PG-VLM: A Multi-Stage Panoptic-Graph Architecture for Detailed Visual-Linguistic Grounding in Urban Scenes

Anonymous authors
Paper under double-blind review

Abstract

Describing complex urban scenes with coherent paragraphs that are both semantically rich and spatially grounded is a key challenge for vision–language research. We present PG-VLM, a modular framework that (i) builds a Hierarchical Panoptic Scene Graph (HPSG) from panoptic segmentation, (ii) distills the graph into semantic triplets using a local instruction model, and (iii) generates narratives with a structured-to-text T5 generator. We assess text quality with standard captioning metrics and grounding with a new Narrative Relevance Detection Score (NRDS) that ties detection correctness to textual mention quality. On Cityscapes, PG-VLM surpasses recent vision-language baselines (BLIP-2, LLaVA-1.5 7B, SpatialVLM) across all metrics: CIDEr 135.0 (vs. 88.0/104.5/118.2), SPICE 28.8 (vs. 19.5/21.2/23.6), and BERTScore-F1 92.5 (vs. 88.0/89.0/90.1). Hallucination is reduced, with CHAIR-s 7.2 and CHAIR-i 9.5 (vs. 16.8/20.5 for BLIP-2, 13.0/16.2 for LLaVA-1.5, 11.4/14.8 for SpatialVLM). PG-VLM achieves substantially higher grounding via NRDS 0.76 compared to BLIP-2 at 0.52. A zero-shot check on BDD100K (50 images) indicates cross-dataset generalization (CIDEr 108.4, SPICE 24.1, NRDS-ZS 0.68), maintaining margins over all baselines. These results show that enforcing a symbolic bottleneck (HPSG to triplets) before generation improves both descriptive quality and faithfulness, offering a reproducible and extensible route to interpretable visual–language grounding in urban scenes.

1 Introduction

Generating paragraph-level descriptions for complex scenes remains challenging. Short captions miss object relations, spatial layout, and context typical of urban images (Ye et al., 2025; Zhang et al., 2024b; Zheng et al., 2025). Modern vision–language models (VLMs) such as BLIP-2, LLaVA, and Flamingo improve general multimodal reasoning (Zhang et al., 2024a; Yin et al., 2023) yet often produce generic, object-centric sentences with weak spatial fidelity (Singh et al., 2024; Qi et al., 2025). Even spatially aware systems including SpatialVLM (Chen et al., 2024) and ARPGrounding (Zeng et al., 2024b) struggle with occlusion, adjacency, and containment in dense layouts (Chan et al., 2023; Peng et al., 2024; Wu et al., 2024; Zhao et al., 2024a). Hallucinations remain a recurrent failure mode across model sizes and training regimes (Li et al., 2023b; Hao et al., 2025; Chang et al., 2024; Leng et al., 2024; Zhao et al., 2024b; Wang et al., 2024). Benchmarks focused on paragraph quality show that coherence and grounded relations are still limited despite large-scale pretraining (Ye et al., 2025; Zheng et al., 2025; Sarto et al., 2025; Liao et al., 2025; Li et al., 2024b; Bieri et al., 2025; Sima et al., 2024).

We present **PG-VLM**, which inserts an explicit symbolic bottleneck between vision and language. From a single urban image, Mask2Former produces panoptic outputs (Cheng et al., 2022) that we lift into a *Hierarchical Panoptic Scene Graph* (HPSG) with explicit spatial and hierarchical edges (Zhou et al., 2024; Miyanishi et al., 2023). A local instruction model converts the HPSG into canonical triplets, and a T5-based decoder maps the triplets to a multi-sentence paragraph. This design targets three pain points: spatial grounding, hallucination, and narrative flow.

On Cityscapes (Cordts et al., 2016), PG-VLM improves all automatic metrics over strong baselines (BLIP-2, LLaVA-1.5, SpatialVLM) and achieves a substantial gain on the *Narrative Relevance*

Detection Score (NRDS), a metric that scores whether narratively important detections are realized in text. A zero-shot check on BDD100K (Yu et al., 2020) indicates robustness to distribution shift. PG-VLM advances structured visual–linguistic grounding by describing not only what is present but also how entities relate within the scene.

1.1 Motivation and Research Gap

Urban driving scenes are visually dense: multiple agents share the road, static infrastructure shapes the layout, and small details such as traffic lights or signs change the interpretation of the scene. Existing vision–language models (VLMs) can recognize many object categories, but their paragraph descriptions still suffer from three recurring issues:

1. **Weak spatial grounding.** Relative positions (“in front of”, “to the left of”, “on the sidewalk”) are often incorrect or omitted, especially in crowded layouts.
2. **Hallucinated content.** Models describe objects or events that are not supported by the image, for example an extra pedestrian or a misplaced vehicle.
3. **Limited interpretability.** End-to-end models map pixels directly to text, making it hard to inspect which visual evidence supports each sentence or to debug errors.

Prior work on detailed captioning and spatial VLMs has partially addressed these issues, for example by injecting bounding boxes into the language model or by conditioning on scene graphs. However, most methods either (i) rely on instance-only graphs without panoptic coverage of stuff regions, (ii) treat the graph as an auxiliary input rather than a true bottleneck, or (iii) evaluate mainly with surface-level metrics such as BLEU or CIDEr that do not directly reflect spatial correctness.

PG-VLM targets this gap with a multi-stage pipeline that *enforces* a symbolic bottleneck between perception and generation. A panoptic backbone builds a hierarchical panoptic scene graph (HPSG) with explicit spatial relations, a local instruction model converts the HPSG into a compact set of semantic triplets, and a T5-based decoder generates a paragraph from these triplets. This design makes it possible to (i) explicitly control which entities and relations enter the narrative, (ii) trace each sentence back to graph elements, and (iii) define an instance-level grounding metric (NRDS) that measures how well narratively important detections are realized in text.

2 Literature Review

VLMs and paragraph generation. Foundational VLMs couple visual encoders with language decoders and show strong zero- and few-shot performance across captioning, VQA, and retrieval (Li et al., 2023a; Alayrac et al., 2022; Koh et al., 2023; Liu et al., 2023a). They typically underperform on multi-sentence description, producing single sentences without narrative coherence or structured grounding (Ye et al., 2025; Singh et al., 2024). Long-form prompting and alignment help (Zhu et al., 2023; Yang et al., 2023) yet spatial reasoning in crowded scenes remains weak (Zhang et al., 2024a; Yin et al., 2023; Wu et al., 2024). Instruction-tuned and spatially aware models, including SpatialVLM (Chen et al., 2024), LLaVA (Liu et al., 2023c), and InstructBLIP (Dai et al., 2023), improve grounding but still miss many relations in dense layouts.

Panoptic segmentation and scene graphs. Transformer-based panoptic models provide strong dense predictions for both stuff and thing categories (Cheng et al., 2022). These outputs motivate panoptic scene graphs that integrate region and instance nodes with relations (Zeng et al., 2024b; Chan et al., 2023; Kim et al., 2024; Nguyen et al., 2024), sometimes aided by LLMs for relation induction (Hayder & He, 2024) or hierarchical refinement (Lee et al., 2024). We adopt an HPSG to encode spatial layout, attributes, and part–whole structure for downstream text generation (Zhou et al., 2024).

Structured generation. Data-to-text methods use intermediate structure such as triples or tables to improve factuality and fluency (Alaçam et al., 2024; Li et al., 2024c;b; Xue et al., 2025). T5-style encoder–decoder transformers are effective for structured-to-text mapping (Liu et al., 2023b). PG-VLM follows this recipe: HPSG to triplets to paragraph, with lightweight planning tags to improve discourse flow.

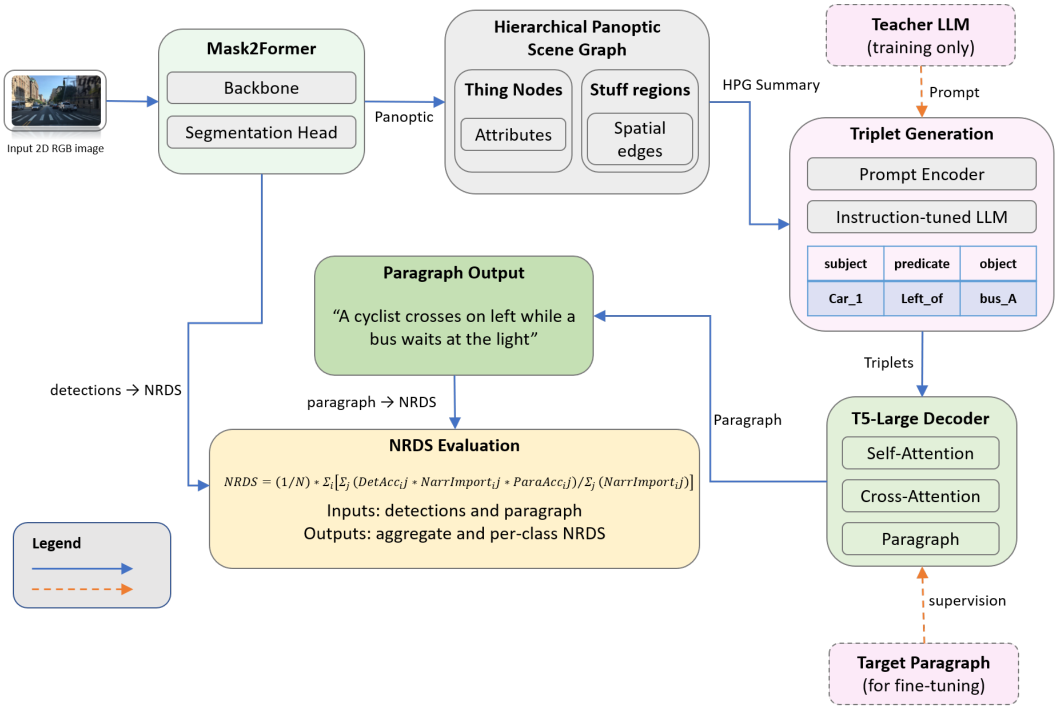


Figure 1: End-to-end PG-VLM pipeline. Mask2Former produces panoptic predictions and multi-scale features. We build an HPSG with thing and stuff nodes and explicit spatial edges, serialize salient nodes and relations, and prompt a local instruction model to produce semantic triplets. A T5-based decoder generates a paragraph from the triplets. NRDS evaluates alignment between detected entities and their narrative realization.

Teacher supervision. Instruction-tuned LLMs provide scalable pseudo-labels when human text is scarce (Chang et al., 2024; Yu et al., 2024; Yanuka et al., 2025; Ma et al., 2024). We use a local instruction model to synthesize paragraph targets from triplets for T5 fine-tuning, aligned with current multimodal instruction practices (Liu et al., 2023c; Dai et al., 2023; Cordts et al., 2016).

Spatial reasoning and evaluation. VLMs frequently mis-handle relative position, occlusion, and containment (Zheng et al., 2025; Peng et al., 2024; Zhao et al., 2024a); geometric priors or prompt design improve some cases (Dorkenwald et al., 2024; Zeng et al., 2024a; González-Chávez et al., 2023). Standard text metrics emphasize lexical overlap and can miss grounding errors (Dorkenwald et al., 2024; Sharma, 2021). Hallucination measures such as CHAIR capture unsupported mentions (Li et al., 2023b;c). We introduce NRDS to reward faithful realization of narratively important detections, complementing surface-level metrics.

3 Methodology: The PG-VLM Framework

PG-VLM converts a single RGB image into a coherent paragraph through three stages: (i) Hierarchical Panoptic Scene Graph (HPSG) construction, (ii) semantic triplet extraction with a local instruction model, and (iii) paragraph decoding with a sequence-to-sequence generator. A brief review of panoptic segmentation, scene graphs, and encoder–decoder transformers is given in Appendix A.1; here we focus on how these components are composed into our end-to-end framework. An overview appears in Fig. 1; a zoomed view of the HPSG appears in Fig. 2.

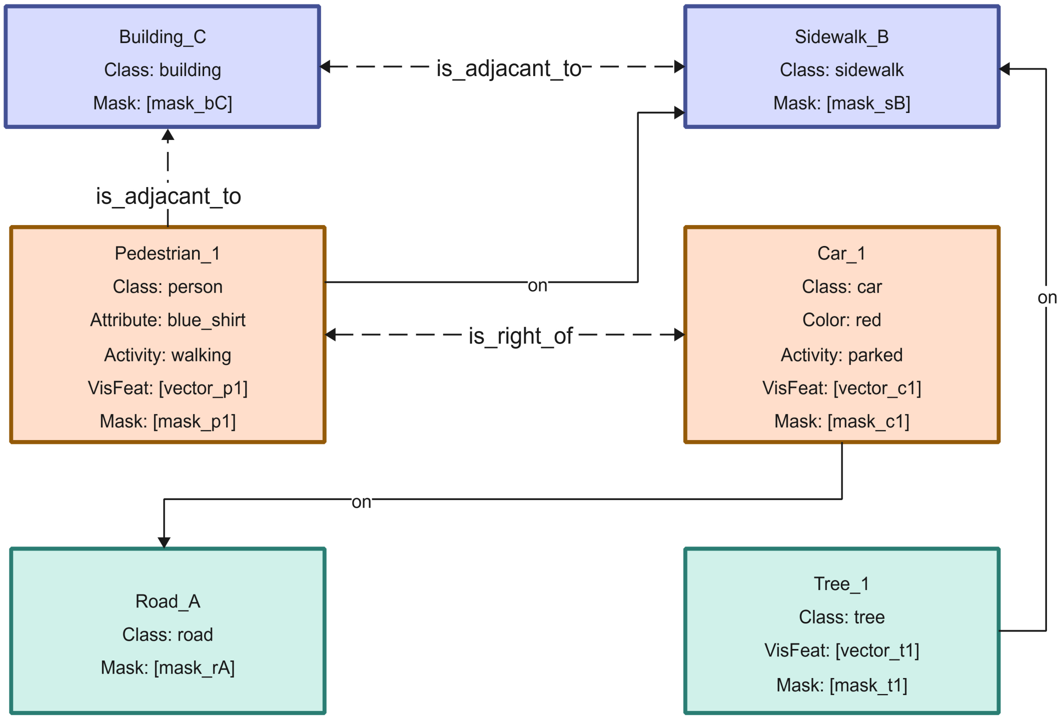


Figure 2: Zoomed view of the HPSG. Nodes store masks, boxes, features, and attributes. Edges encode spatial and hierarchical relations that drive triplet extraction and narrative planning.

3.1 Hierarchical Panoptic Scene Graph Construction

Panoptic backbone. Given a Cityscapes image I (Cordts et al., 2016), we apply Mask2Former with a Swin-L backbone (Cheng et al., 2022). The model outputs segmentation masks $\{M_k\}_{k=1}^{K}$, boxes $\{B_k\}$, class labels $\{c_k\}$, confidence scores $\{s_k\}$, and pyramid features $\{\Phi_\ell\}$.

Nodes. We create nodes for all *thing* instances and *stuff* regions. For node n we store: class $c(n)$, mask $M(n)$, box $B(n)$, a pooled feature

$$f(n) = \mathrm{RoIAlign}(\{\Phi_\ell\},\, B(n))\,,$$

and optional unary attributes $\mathcal{A}(n)$ (for example color or state) derived from rule-based probes applied to $f(n)$.

Edges. We compute pairwise spatial relations using mask and box geometry. The predicate set is

$$\begin{aligned}\mathcal{R} = \{&\texttt{left\_of},\ \texttt{right\_of},\ \texttt{in\_front\_of},\ \texttt{behind},\\ &\texttt{on},\ \texttt{inside},\ \texttt{adjacent\_to},\ \texttt{near},\\ &\texttt{overlaps},\ \texttt{occludes},\ \texttt{part\_of},\ \texttt{contacts}\}\end{aligned}$$

Each relation $r \in \mathcal{R}$ is scored by a differentiable geometric functional $\gamma_r(n_i, n_j)$; we keep the top-k outgoing edges per node and predicate, and suppress symmetric duplicates. The resulting HPSG $G = (N, E)$ preserves layout while remaining compact for downstream text generation. A conceptual view is shown in Fig. 2.

3.2 Semantic Triplet Generation via a Local Instruction Model

Serialization. We linearize G into a compact summary that lists salient nodes with aliases, attributes, and high-confidence edges. The summary groups facts by layout elements (road, sidewalk, building) and by agents (car, bus, person, rider), and orders agent entries from left to right using instance centroids.

Instruction model. A locally hosted instruction-tuned LLM (see Section 4.5) receives the summary and emits a set of canonical *semantic triplets* $\mathcal{T} = \{(s, p, o)\}$ where $p \in \mathcal{R}$ and s, o are node aliases. We constrain the predicate vocabulary to $\mathcal{R}$ and provide few-shot exemplars for format stability, following instruction-tuning practices used in LLaVA and InstructBLIP (Liu et al., 2023c; Dai et al., 2023).

Filtering and canonicalization. Naively forming all subject–predicate–object combinations from G leads to a large and noisy set of triplets, many of which correspond to redundant or visually marginal relations. We therefore apply a two-stage filter. First, we discard triplets whose predicate is outside $\mathcal{R}$, whose arguments are not present in G, or whose edge score falls below a predicate-specific threshold. Symmetric predicates (e.g., `left_of`/`right_of`) are canonicalized and, when we obtain contradictory pairs, we keep the higher-confidence item.

Second, we rank the remaining triplets by a salience score that combines (i) the geometric confidence of the underlying edge, (ii) a predicate prior that favours relations such as `on`, `inside` and `in_front_of`, and (iii) the degree centrality of the incident nodes in the HPSG. We keep the top $K = 40$ triplets per image. This compact, high-confidence set $\mathcal{T}$ is then linearized into $\mathrm{Lin}(\mathcal{T})$, which forms the input to the decoder.

3.3 Structured Paragraph Generation with T5-Large

Model and input format. We use T5-Large (Liu et al., 2023b). The input is the ordered list of filtered triplets interleaved with lightweight planning tags:

$$\underbrace{[\text{LAYOUT}]\ (\text{road}, \texttt{adjacent\_to}, \text{sidewalk})\ \cdots}_{\text{layout block}}\ \ \underbrace{[\text{AGENTS}]\ (\text{car\_1}, \texttt{on}, \text{road})\ \cdots}_{\text{agent block}}$$

followed by compact unary attribute hints (for example `car_1[color=blue]`). The vocabulary includes class aliases and predicate tokens, which encourages faithful lexicalization.

Teacher-generated targets. For each training image, the same local instruction model is prompted with $\mathrm{Lin}(\mathcal{T})$ to produce a target paragraph that is concise and spatially grounded. A validator removes sentences mentioning entities absent from $\mathcal{T}$, which yields consistent $(\mathrm{Lin}(\mathcal{T}) \rightarrow P)$ pairs for supervision. This follows data-to-text planning and entity modeling best practices (Puduppully et al., 2022; 2019).

Learning and decoding. We minimize token-level negative log-likelihood with label smoothing 0.1. At inference we use beam search ($B = 4$) with length penalty 0.8. A constrained post-checker downranks beams that introduce unsupported entities relative to $\mathcal{T}$ while preserving fluency.

Hyperparameters. Inputs and targets are limited to 512 tokens. Training uses AdamW with learning rate 3×10^{-5}, weight decay 0.01, and an effective batch size of 32 via gradient accumulation, for 20 epochs. HPSG thresholds and the predicate set $\mathcal{R}$ are fixed before decoder training.

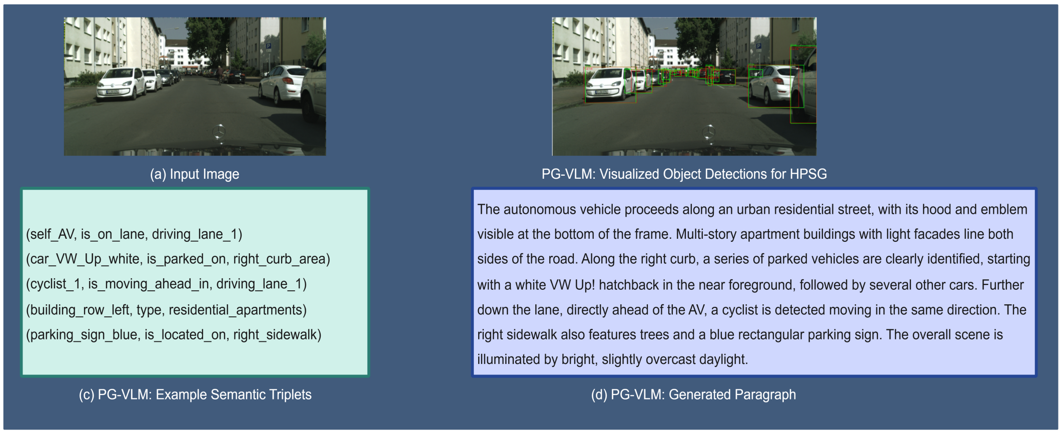
3.4 Qualitative Illustration

Fig. 3 presents an example image with detected instances, filtered triplets, and the final paragraph. The narrative reflects key entities and relations from the HPSG and avoids unsupported mentions.

The T5-Large decoder is trained with a standard token-level cross-entropy loss with label smoothing; full optimization and loss details are provided in Appendix A.2.

4 Experimental Setup

We evaluate PG-VLM on Cityscapes with a zero-shot transfer check on BDD100K. This section details datasets, baselines, training configuration, and metrics.



(a) Input Image

PG-VLM: Visualized Object Detections for HPSG

(self_AV, is_on_lane, driving_lane_1)
(car_VW_Up_white, is_parked_on, right_curb_area)
(cyclist_1, is_moving_ahead_in, driving_lane_1)
(building_row_left, type, residential_apartments)
(parking_sign_blue, is_located_on, right_sidewalk)

(c) PG-VLM: Example Semantic Triplets

The autonomous vehicle proceeds along an urban residential street, with its hood and emblem visible at the bottom of the frame. Multi-story apartment buildings with light facades line both sides of the road. Along the right curb, a series of parked vehicles are clearly identified, starting with a white VW Up! hatchback in the near foreground, followed by several other cars. Further down the lane, directly ahead of the AV, a cyclist is detected moving in the same direction. The right sidewalk also features trees and a blue rectangular parking sign. The overall scene is illuminated by bright, slightly overcast daylight.

(d) PG-VLM: Generated Paragraph

Figure 3: Qualitative example. (a) Input image, (b) detected instances and stuff regions, (c) filtered triplets, (d) generated paragraph.

4.1 Datasets

Cityscapes. 5,000 finely annotated street-view images with 30 semantic classes (Cordts et al., 2016). We use the standard split: 2,975 train, 500 val, 1,525 test. Panoptic labels directly support HPSG construction.

BDD100K. A diverse driving dataset; we use a 50-image daytime subset for cross-dataset evaluation without tuning (Yu et al., 2020).

4.2 Baselines

- **BLIP-2** (Li et al., 2023a): vision–language model with a Q-Former bridge.
- **LLaVA-1.5 (7B)** (Liu et al., 2023c): instruction-tuned multimodal model.
- **SpatialVLM** (Chen et al., 2024): VLM with explicit spatial grounding.

All baselines share the same preprocessing, decoding, and metric scripts.

4.3 Training Configuration

The triplet-to-paragraph decoder is T5-Large (770M). We train with AdamW (lr 3×10^{-5}, weight decay 0.01), effective batch size 32 via gradient accumulation, sequence length 512 (input/output), for 20 epochs with early stopping on validation CIDEr. The predicate set and HPSG thresholds are fixed before training. Experiments run on a single NVIDIA RTX 4090 (24 GB).

4.4 Evaluation Metrics

Text-based metrics. We first report standard captioning metrics: BLEU-1–4, ROUGE-L, METEOR, CIDEr, and SPICE, together with BERTScore-F1. These scores capture lexical overlap and semantic similarity between generated paragraphs and the reference texts.

Hallucination metrics. To assess hallucination, we use CHAIR-s/i and Entity-Precision. CHAIR-s measures the fraction of sentences that mention at least one unsupported object, while CHAIR-i measures the fraction of object mentions that are not grounded in the image. Entity-Precision counts the proportion of mentioned entities that correspond to ground-truth instances.

Narrative Relevance Detection Score (NRDS). The above metrics are largely text-based and do not directly test whether visually important instances are described correctly. To address this, we use the *Narrative Relevance Detection Score* (NRDS), an instance-level grounding metric that links panoptic detections to noun phrases in the paragraph.

For a validation set of N images with detection sets $\{D_i\}_{i=1}^N$, the overall NRDS is

$$\mathrm{NRDS} \;=\; \frac{1}{N} \sum_{i=1}^{N} \frac{\sum_{j \in D_i} \big(\mathrm{DetAcc}_j \cdot \mathrm{NarrImport}_j \cdot \mathrm{ParaAcc}_j\big)}{\mathrm{TotalNarrImport}_i}. \tag{1}$$

Here, $\mathrm{DetAcc}_j \in \{0, 1\}$ indicates whether detection j matches a ground-truth instance of the same class with IoU > 0.5. $\mathrm{NarrImport}_j$ weights the narrative importance of the class of j; less frequent classes receive higher weight, so rare but important entities (e.g., buses, riders) contribute more. $\mathrm{ParaAcc}_j \in [0, 1]$ measures how well j is realized in the paragraph: we build a short reference phrase from the HPSG attributes of j, find matching noun phrases in the generated text using class-specific aliases, and take the maximum CLIP similarity between the image crop of j and those spans. The denominator $\mathrm{TotalNarrImport}_i$ is the sum of narrative weights over all narratively relevant ground-truth instances in image i.

NRDS is related to CLIPScore but differs in two ways. CLIPScore computes a single image–paragraph similarity, whereas NRDS operates at the level of individual instances and aggregates their contributions. In addition, NRDS explicitly factors in detection correctness and class-dependent importance, so a model cannot obtain a high NRDS by describing irrelevant or incorrectly detected content. Unlike CIDEr and other text-only metrics, NRDS couples visual detection quality with the narrative and directly reflects whether important objects in the scene are mentioned accurately.

4.5 Training Data, Pseudo-Labels, and Evaluation Protocol

Cityscapes does not provide paragraph-level descriptions, so we adopt a teacher–student scheme for structured-to-text training. For each training image, we first build an HPSG and extract semantic triplets as described in Section 3. Unless otherwise stated, the local instruction model (the *teacher*) is `LLaMA-2-7B-Chat` (**?**), an instruction-tuned 7B LLaMA-family model run locally with 16-bit precision. We use the public checkpoint without any additional fine-tuning on Cityscapes or other driving datasets. The teacher receives the triplets and produces a concise, spatially grounded paragraph; an automatic validator removes sentences that mention entities not present in the input triplets. The resulting paragraphs serve as *pseudo-labels* for the T5 decoder and as reference texts for all models (PG-VLM and baselines) when computing automatic captioning metrics (CIDEr, SPICE, BERTScore, BLEU, ROUGE-L, METEOR). Baselines (BLIP-2, LLaVA-1.5, SpatialVLM) are evaluated in their publicly released form without any fine-tuning on these references.

This protocol can introduce a bias in favour of PG-VLM, since the pseudo-labels reflect the teacher's style and the structure of the triplets. To partially control for this, we (i) report hallucination metrics (CHAIR-s/i, Entity-Precision) and NRDS, which depend on image–text alignment rather than lexical similarity to the teacher, and (ii) conduct a blind human evaluation in which annotators compare model outputs given the *image* rather than the teacher text; results are reported in Appendix A.8.

5 Results

We report automatic metrics on Cityscapes and a cross-dataset check on BDD100K. PG-VLM is compared against SpatialVLM, LLaVA-1.5, and BLIP-2 under identical evaluation settings.

5.1 Automatic Evaluation on Cityscapes

PG-VLM yields consistent gains across n-gram overlap (BLEU), long-form coverage (ROUGE-L), semantic adequacy (METEOR, SPICE, BERTScore), and consensus-based relevance (CIDEr). Improvements are most pronounced in CIDEr and SPICE, which are sensitive to content selection and relational details. The pattern suggests that the HPSG→triplet bottleneck improves both *what* is said (entity/attribute coverage) and *how* it is organized (coherent, multi-sentence narratives). Qualitatively, we observe better grounding of spatial predicates and fewer missing scene elements, which matches the larger margins on SPICE and BERTScore.

Table 1: Automatic evaluation of paragraph generation on Cityscapes. Higher is better.

Metric	PG-VLM	SpatialVLM	LLaVA-1.5	BLIP-2
BLEU-1	78.2	72.0	69.5	65.1
BLEU-2	63.5	57.8	54.2	48.8
BLEU-3	51.0	45.6	42.1	36.0
BLEU-4	40.8	36.0	32.8	27.3
ROUGE-L	60.5	55.2	52.4	49.2
METEOR	31.5	28.1	26.3	24.0
CIDEr	135.0	118.2	104.5	88.0
SPICE	28.8	23.6	21.2	19.5
BERTScore (F1)	92.5	90.1	89.0	88.0

Table 2: Hallucination and entity grounding on Cityscapes. Lower is better for CHAIR; higher is better for Entity-Precision.

Metric	PG-VLM	SpatialVLM	LLaVA-1.5	BLIP-2
CHAIR-s ↓	7.2	11.4	13.0	16.8
CHAIR-i ↓	9.5	14.8	16.2	20.5
Entity-Precision ↑	89.6	84.2	81.5	77.0

5.2 Hallucination and Entity Grounding

PG-VLM reduces unsupported entity mentions (CHAIR-s/i) while improving correct mentions (Entity-Precision). This trend aligns with the model's explicit symbolic bottleneck: triplets constrain lexical realization to entities and relations present in the graph, curbing free-form drift and yielding stronger image–text fidelity.

5.3 Narrative Relevance Detection Score

NRDS measures whether narratively important visual entities are realized faithfully in text (Section 4.4). On Cityscapes validation, PG-VLM achieves **0.76** versus **0.52** for BLIP-2, indicating better alignment between detection, narrative salience, and paragraph realization. We observe a strong correlation between NRDS and CIDEr across samples, while cases with high CIDEr but low NRDS typically omit salient instances or introduce subtle hallucinations—scenarios where PG-VLM shows fewer failures.

5.4 Cross-dataset Generalization

PG-VLM maintains an advantage without dataset-specific tuning. The HPSG→triplet abstraction helps preserve core semantics and spatial relations under shifts in camera geometry and urban context, supporting generalization beyond the training distribution.

6 Ablations and Analysis

We conduct controlled experiments on Cityscapes validation to isolate contributions of each PG-VLM component, keeping training settings constant with the main model.

6.1 Effect of the HPSG Bottleneck

Removing the graph stage and feeding visual features directly to the decoder reduces content selection quality (CIDEr), relational coverage (SPICE), and faithfulness (NRDS), while doubling unsupported mentions (CHAIR). The symbolic bottleneck forces early disambiguation of entities and relations, which translates into more reliable downstream text.

Table 3: Zero-shot transfer to BDD100K daytime subset (50 images).

Model	**CIDEr** ↑	**SPICE** ↑	**NRDS-ZS** ↑
PG-VLM	108.4	24.1	0.68
SpatialVLM	95.7	20.3	0.57
LLaVA-1.5 (7B)	88.9	18.4	0.51
BLIP-2	73.2	16.2	0.42

Table 4: Impact of the structured HPSG bottleneck. Removing HPSG weakens spatial grounding and increases hallucination.

Model Variant	CIDEr ↑	SPICE ↑	CHAIR-s ↓	NRDS ↑
PG-VLM (full)	135.0	28.8	7.2	0.76
Direct ViT → T5	112.3	22.5	13.4	0.59

6.2 Triplet Filtering and Salience

Restricting the input to high-confidence, non-redundant triplets improves precision and NRDS. The decoder benefits from a concise fact plan rather than an over-complete set that can encourage repetition or off-target elaboration.

6.3 Decoder Variants and Planning Signals

We compare decoder scales and planning tags in Appendix A.5, showing that lightweight layout-first, left-to-right agent ordering improves ROUGE-L, SPICE, and NRDS without adding trainable planners.

6.4 Sensitivity to Predicate Inventory and Triplet Budget

We vary the predicate inventory and the triplet budget K. Extremely small inventories collapse distinct spatial cues (e.g., merging `in_front_of` with `left_of`), reducing SPICE and NRDS. Increasing K beyond the chosen setting offers little benefit and can slightly degrade fluency due to verbosity. The default inventory and K strike a practical balance between coverage and concision.

6.5 Qualitative Analysis

Qualitative inspection (see Figure 3) shows that PG-VLM consistently realizes spatial relations (e.g., *car beside bus*, *pedestrian on sidewalk*) and avoids unsupported mentions that appear in end-to-end baselines. We also observe better narrative cohesion across sentences, aided by triplet ordering and grouping.

7 Conclusion

PG-VLM introduces a structured pipeline for paragraph-level scene understanding that combines panoptic segmentation, a hierarchical panoptic scene graph, semantic triplets, and a sequence-to-sequence decoder. By grounding generation in explicit symbolic facts, PG-VLM improves spatial fidelity and reduces hallucination compared with recent vision–language models. On Cityscapes, it delivers consistent gains across captioning and semantic metrics and lowers hallucination rates, while maintaining an advantage under zero-shot transfer to BDD100K. The Narrative Relevance Detection Score complements traditional metrics by emphasizing alignment between detected entities and their textual realization. Altogether, the results support the value of an HPSG→triplet bottleneck for reliable, coherent narratives of complex urban scenes.

References

Özge Alaçam, Ronja Utescher, Hannes Grönner, Judith Sieker, and Sina Zarrieß. WikiScenes with descriptions: Aligning paragraphs and sentences with images in Wikipedia articles. In Danushka Bollegala and Vered Shwartz (eds.), *Proceedings of the 13th Joint Conference on Lexical and Computational Semantics (*SEM 2024)*, pp. 93–105, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.starsem-1.8. URL `https://aclanthology.org/2024.starsem-1.8/`.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

Valentin Bieri, Marco Zamboni, Nicolas S Blumer, Qingxuan Chen, and Francis Engelmann. Opencity3d: What do vision-language models know about urban environments? In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 5147–5155. IEEE, 2025.

David Chan, Suzanne Petryk, Joseph E Gonzalez, Trevor Darrell, and John Canny. Clair: Evaluating image captions with large language models. *arXiv preprint arXiv:2310.12971*, 2023.

Chun-Peng Chang, Shaoxiang Wang, Alain Pagani, and Didier Stricker. Mikasa: Multi-key-anchor & scene-aware transformer for 3d visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14131–14140, 2024.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14455–14465, 2024.

Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1290–1299, 2022.

Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL `https://openreview.net/forum?id=vvoWPYqZJA`.

Michael Dorkenwald, Nimrod Barazani, Cees G. M. Snoek, and Yuki M. Asano. Pin: Positional insert unlocks object localisation abilities in vlms, 2024. URL `https://arxiv.org/abs/2402.08657`.

Othón González-Chávez, Guillermo Ruiz, Daniela Moctezuma, and Tania A. Ramirez-delReal. Are metrics measuring what they should? an evaluation of image captioning task metrics, 2023. URL `https://arxiv.org/abs/2207.01733`.

Yunzhuo Hao, Jiawei Gu, Huichen Will Wang, Linjie Li, Zhengyuan Yang, Lijuan Wang, and Yu Cheng. Can mllms reason in multimodality? emma: An enhanced multimodal reasoning benchmark. *arXiv preprint arXiv:2501.05444*, 2025.

Zeeshan Hayder and Xuming He. Dsgg: Dense relation transformer for an end-to-end scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28317–28326, 2024.

Kibum Kim, Kanghoon Yoon, Jaehyeong Jeon, Yeonjun In, Jinyoung Moon, Donghyun 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*, pp. 28306–28316, 2024.

Jing Yu Koh, Ruslan Salakhutdinov, and Daniel Fried. Grounding language models to images for multimodal inputs and outputs. In *International Conference on Machine Learning*, pp. 17283–17300. PMLR, 2023.

Yukyung Lee, Soonwon Ka, Bokyung Son, Pilsung Kang, and Jaewook Kang. Navigating the path of writing: Outline-guided text generation with large language models. *CoRR*, abs/2404.13919, 2024. URL https://doi.org/10.48550/arXiv.2404.13919.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13872–13882, 2024.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PMLR, 2023a.

Li Li, Wei Ji, Yiming Wu, Mengze Li, You Qin, Lina Wei, and Roger Zimmermann. Panoptic scene graph generation with semantics-prototype learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 3145–3153, 2024a.

Rui Li, Qi Liu, Liyang He, Zheng Zhang, Hao Zhang, Shengyu Ye, Junyu Lu, and Zhenya Huang. Optimizing code retrieval: High-quality and scalable dataset annotation through large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 2053–2065, 2024b.

Rui Li, Qi Liu, Liyang He, Zheng Zhang, Hao Zhang, Shengyu Ye, Junyu Lu, and Zhenya Huang. Optimizing code retrieval: High-quality and scalable dataset annotation through large language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 2053–2065, Miami, Florida, USA, November 2024c. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.123. URL https://aclanthology.org/2024.emnlp-main.123/.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023b.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models, 2023c. URL https://arxiv.org/abs/2305.10355.

Xiaoxuan Liao, Binrong Zhu, Jacky He, Guiran Liu, Hongye Zheng, and Jia Gao. A fine-tuning approach for t5 using knowledge graphs to address complex tasks. *arXiv preprint arXiv:2502.16484*, 2025.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023a. URL https://arxiv.org/abs/2304.08485.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023b. URL https://arxiv.org/abs/2304.08485.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023c. URL https://arxiv.org/abs/2304.08485.

Chenyang Ma, Kai Lu, Ta-Ying Cheng, Niki Trigoni, and Andrew Markham. Spatialpin: Enhancing spatial reasoning capabilities of vision-language models through prompting and interacting 3d priors, 2024. URL https://arxiv.org/abs/2403.13438.

Taiki Miyanishi, Fumiya Kitamori, Shuhei Kurita, Jungdae Lee, Motoaki Kawanabe, and Nakamasa Inoue. Cityrefer: geography-aware 3d visual grounding dataset on city-scale point cloud data. *arXiv preprint arXiv:2310.18773*, 2023.

Trong-Thuan Nguyen, Pha Nguyen, and Khoa Luu. Hig: Hierarchical interlacement graph approach to scene graph generation in video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18384–18394, 2024.

Wujian Peng, Sicheng Xie, Zuyao You, Shiyi Lan, and Zuxuan Wu. Synthesize diagnose and optimize: Towards fine-grained vision-language understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13279–13288, 2024.

Ratish Puduppully, Li Dong, and Mirella Lapata. Data-to-text generation with entity modeling, 2019. URL `https://arxiv.org/abs/1906.03221`.

Ratish Puduppully, Yao Fu, and Mirella Lapata. Data-to-text generation with variational sequential planning, 2022. URL `https://arxiv.org/abs/2202.13756`.

Jianing Qi, Jiawei Liu, Hao Tang, and Zhigang Zhu. Beyond semantics: Rediscovering spatial awareness in vision-language models. *arXiv preprint arXiv:2503.17349*, 2025.

Sara Sarto, Marcella Cornia, and Rita Cucchiara. Image captioning evaluation in the age of multimodal llms: Challenges and future perspectives. *arXiv preprint arXiv:2503.14604*, 2025.

Himanshu Sharma. A survey on image captioning datasets and evaluation metrics. *IOP Conference Series: Materials Science and Engineering*, 1116, 2021. URL `https://api.semanticscholar.org/CorpusID:235292921`.

Chonghao Sima, Katrin Renz, Kashyap Chitta, Li Chen, Hanxue Zhang, Chengen Xie, Jens Beißwenger, Ping Luo, Andreas Geiger, and Hongyang Li. Drivelm: Driving with graph visual question answering. In *European Conference on Computer Vision*, pp. 256–274. Springer, 2024.

Hrishikesh Singh, Aarti Sharma, and Millie Pant. Pixels to prose: Understanding the art of image captioning. *arXiv preprint arXiv:2408.15714*, 2024.

Xintong Wang, Jingheng Pan, Liang Ding, and Chris Biemann. Mitigating hallucinations in large vision-language models with instruction contrastive decoding. *arXiv preprint arXiv:2403.18715*, 2024.

Junfei Wu, Qiang Liu, Ding Wang, Jinghao Zhang, Shu Wu, Liang Wang, and Tieniu Tan. Logical closed loop: Uncovering object hallucinations in large vision-language models. *arXiv preprint arXiv:2402.11622*, 2024.

Haochen Xue, Feilong Tang, Ming Hu, Yexin Liu, Qidong Huang, Yulong Li, Chengzhi Liu, Zhongxing Xu, Chong Zhang, Chun-Mei Feng, Yutong Xie, Imran Razzak, Zongyuan Ge, Jionglong Su, Junjun He, and Yu Qiao. Mmrc: A large-scale benchmark for understanding multimodal large language model in real-world conversation, 2025. URL `https://arxiv.org/abs/2502.11903`.

Xu Yang, Yongliang Wu, Mingzhuo Yang, Haokun Chen, and Xin Geng. Exploring diverse in-context configurations for image captioning. *Advances in Neural Information Processing Systems*, 36: 40924–40943, 2023.

Moran Yanuka, Assaf Ben Kish, Yonatan Bitton, Idan Szpektor, and Raja Giryes. Bridging the visual gap: Fine-tuning multimodal models with knowledge-adapted captions, 2025. URL `https://arxiv.org/abs/2411.09018`.

Qinghao Ye, Xianhan Zeng, Fu Li, Chunyuan Li, and Haoqi Fan. Painting with words: Elevating detailed image captioning with benchmark and alignment learning. *arXiv preprint arXiv:2503.07906*, 2025.

Zhenfei Yin, Jiong Wang, Jianjian Cao, Zhelun Shi, Dingning Liu, Mukai Li, Xiaoshui Huang, Zhiyong Wang, Lu Sheng, Lei Bai, et al. Lamm: Language-assisted multi-modal instruction-tuning dataset, framework, and benchmark. *Advances in Neural Information Processing Systems*, 36: 26650–26685, 2023.

Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning, 2020. URL `https://arxiv.org/abs/1805.04687`.

Qiying Yu, Quan Sun, Xiaosong Zhang, Yufeng Cui, Fan Zhang, Yue Cao, Xinlong Wang, and Jingjing Liu. Capsfusion: Rethinking image-text data at scale, 2024. URL https://arxiv.org/abs/2310.20550.

Yunan Zeng, Yan Huang, Jinjin Zhang, Zequn Jie, Zhenhua Chai, and Liang Wang. Investigating compositional challenges in vision-language models for visual grounding. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14141–14151, 2024a. doi: 10.1109/CVPR52733.2024.01341.

Yunan Zeng, Yan Huang, Jinjin Zhang, Zequn Jie, Zhenhua Chai, and Liang Wang. Investigating compositional challenges in vision-language models for visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14141–14151, 2024b.

Beichen Zhang, Pan Zhang, Xiaoyi Dong, Yuhang Zang, and Jiaqi Wang. Long-clip: Unlocking the long-text capability of clip. In *European Conference on Computer Vision*, pp. 310–325. Springer, 2024a.

Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024b.

Fengzhi Zhao, Zhezhou Yu, Tao Wang, and Yi Lv. Image captioning based on semantic scenes. *Entropy*, 26(10):876, 2024a.

Linxi Zhao, Yihe Deng, Weitong Zhang, and Quanquan Gu. Mitigating object hallucination in large vision-language models via classifier-free guidance. *arXiv preprint arXiv:2402.08680*, 2024b.

Xu Zheng, Ziqiao Weng, Yuanhuiyi Lyu, Lutao Jiang, Haiwei Xue, Bin Ren, Danda Paudel, Nicu Sebe, Luc Van Gool, and Xuming Hu. Retrieval augmented generation and understanding in vision: A survey and new outlook. *arXiv preprint arXiv:2503.18016*, 2025.

Zijian Zhou, Zheng Zhu, Holger Caesar, and Miaojing Shi. Openpsg: Open-set panoptic scene graph generation via large multimodal models. In *European Conference on Computer Vision*, pp. 199–215. Springer, 2024.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

A Appendix

A.1 Preliminaries

This section summarizes the standard building blocks used by PG-VLM: panoptic segmentation for dense visual recognition, scene graphs for structured reasoning, and encoder–decoder transformers for text generation. The goal is to fix notation and clarify how these components interface in our pipeline.

Panoptic segmentation. Panoptic segmentation assigns a semantic label to every pixel while distinguishing instance identities for countable categories. Let $I \in \mathbb{R}^{H \times W \times 3}$ be an input image. A panoptic model produces $(\boldsymbol{M}, \boldsymbol{Y})$, where $\boldsymbol{M} = \{M_k\}_{k=1}^{K}$ are binary masks and $\boldsymbol{Y} = \{(c_k, s_k)\}_{k=1}^{K}$ are per-instance class labels c_k and scores s_k. Stuff regions such as road or sidewalk are represented as single masks without an instance index. We use Mask2Former with a Swin backbone (Cheng et al., 2022) to obtain masks, boxes, and multi-scale features that seed graph construction. In PG-VLM, these panoptic outputs are the only direct interface between raw pixels and the symbolic scene graph.

Scene graph representations. A scene graph is a directed multigraph $G = (N, E)$ with node set N and labeled edges $E \subseteq N \times \mathcal{R} \times N$, where $\mathcal{R}$ is a predicate set. In a panoptic scene graph (Li et al., 2024a), nodes include both *thing* instances (cars, pedestrians, riders) and *stuff* regions (road, sidewalk, building). For a node $n \in N$, we store $(c(n), M(n), B(n), f(n))$: semantic class, mask, bounding box, and a pooled visual feature. PG-VLM constructs a *Hierarchical Panoptic Scene Graph*

(HPSG) in which edges cover spatial relations (e.g., `left_of`, `on`, `in_front_of`), containment and part–whole links, and selected unary attributes as key–value tags. The HPSG serves as the sole structured input to the language stack and provides explicit hooks for reasoning about layout and interactions.

Sequence-to-sequence models. Given a linearized structured input X_{struct} and a target paragraph $P = (s_1, \ldots, s_T)$, an encoder–decoder transformer models

$$p(P \mid X_{\text{struct}}) = \prod_{t=1}^{T} p(s_t \mid s_{<t}, X_{\text{struct}}).$$

We instantiate the decoder with T5-Large (Liu et al., 2023b) and format X_{struct} as a sequence of canonical triplets plus compact attribute hints. This separation of perception (panoptic segmentation) from generation (seq2seq) allows the model to focus on content selection and discourse structure while preserving explicit grounding hooks through the triplets. In particular, entities and relations that are not present in the HPSG cannot enter the input sequence, which constrains the decoder and reduces free-form hallucination.

A.2 Reproducibility and Implementation Details

Environment. PyTorch 2.2, CUDA 12.x, Transformers 4.41, Accelerate 0.30, SentencePiece 0.1.99. Mixed precision (bfloat16) for all model stages. Fixed global seed 1234 for data shuffling, sampler, and decoder initialization.

Preprocessing. Cityscapes frames are resized to 1024×512 (longer side 1024), normalized to backbone statistics, and batched by image area. Panoptic maps use the official evaluation format; category aliases follow the dataset label policy.

HPSG construction. Mask2Former (Swin-L) outputs are filtered at confidence $\geq$ 0.5 for *thing* instances. Edges are created from masks and boxes: `on`, `inside`, `left_of`, `right_of`, `in_front_of`, `behind`, `adjacent_to`, `near`, `overlaps`, `occludes`, `part_of`, `contacts`. We apply per-predicate non-maximum suppression using IoU on relation supports (union of incident masks) and retain top-$k = 6$ outgoing edges per node per predicate.

Triplet serializer and filter. Nodes are serialized with {id, class, mask-derived attributes (e.g., color hue bin), size bin, centroid}. We keep up to $K = 40$ triplets ranked by (predicate prior, edge confidence, degree centrality). Symmetric duplicates (`left_of`/`right_of`) are canonicalized; cycles on `on`/`inside` are pruned.

Decoder training. T5-Large; AdamW (3×10^{-5}, weight decay 0.01), label smoothing 0.1, batch size 32 (via gradient accumulation), 512 tokens input/output, 20 epochs with early stopping on validation CIDEr. Beam search size 4, length penalty 0.8. A constrained post-checker downranks beams with entities not present in the input triplets.

Training objective. The teacher LLM is kept fixed; we do not fine-tune it on Cityscapes. Its role is to (i) generate semantic triplets from HPSG summaries during training-time data preparation and (ii) synthesize target paragraphs from those triplets. The learnable component in PG-VLM is the T5-Large decoder. Given an input sequence $\mathrm{Lin}(\mathcal{T})$ of filtered triplets and a target paragraph P, the model is trained with a standard token-level cross-entropy loss

$$\mathcal{L} = -\sum_{t=1}^{T} \log p_\theta(s_t \mid s_{<t}, \mathrm{Lin}(\mathcal{T})),$$

with label smoothing 0.1. We use AdamW with learning rate 3×10^{-5}, weight decay 0.01, and an effective batch size of 32. NRDS is used only for evaluation and model selection; it does not appear in the loss.

A.3 Prompt Templates

Triplet extraction (HPSG→Triplets).

> **System:** You receive a scene summary with nodes and relations. Return a list of triplets in the canonical form `(subject, predicate, object)` using this predicate set: `on, inside, left_of, right_of, in_front_of, behind, adjacent_to, near, overlaps, occludes, part_of, contacts.` Use node ids (e.g., `car_3`, `road_1`). Do not invent objects.
>
> **User:** `NODES:` {...} `EDGES:` {...}
> **Assistant:** `(car_3, on, road_1);(person_5, near, crosswalk_1);`...

Paragraph teacher (Triplets→Paragraph).

> **System:** Write a concise paragraph (3–6 sentences) describing the scene using only the provided triplets and attributes. Preserve spatial relations and avoid objects not listed. Maintain a natural flow: layout first, then agents left to right.
>
> **User:** `TRIPLETS:` {...} `ATTRS:` {...}
> **Assistant:** *The road stretches ahead with buildings on both sides...*

A.4 Narrative Relevance Detection Score (NRDS)

Standard captioning metrics emphasize lexical overlap and often miss whether narratively important visual entities are faithfully realized in text. NRDS quantifies visual–text alignment by combining detection correctness, class-level narrative importance, and paragraph realization.

For a validation set of N images with detection sets $\{D_i\}_{i=1}^{N}$, the overall NRDS is

$$\text{NRDS} \;=\; \frac{1}{N}\sum_{i=1}^{N}\frac{\sum_{j\in D_i}\big(\text{DetAcc}_j\cdot\text{NarrImport}_j\cdot\text{ParaAcc}_j\big)}{\text{TotalNarrImport}_i}. \tag{2}$$

Components.

- D_i is the set of detected instances for image i.
- $\text{DetAcc}_j \in \{0,1\}$ is 1 if detected instance j matches a ground-truth instance of the same class with IoU > 0.5, else 0.
- NarrImport_j weights class importance. Let c be the class of instance j and F_c the number of class-c instances in training. We set $\text{NarrImport}_j = 1/\sqrt{F_c}$ to emphasize less frequent classes without extreme weights.
- $\text{ParaAcc}_j \in [0,1]$ measures textual realization. We construct a short reference phrase from HPSG attributes/relations for j, locate candidate spans via class aliases and noun-phrase chunking, compute CLIP cosine similarity between the image crop of j and each span, take the maximum, and normalize to $[0,1]$.
- TotalNarrImport_i is the sum of NarrImport_j over narratively relevant ground-truth instances in image i.

Implementation details. Detections and ground truth are matched per class with one-to-one assignment. Crops are taken from instance masks with small padding; class aliases include common synonyms for Cityscapes classes. We report aggregate NRDS and per-class NRDS.

A.5 Additional Quantitative Results

Table 6: Per-class NRDS for key *thing* categories. Higher is better.

Class	Car	Person	Bicycle	Bus	Truck	Rider	Motorcycle	Train
NRDS	0.82	0.73	0.65	0.80	0.60	0.58	0.55	0.75

Per-class NRDS on Cityscapes (validation).

Table 7: Paragraph length control with beam search. Tokens approximate mean output length.

Setting	Tokens	CIDEr ↑	SPICE ↑	CHAIR-s ↓
Short (minlen)	85	126.7	27.1	8.0
Default	120	135.0	28.8	7.2
Long (no cap)	165	133.9	28.6	7.8

Length control versus quality.

Table 8: Effect of ordering/grouping on discourse and grounding.

Input Order	ROUGE-L ↑	SPICE ↑	CHAIR-s ↓	NRDS ↑
Shuffled	58.9	27.4	8.1	0.72
Layout→Agents (ours)	60.5	28.8	7.2	0.76

Triplet ordering and grouping.

CIDEr–NRDS correlation. Across validation samples, Pearson $r = 0.88$, indicating strong alignment. Outliers with high CIDEr but lower NRDS typically omit a salient entity or conflate two adjacent agents, which the symbolic bottleneck mitigates.

A.6 Robustness Analyses

Occlusion sensitivity. We bin instances by occlusion ratio (mask overlap with other instances).

Table 9: NRDS versus occlusion level.

Occlusion bin	$[0, 0.2)$	$[0.2, 0.5)$	$[0.5, 1.0]$
NRDS	0.79	0.74	0.66

Lighting/contrast perturbation. We apply uniform brightness/contrast jitters at inference time (no retraining).

A.7 Runtime and Memory

Throughput is ∼4.1 FPS in steady state with batched Mask2Former inference. VRAM peaks when panoptic features and the decoder are both resident; gradient-free inference limits memory pressure.

A.8 Human Evaluation Protocol

Three annotators evaluated 150 validation images (balanced across density). Each paragraph was scored on 1–5 Likert scales for *Fluency*, *Relevance*, and *Spatial correctness*. Items were randomized per annotator; model identities were blinded. Krippendorff's $\alpha = 0.62$ averaged across dimensions.

A.9 Error Analysis

- **Small-instance miss:** rare classes (e.g., *rider*, *motorcycle*) missed under heavy occlusion can lead to under-reporting. Mitigation: predicate smoothing and per-class triplet priors.
- **Attribute drift:** color/state mismatches for traffic lights at distance. Mitigation: attribute confidence gating and span-level fallback.
- **Relation ambiguity:** camera parallax confuses `in_front_of` vs. `left_of`. Mitigation: tie-break by orientation priors and vanishing-point heuristics.

Table 10: NRDS under test-time brightness/contrast jitter.

Perturbation	None	Mild	Strong
NRDS	0.76	0.74	0.70

Table 11: Per-image inference on RTX 4090 (24GB) at 1024×512.

Stage	Latency (ms)	Peak VRAM (GB)
Mask2Former panoptic	85	9.2
HPSG build (CPU+GPU mix)	15	0.4
Triplet LLM (local)	90	2.1
T5-Large decoding (beam=4)	55	3.1
Total	**245**	**14.8**

- **Noun-phrase alignment collisions:** aliases overlapping with other classes (e.g., *rider*/*person*) reduce ParaAcc. Mitigation: class-conditional lexicons.

A.10 Post-checker and Re-ranking Details

The post-checker computes entity coverage $\mathcal{C}$ and unsupported mentions $\mathcal{U}$ per candidate beam. The final score is

$$s_{\text{final}} = s_{\text{beam}} + \lambda_{\text{cov}} \cdot \mathcal{C} - \lambda_{\text{unsup}} \cdot \mathcal{U},$$

with $\lambda_{\text{cov}} = 0.6$, $\lambda_{\text{unsup}} = 0.8$. $\mathcal{C}$ counts unique input entities realized at least once; $\mathcal{U}$ counts span-level entities not in the input triplets (exact/alias match). This re-ranking consistently reduces CHAIR while preserving fluency.

A.11 Dataset and Licensing Notes

Cityscapes (Cordts et al., 2016) and BDD100K (Yu et al., 2020) are used under their respective academic licenses. Faces and license plates in figures are blurred per dataset policy. No personal data is newly collected.

A.12 Broader Impact and Risk Controls (Extended)

We include a brief operational guidance: require minimum thresholds on CHAIR and NRDS before any deployment; maintain human-in-the-loop review; log coverage and confidence per paragraph; apply privacy filters to sensitive details; continuously evaluate across diverse cities, weather, and times of day.

A.13 Additional Qualitative Examples

We provide more end-to-end examples (image, HPSG snippet, triplets, paragraph) to illustrate typical successes and remaining challenges. See Figure 3 in the main paper and the supplemental gallery provided with the code release.

Limitations and future work. PG-VLM depends on the quality of the panoptic backbone and the teacher LLM: severe panoptic failures or biased teacher paragraphs can propagate to the triplets and the final narrative. Our evaluation is also centered on urban driving scenes (Cityscapes and a small BDD100K subset), so generalization to other domains remains to be tested. Finally, NRDS, while more grounded than text-only metrics, still relies on CLIP and heuristic phrase matching. Extending PG-VLM and NRDS to a broader set of datasets and detection backbones is an important direction for future work.

Code and model release. Upon publication we will release the full PG-VLM codebase, including scripts for HPSG construction, triplet extraction, paragraph generation, and NRDS computation. We

Table 12: Mean human ratings (1–5). Higher is better.

Model	Fluency	Relevance	Spatial Corr.
PG-VLM	4.3	4.1	4.0
SpatialVLM	3.9	3.7	3.6
LLaVA-1.5	3.8	3.6	3.4
BLIP-2	3.6	3.4	3.1

will also release configuration files, pretrained weights for the T5-Large decoder, and evaluation scripts used to obtain all reported numbers. Where licenses permit, we will additionally provide the panoptic backbone checkpoint and teacher- generated training paragraphs for Cityscapes to facilitate exact reproduction of our experiments.