near", "has") that carry little relational specificity [Zellers *et al.*, 2018]. Recent work has begun to address this limitation by incorporating open-vocabulary reasoning into the SGG pipeline. Li et al. [2024] generate scene graph sequences via vision-language models, enabling novel concept prediction without a fixed output space. Kim et al. [2024] use an LLM to synthesise weakly supervised triplet annotations from captions, achieving strong results in the weakly supervised setting. Liu et al. [2025] enhance text representations with LLM-generated relation-aware prompts for contrastive alignment. Most recently, OvSGTR [Chen *et al.*, 2025] proposes a DETR-like transformer with a frozen CLIP backbone and relation-aware pre-training, achieving state-of-the-art performance across both closed-set and open-vocabulary VG150 settings. Critically, however, all of these approaches require task-specific training or fine-tuning as they are not applicable out-of-the-box to an *arbitrary* open-world image without access to annotated SGG data.

We propose a complementary direction: a **training-free, neurosymbolic pipeline** for open-vocabulary SGG, illustrated in Figure 2. Our approach is grounded in a clean architectural principle: the LLM handles all semantic decisions (what objects are present, which are salient, what predicates apply), while deterministic Python code handles all geometric decisions (mask deduplication, IoU-based merging, reading-order assignment, triplet validation). No ungrounded node can appear in the output graph; every object in the final scene graph corresponds to a predicted instance mask in the image.

Our main contributions include:

- We introduce the first **training-free** open-vocabulary SGG pipeline that uses an LLM as the semantic driver of a concept-level segmentation model (SAM 3), replacing both the trained object detector and the trained relation classifier found in prior work.

(a) input image of a building façade (b) segmented instance masks of façade (c) generated scene graph of façade

Figure 1: Overview of the proposed open-vocabulary scene graph generation approach, NEUROSGG, illustrated on an exemplary street-view image of a building façade. The generated scene graph contains 76 object nodes and 302 relational edges, including two disconnected subgraphs.

- We demonstrate that **Set-of-Marks prompting** [Yang *et al.*, 2023] provides sufficient visual grounding for an LLM to predict scene graph relations directly from segmented instances, without any relation-specific fine-tuning.
- We provide an initial quantitative analysis of a **metric bias in VG150**, showing that the annotation conventions of VG150 concentrate more than 50% of ground-truth triplets on only three spatial predicates. This distribution systematically disadvantages open-vocabulary approaches, which tend to generate richer and more diverse predicate descriptions than those represented in the benchmark annotations.
- We report an **LLM ablation** across Gemma 4 (31B) and Qwen 3.5 (122B), providing the first systematic comparison of locally hosted open-weight LLMs as zero-shot relation predictors within a scene graph generation pipeline. The current evaluation is conducted on a small subset of VG150 samples as a preliminary analysis.

2 Related Work

Closed-set scene graph generation. Early SGG methods frame relation prediction as message passing over a detected object graph [Xu *et al.*, 2017], who also established the widely-adopted VG150 evaluation. Neural Motifs [Zellers *et al.*, 2018] demonstrated that a simple frequency-prior baseline—predicting the most common predicate per object-pair label—is highly competitive on VG150, revealing the dataset's strong statistical biases. VCTree [Tang *et al.*, 2019] models visual context using dynamic tree structures, while RelTR [Cong *et al.*, 2023] formulates scene graph generation as an end-to-end transformer-based framework evaluated on the Visual Genome and Open Images benchmarks. All of these methods are confined to a fixed vocabulary at both training and test time.

Open-vocabulary SGG. Pix2Grp [Li *et al.*, 2024] generates scene graph sequences using a VLM, enabling prediction of novel concepts without a fixed output space. LLM4SGG [Kim *et al.*, 2024] uses an LLM to convert image captions into triplet supervision, avoiding manually annotated SGG data but still requiring a training phase. RAHP [Liu *et al.*, 2025] enriches text representations with LLM-generated region-aware prompts inside a contrastive alignment framework. OvSGTR [Chen *et al.*, 2025] investigates three pre-training pipelines (scene-parser-based, LLM-based, and multimodal-LLM-based) to generate weakly supervised annotations and trains a DETR-like model end-to-end, achieving state-of-the-art on VG150. Unlike all of these methods, our system requires **no training**.

Text-prompted segmentation. SAM [Kirillov *et al.*, 2023] established a promptable segmentation paradigm. SAM 3 [Carion *et al.*, 2026] extends this to concept-level prompts, accepting short noun phrases and returning all matching instances in one forward pass. Grounded SAM [Ren *et al.*, 2024] couples Grounding DINO with SAM to enable text-to-segment pipelines via an intermediate detection step. Our pipeline bypasses the detector: LLM-generated noun-phrase prompts feed directly into SAM 3's concept segmentation interface.

Set-of-Marks prompting. Yang et al. [2023] showed that overlaying numbered region marks on an image dramatically improves visual grounding in large multimodal models. We adopt this technique in Stage 5 (Section 3.5), where numbered bounding boxes over SAM 3 instances allow the LLM to unambiguously refer to specific objects when predicting predicates.

3 Method

Our pipeline comprises eight stages organized into three functional blocks: *Segmentation* (Stages 1–2), *Graph Construction* (Stages 3–5), and *Relation Prediction* (Stages 6–8). All LLM calls use temperature $= 0$ for deterministic outputs and all geometric operations are implemented in Python 3.12.

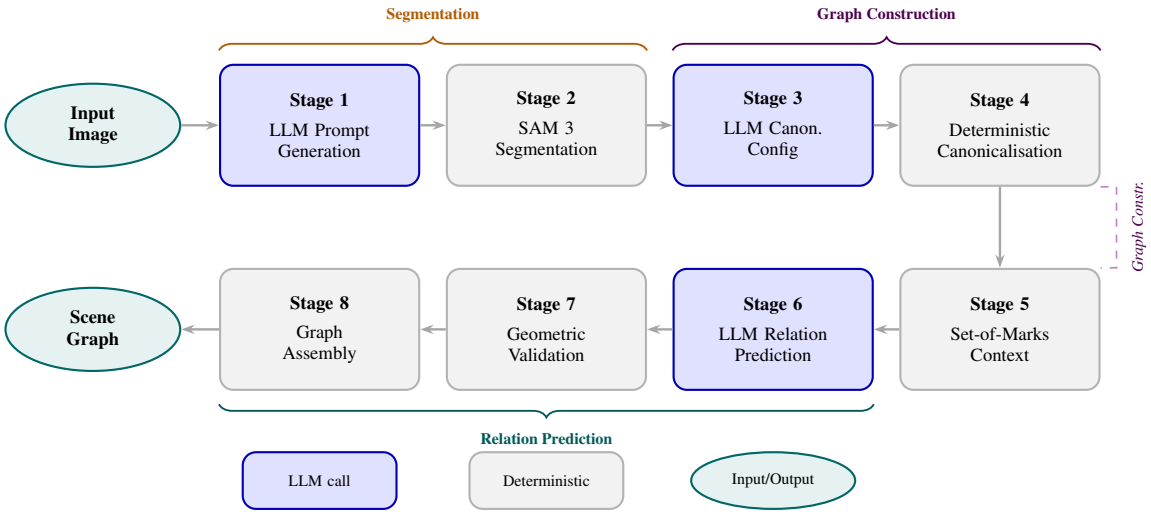

Figure 2: The NEUROSGG pipeline. **Blue** nodes are LLM calls (high-level semantic reasoning); **grey** nodes are deterministic Python operations (geometric grounding); **teal** ellipses are Input/Output endpoints. Stages 1–2 form the *Segmentation* block; Stages 3–5 form the *Graph Construction* block (Stage 5 is in the bottom row but belongs to this block, as indicated by the dashed connector); Stages 6–8 form the *Relation Prediction* block.

3.1 LLM-Driven Prompt Generation

Given an input image, the LLM is prompted to reason about the visual content and return a JSON list of candidate object labels—more concrete, visually grounded noun phrases suitable as SAM 3 concept prompts (*e.g.,* `["wooden desk", "laptop", "person", "window"]`). The LLM is instructed to focus on objects likely to participate in visual relations. This stage is unconstrained by any fixed object taxonomy and may produce multi-word phrases, material-qualified labels, or instance-specific descriptions.

3.2 Concept-Driven Segmentation via SAM 3

The candidate labels are passed as concept prompts to SAM 3 [Carion *et al.*, 2026], which returns instance masks, bounding boxes, and confidence scores for all detected matches as vector geometries. SAM 3's Promptable Concept Segmentation interface provides full-image coverage without requiring a separate object detector. Candidates are filtered by confidence threshold (≥ 0.5) and minimum area ($\geq 25\,\text{px}^2$).

3.3 LLM Canonicalisation Configuration

The raw SAM 3 output often contains label variants, duplicates, and spurious detections. The LLM receives a summary of all candidates (label, score, area, bounding box) and returns a canonicalisation configuration: a mapping from raw labels to canonical node names, together with per-category confidence and area thresholds. Note, that for the sake of evaluation we developed a VG150-constrained mode, in which canonical labels are validated against the VG150 vocabulary, while in the default open-vocabulary mode, any label string is accepted.

3.4 Deterministic Canonicalisation

The LLM-generated configuration is applied deterministically. Candidate masks with IoU > 0.55 are merged, retaining the highest-confidence instance, while masks with a containment ratio > 0.85 are also merged. Each surviving mask becomes a graph node with a unique ID, guaranteeing that every node corresponds to exactly one grounded SAM 3 mask.

3.5 Set-of-Marks Context Construction and LLM Relation Prediction

Canonical nodes are ordered in reading order (top-to-bottom, left-to-right) and assigned numeric identifiers (*e.g.,* `object_001`). Following Yang et al. [2023], numbered bounding boxes are overlaid on the original image to produce a *Set-of-Marks visual prompt*. This image and a compact structured object list are passed to the LLM in Stage 6. The LLM receives the Set-of-Marks image and the compact object list, and predicts pairwise relations as a JSON array of $(\mathit{subject_id},\ \mathit{predicate},\ \mathit{object_id},\ \mathit{confidence})$ tuples. In open-vocabulary mode the LLM may use any predicate string, while in VG150-constrained mode it draws from the VG150 predicate list.

3.6 Deterministic Validation and Scene Graph Assembly

When all predicted triplets are validated their IDs must map to existing nodes, while self-relations are rejected, and confidence scores are clamped to $[0, 1]$. Each remaining relation is enriched with geometric features derived from the corresponding bounding boxes, including IoU, containment ratios, centroid offsets (Δx, Δy, and Euclidean distance), bounding-box distance, overlap fractions, and directional indicators such as *subject-left-of* and *subject-above*. These features are not used for predicate prediction in the current pipeline but are attached for downstream analysis and constitute a foundation for future geometry-supervised relation scoring. Finally, the validated nodes and edges are assembled into a

scene graph and serialized as a JSON file. In cases where the generated output violates the predefined JSON schema, an issue that may arise due to malformed structured generation, incomplete decoding, or inconsistencies in the model output, the LLM performs a single automated repair attempt to restore schema compliance.

4 Experiments

4.1 Setup

As mentioned, we evaluate our approach on the Visual Genome 150 (VG150) benchmark [Krishna *et al.*, 2017; Xu *et al.*, 2017], comprising 150 object categories and 50 predicate types. For each image we report the **Node Recall** which is the fraction of ground-truth objects matched at IoU ≥ 0.5; the **Triplet Precision** describing the fraction of predicted (subject, predicate, object) triplets that are correct; and the **Triplet Recall** representing the fraction of ground-truth triplets recovered. We evaluate three locally hosted open-weights LLMs served via Ollama[1], namely, **Gemma 4 (31B)** and **Qwen 3.5 (122B)**, while no model is fine-tuned and all results are zero-shot.

4.2 Results

Table 1 presents the experimental results of our approach on VG150 across all three evaluated models. To obtain an initial comparison across models with differing characteristics, the evaluation was conducted on the first 10 images of the benchmark only. A more comprehensive evaluation on the full benchmark will be performed for the most promising models in future experiments. Node recall remains relatively stable at 50–53% for all three LLMs, indicating that overall performance is primarily limited by the quality of the SAM 3 segmentations generated from the LLM-conditioned prompts. Gemma 4 achieves the highest triplet recall (7.8%), yielding performance competitive with OvSGTR's Mean R@50 of 7.4%, despite requiring no scene graph generation-specific training. This comparison is particularly informative at the mean recall level, as Mean R@50 weights all 50 predicate categories equally and is therefore less affected by the head-predicate bias inherent to the VG150 annotation distribution. The qualitative analysis shown in Figure 3 is consistent with the quantitative results reported in Table 1. Our approach detects a comparable number of objects to those present in the VG150 ground truth, as illustrated across all examples in Figure 3. However, a substantial proportion of the detected objects is not represented in the benchmark annotations. This behavior is expected, as the model was not explicitly trained on the VG150 object categories and was only prompted to restrict its predictions to the classes defined in the dataset. A similar effect can be observed for relational predicates, which were likewise constrained to the VG150 predicate vocabulary to ensure comparability with the benchmark evaluation protocol.

4.3 VG150 Bias Against Open-Vocabulary

To analyze the bias of VG150 against our open-vocabulary scene graph generation approach, Table 2 shows that the

[1]https://ollama.com/

Figure 3: Exemplary results of NEUROSGG on randomly sampled VG150 images. The left side shows the predicted objects and predicates, while the right side shows the corresponding ground truth annotations. Green indicates correctly detected objects or relational predicates. Red denotes false positive predictions, and blue represents detected objects that do not have a corresponding match in the ground truth. Conversely, orange highlights ground-truth objects and relational predicates that were not detected by the model.

majority of prediction errors produced by our method are concentrated among the four most frequent predicates in the dataset. This observation is consistent with the findings of Zellers et al. [2018], who demonstrated that a frequency-prior baseline, which predicts the most common predicate for each object-pair category, achieves surprisingly strong performance on VG150. Notably, these four predicates account for the majority of all annotations in the dataset. Consequently, a model that predicts only predicates such as on" and in" can obtain deceptively high recall despite exhibiting limited relational understanding. Although the LLM in our approach was explicitly constrained to use the full set of 50 VG150

Table 1: VG150 results for NEUROSGG across three LLMs (10 images each) and OvSGTR [Chen *et al.*, 2025] as a trained reference. †OvSGTR numbers are on Mean R@50 over the full benchmark; our metrics cover all predicted triplets. Best training-free result per column in **bold**.

	Approach	Node Rec.	Trip. Prec.	Trip. Rec.
Ours	Gemma 4 (31B)	52.8	**5.9**	**7.8**
	Qwen 3.5 (122B)	**53.1**	3.9	4.7
	OvSGTR† [Chen *et al.*, 2025]	—	—	7.4

Table 2: Most common false-predicate errors. The top errors are all high-frequency head predicates in VG150, illustrating the annotation-convention mismatch discussed in Section 4.3. "—" indicates the predicate was not among the top errors for that model.

Predicate	Gemma 4	Qwen 3.5
on	76	59
in	46	—
near	10	17
has	10	—

predicate categories, this observation remains important, as restricting the model to only the most frequent predicates would likely yield higher benchmark scores while reducing the semantic richness and expressiveness of the generated scene graphs.

While our approach correctly describes a relation using predicates such as *leaning against*, *resting on top of*, or *partially occluding*, these predictions are still counted as errors under the exact-match triplet evaluation used by VG150. This creates a structural mismatch in which VG150 Recall@K conflates two distinct failure modes: (1) *genuine relational prediction errors*, and (2) *vocabulary misalignment* between the expressive open-vocabulary outputs generated by the LLM and the restricted predicate vocabulary defined by the benchmark annotations. Consequently, we argue that the raw triplet recall values reported in Table 1 should be interpreted as *lower bounds* on the true relational accuracy of our approach and that a fair evaluation of open-vocabulary scene graph generation methods therefore requires a vocabulary-normalized evaluation that maps semantically equivalent predicted predicates to their closest VG150 annotation prior to scoring.

4.4 LLM Ablation Discussion

Our experimental results indicate that Gemma 4 (31B)[2] outperforms the substantially larger Qwen 3.5 (122B)[3] on triplet recall, suggesting that model size alone does not determine relation prediction accuracy. Gemma 4 shows the lower node recall (52.8%), pointing to weaker prompt generation in Stage 1 as a primary upstream bottleneck. These findings motivate future work on optimizing the prompt-generation stage as a first-class concern in training-free SGG pipelines.

[2]https://ai.google.dev/gemma/docs/core?hl=de

[3]https://qwen.ai/blog?id=qwen3.5

5 Conclusion

In this work we present NEUROSGG, a training-free neurosymbolic pipeline for open-vocabulary scene graph generation. By using an LLM to drive concept-level segmentation with SAM 3 and adopting Set-of-Marks visual prompting for relation prediction, NEUROSGG produces fully grounded scene graphs without any task-specific training. Evaluation on VG150 shows node recall competitive with SAM 3's segmentation ceiling ($\sim$53%) and triplet recall comparable to the mean recall of state-of-the-art trained methods—while operating zero-shot on any image and any open vocabulary. Our analysis further demonstrates that standard VG150 metrics structurally underestimate open-vocabulary system performance—a finding with broader implications for how the field benchmarks the next generation of open-world SGG systems. Two natural directions for future work emerge: (i) **geometry-supervised predicate scoring**, using the rich geometric features already attached to each relation in Stage 7 to guide or re-rank LLM predicate predictions; and (ii) **vocabulary-normalised evaluation**, a metric protocol suitable for open-vocabulary SGG that does not conflate genuine errors with annotation-convention mismatches.

References

[Carion *et al.*, 2026] Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. SAM 3: Segment Anything with Concepts. *arXiv preprint arXiv:2511.16719*, 2026.

[Chen *et al.*, 2025] Zuyao Chen, Jinlin Wu, Zhen Lei, Zhaoxiang Zhang, and Chang Wen Chen. Expanding scene graph boundaries: Fully open-vocabulary scene graph generation via visual-concept alignment and retention. In *Computer Vision – ECCV 2024*, pages 108–124, 2025.

[Cong *et al.*, 2023] Yuren Cong, Michael Ying Yang, and Bodo Rosenhahn. RelTR: Relation Transformer for Scene Graph Generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):11169–11183, 2023.

[Hildebrandt *et al.*, 2020] Marcel Hildebrandt, Hang Li, Rajat Koner, Volker Tresp, and Stephan Günnemann. Scene Graph Reasoning for Visual Question Answering. *arXiv preprint arXiv:2007.01072*, 2020.

[Kim *et al.*, 2024] Kibum Kim, Kanghee Yoon, Jaehyeong Park, Seongsu Hong, Suchan Lee, Hyunwoo Cho, Sungho

Kim, and Chanyoung Park. LLM4SGG: Large Language Models for Weakly Supervised Scene Graph Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 28306–28316, 2024.

[Kirillov *et al.*, 2023] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment Anything . In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3992–4003, 2023.

[Krishna *et al.*, 2017] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017.

[Li *et al.*, 2024] Rongjie Li, Songyang Zhang, Dahua Lin, Kai Chen, and Xuming He. From Pixels to Graphs: Open-Vocabulary Scene Graph Generation with Vision-Language Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 28076–28086, 2024.

[Liu *et al.*, 2025] Tao Liu, Rongjie Li, Weidong Wang, and Xuming He. Relation-aware Hierarchical Prompt for Open-vocabulary Scene Graph Generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5576–5584, 2025.

[Ren *et al.*, 2024] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. Grounded SAM: Assembling Open-World Models for Diverse Visual Tasks. *arXiv preprint arXiv:2401.14159*, 2024.

[Singh *et al.*, 2023] Kunal Pratap Singh, Jordi Salvador, Luca Weihs, and Aniruddha Kembhavi. Scene Graph Contrastive Learning for Embodied Navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10884–10894, October 2023.

[Tang *et al.*, 2019] Kaihua Tang, Hanwang Zhang, Baoyuan Wu, Wenhan Luo, and Wei Liu. Learning to Compose Dynamic Tree Structures for Visual Contexts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6619–6628, 2019.

[Xu *et al.*, 2017] Danfei Xu, Yuke Zhu, Christopher B. Choy, and Li Fei-Fei. Scene Graph Generation by Iterative Message Passing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5410–5419, 2017.

[Yang *et al.*, 2019] Xu Yang, Kaihua Tang, Hanwang Zhang, and Jianfei Cai. Auto-Encoding Scene Graphs for Image Captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[Yang *et al.*, 2023] Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chunyuan Li, and Jianfeng Gao. Set-of-Mark Prompting Unleashes Extraordinary Visual Grounding in GPT-4V. *arXiv preprint arXiv:2310.11441*, 2023.

[Zellers *et al.*, 2018] Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. Neural Motifs: Scene Graph Parsing with Global Context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5831–5840, 2018.