SD-LLaVA: A Dual-Stream Framework for Semantic-Augmented BEV Understanding

Anonymous authors
Paper under double-blind review

Abstract

Large vision-language models (LVLMs) remain unreliable on Bird's-Eye-View (BEV) semantic maps because their visual encoders are optimized for natural-image appearance, whereas BEV semantics are represented through discrete color labels, spatial layouts, and geometric conventions. We present SD-LLaVA, a dual-stream architecture that augments CLIP patch tokens with structured, region-aware semantic prior tokens. Its SemanticPriorEncoder partitions each raw BEV map into four diagonal regions, extracts class-specific color masks, represents object-like categories using regional counts and stuff categories using presence indicators, and projects the resulting statistics into the language-model embedding space. We further formulate BEV question answering with objective factual queries and subjective driving-oriented queries, enabling separate evaluation of factual grounding and open-ended response quality. On a BEVCar-derived benchmark, SD-LLaVA achieves 90.97% semantic accuracy, compared with 29.59% for an independently trained CLIP-only variant and 73.42% for a semantic-only variant. For subjective QA, it obtains an average rubric-based judge score of 69.96, outperforming CLIP-only (45.30) and semantic-only (46.86). These results demonstrate that continuous visual representations and explicit semantic priors provide complementary signals for BEV-grounded vision-language reasoning.

1 Introduction

Bird's-Eye-View (BEV)representations Philion & Fidler (2020); Li et al. (2024b); Liu et al. (2023c); Chen et al. (2017); Yang et al. (2018) have become a central scene abstraction in autonomous driving, as they provide a unified top-down coordinate system for organizing road layouts, drivable areas, lane structures, traffic participants, and surrounding obstacles. Compared with perspective images, BEV semantic maps offer a more spatially aligned representation for downstream driving tasks, making them suitable for region-level understanding, spatial relation reasoning, and short-horizon decision support Li et al. (2023); Sima et al. (2024); Tian et al. (2025). Meanwhile, Large Vision-Language Models (LVLMs), such as LLaVA Liu et al. (2023a; 2024b); Li et al. (2024a), have demonstrated strong instruction-following and natural-language reasoning capabilities, suggesting a promising interface for interpreting BEV driving scenes through language Sima et al. (2024); Ishaq et al. (2025); Mohan et al. (2025). However, directly applying general-purpose LVLMs to BEV semantic maps Choudhary et al. (2024); Sima et al. (2024); Mao et al. (2023); Shao et al. (2024); Xu et al. (2024) remains unreliable. Standard visual encoders are primarily trained on natural images, where texture, object appearance, and local visual patterns dominate representation learning, whereas BEV maps encode semantics through structured color labels, geometric layouts, and domain-specific spatial conventions. This structural mismatch makes it difficult for conventional LVLMs to preserve explicit region-level facts, spatial orientation, and safety-relevant relations, which are essential for reliable BEV-grounded driving understanding Liu et al. (2024a;c); Li et al. (2025); Choudhary et al. (2024); Ding et al. (2024); Xu et al. (2025). In this work, we make three contributions. First, we propose SD-LLaVA, a semantic dual-stream LVLM framework that combines CLIP visual tokens with explicit BEV semantic prior tokens. Second, we introduce a SemanticPriorEncoder that converts color-coded BEV maps into region-aware learnable semantic tokens. Third, we construct a task-aware BEV QA formulation covering objective factual questions and subjective driving-oriented reasoning, enabling separate evaluation of factual grounding and open-ended response quality.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

2 METHOD

2.1 OVERVIEW: SEMANTIC DUAL-STREAM BEV REASONING

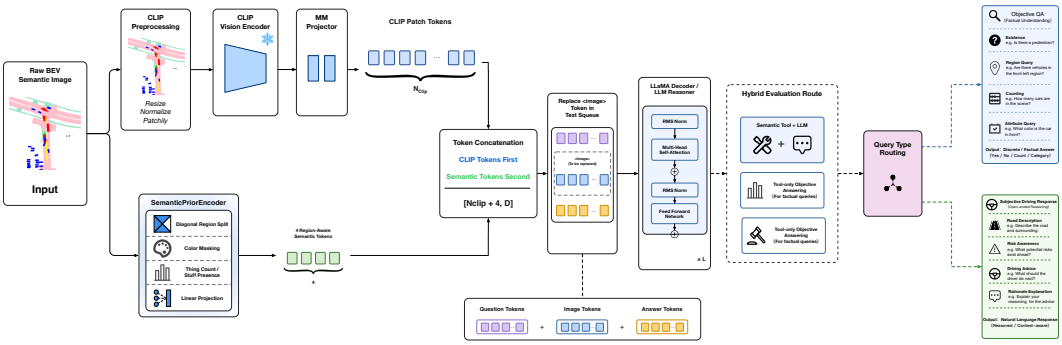


Figure 1: Overall architecture of SD-LLaVA. The raw BEV semantic image is processed by two complementary streams: a CLIP visual stream that produces continuous patch tokens and a SemanticPriorEncoder stream that produces region-aware semantic prior tokens. The two token sequences are concatenated and inserted into the LLM input sequence for BEV-grounded reasoning. A task-aware QA protocol separates objective factual answering from subjective driving-oriented response generation.

We propose SD-LLaVA, a semantic dual-stream framework for BEV-grounded multimodal reasoning, as illustrated in Figure 1. The key observation is that BEV semantic maps differ from natural images in both representation form and reasoning requirements. Natural images mainly encode object appearance, texture, and local visual patterns, while BEV semantic maps encode scene structure through symbolic color labels, spatial layout, region occupancy, and driving-specific geometric conventions. Therefore, directly compressing a BEV map into standard visual tokens may obscure explicit facts that are crucial for BEV reasoning, such as the existence of vehicles in a particular region, the presence of lane dividers, or the spatial distribution of drivable areas. SD-LLaVA addresses this mismatch by processing the same BEV input through two complementary streams: a CLIP Radford et al. (2021) visual stream that preserves continuous visual context and a SemanticPriorEncoder stream that extracts region-aware semantic priors from the raw BEV map. The resulting CLIP patch tokens and semantic prior tokens are concatenated and inserted into the LLM input sequence, enabling the language model to jointly reason over visual context and explicit BEV semantics.

Given a BEV semantic image I and a natural-language query q, SD-LLaVA generates an answer y by decomposing the image-conditioned representation into two complementary streams:

$$Z_{\text{img}} = [Z_{\text{vis}}; Z_{\text{sem}}], \tag{1}$$

where Z_{vis} denotes continuous visual tokens and Z_{sem} denotes structured semantic prior tokens.

For each sample, the same BEV frame is represented in two synchronized forms: a processed image $I_{proc}$for the visual encoder and a raw RGB image I_{raw} for exact semantic parsing. The visual stream is defined as

$$Z_{\mathbf{vis}} = g_{\mathbf{mm}}(f_{\mathbf{vis}}(I_{\mathbf{proc}})). \tag{2}$$

where f_{vis} is the vision encoder and g_{mm} projects visual features into the language model embedding space. In our implementation with CLIP ViT-L/14-336, the visual stream typically produces 576 patch tokens for a single-view BEV input Liu et al. (2023a); Philion & Fidler (2020). The semantic stream produces four additional semantic prior tokens, resulting in a fused image representation with 580 tokens. The two streams are concatenated in the order of visual tokens followed by semantic tokens.

This dual-stream formulation allows the language model to jointly attend to continuous visual context and explicit BEV-specific symbolic evidence, rather than relying solely on latent visual features.

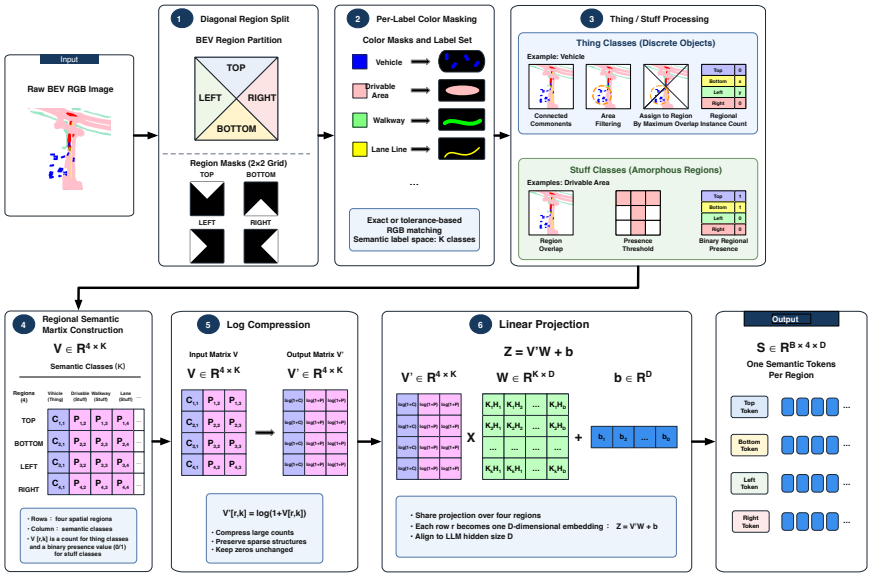


Figure 2: Detailed pipeline of the SemanticPriorEncoder. The raw BEV RGB image is first partitioned into four diagonal spatial regions. Per-label color filtering extracts semantic masks according to the BEV palette. Thing classes are converted into regional object counts, while stuff classes are converted into regional presence indicators. The resulting regional semantic matrix is log-compressed and linearly projected into the LLM embedding space, producing one semantic prior token for each BEV region.

2.2 SEMANTICPRIORENCODER

As illustrated in Figure 2, the SemanticPriorEncoder converts a raw BEV semantic image into compact and learnable semantic prior tokens through a region-aware semantic vectorization pipeline. Instead of treating RGB values as appearance signals, it interprets them as symbolic category labels defined by a BEV semantic palette. The encoder first partitions the BEV image into coarse spatial regions, then extracts per-label masks, computes region-level thing counts and stuff presence indicators, and finally projects the resulting semantic matrix into the LLM embedding space.

The encoder first constructs a fixed label space from the semantic palette. In the current implementation, the palette contains $K = 9$ categories: Ego Car, Pedestrian Crossing, Walkway, Drivable Area, Road Divider, Vehicle, Stop Line, Lane Line, and Carpark Area. For each BEV image, the encoder builds a region-label statistic matrix

$$\mathbf{V} \in \mathbb{R}^{\mathbf{4}\times\mathbf{K}}, \tag{3}$$

where the four rows correspond to top, bottom, left, and right BEV regions, and the columns correspond to the predefined semantic labels. Each entry records either the count or presence of a semantic category in a region.

The spatial regions are obtained by splitting the BEV image with two diagonals, producing top, bottom, left, and right partitions. This provides a coarse but effective spatial reference frame aligned with driving language, such as front, rear, left, and right. For each semantic label, the encoder performs color-mask parsing on the raw RGB image. Object-like categories and region-like categories are then handled differently: object-like categories, currently Vehicle, are counted using 8-connected components with a minimum area threshold; region-like categories are represented by binary presence if their mask area exceeds a predefined threshold.

After this parsing step, the statistic matrix V is logarithmically compressed and projected into the LLM embedding space:

$$\tilde{\mathbf{V}} = \log(1 + \mathbf{V}), \quad \mathbf{Z}_{\text{sem}} = \tilde{\mathbf{V}}\mathbf{W}_{\text{sem}} + \mathbf{1}\mathbf{b}_{\text{sem}}^{\top}, \tag{4}$$

where $\mathbf{V} \in \mathbb{R}^{4\times K}$, $\tilde{\mathbf{V}} \in \mathbb{R}^{4\times K}$, $\mathbf{W}_{\text{sem}} \in \mathbb{R}^{D\times K}$, $\mathbf{b}_{\text{sem}} \in \mathbb{R}^{D}$, and $\mathbf{Z}_{\text{sem}} \in \mathbb{R}^{4\times D}$. Here, K is the number of BEV semantic categories, D is the LLM hidden size, and $\mathbf{1} \in \mathbb{R}^4$ broadcasts the bias term to the four region tokens. For a batch of size B, the semantic stream maps region-label statistics of shape $[B, 4, K]$ into semantic tokens of shape $[B, 4, D]$.

This design preserves interpretability at the input side, since each dimension of V corresponds to a known BEV category, while still allowing the model to learn how these structured priors should be aligned with language-model representations.

2.3 Token Fusion and Training Objective

After obtaining visual tokens and semantic prior tokens, SD-LLaVA inserts the fused representation into the multimodal language model. The original image placeholder in the text sequence is replaced by the fused image representation defined in Equation 1.

If the query contains L text tokens including one image placeholder, the final sequence length becomes $L_{\text{final}} = L - 1 + N_v + 4$, where N_v is the number of visual tokens and 4 is the number of semantic prior tokens. The inserted visual and semantic token spans are masked in the language modeling labels, and the model is trained to generate the target response autoregressively:

$$\mathcal{L}_{\text{LM}} = -\sum_{t=1}^{T} \log p_\theta(y_t \mid y_{<t}, q, \mathbf{Z}_{\text{vis}}, \mathbf{Z}_{\text{sem}}). \tag{5}$$

This training objective remains compatible with standard LVLM instruction tuning, while the image representation is augmented with explicit BEV semantic priors.

This formulation also supports three separately trained variants for controlled comparison: a visual-only model using only Z_{vis}, a semantic-only model using only Z_{sem}, and the full dual-stream model using both. Each variant is trained independently under the same BEV QA setup, isolating the contribution of each stream.

2.4 Task-Aware BEV QA and Model-Internal Answer Routing

To support both factual grounding and open-ended driving reasoning, we construct a task-aware BEV QA formulation in which each BEV image is paired with objective factual questions and subjective driving-oriented questions. Objective questions cover element existence, absent elements, region-level presence, vehicle existence, vehicle counting, and region-element containment. Subjective questions cover road-condition description and short-horizon driving advice.

These two question types require different answer behaviors. Objective questions expect constrained and verifiable outputs, such as element sets, yes/no answers, or integer counts, while subjective questions require free-form natural-language responses grounded in the BEV scene. Instead of introducing a separately trained routing network, we rely on the instruction-following behavior of the LLaVA decoder. Given the fused BEV representation and the task-specific question prompt, the model generates either parseable factual answers for objective QA or open-ended driving-oriented responses for subjective QA.

This protocol connects data construction, inference, and evaluation without adding router parameters or separate router supervision. Objective outputs are evaluated with strict text accuracy, semantic accuracy, and type-wise metrics, whereas subjective outputs are evaluated by a rubric-based judge along understanding, risk awareness, safety actionability, and clarity.

2.5 Ablation Study

We first conduct ablation experiments to examine whether the two streams in SD-LLaVA provide complementary information. The ablation study compares CLIP-Only, Semantic-Only, and the full SD-LLaVA model, each trained independently on the same BEV QA training set under the same hyperparameter setting.

2.5.1 OBJECTIVE QA ABLATION STUDY

Objective QA evaluates verifiable BEV factual reasoning. The questions cover whole-image element existence, absent element recognition, region-level element presence, region-level vehicle existence, regional vehicle counting, whole-image vehicle counting, and random region-element containment. These questions require parseable outputs, such as element sets, yes/no answers, or integer counts. Although the objective QA labels are automatically derived from BEV semantic maps, the ground-truth answers are used only for training supervision and evaluation. During inference, the model receives only the BEV image, the corresponding visual and/or semantic tokens, and the natural-language question. We report four objective metrics: Semantic Accuracy, Elements Accuracy, Yes/No Accuracy, and Counting Accuracy. Semantic Accuracy measures whether the parsed meaning of the model output matches the ground-truth answer. Elements Accuracy evaluates whether the model correctly identifies semantic elements in the BEV map. Yes/No Accuracy evaluates binary factual judgments, and Counting Accuracy evaluates numerical vehicle-counting questions. Table 1

Table 1: Objective QA ablation study based on LLaVA-v1.6-13B.

Experiment	**Semantic Acc.**	**Elements Acc.**	**Yes/No Acc.**	**Counting Acc.**
CLIP-Only	29.59%	1.70%	61.68%	5.31%
Semantic-Only	73.42%	45.75%	83.22%	88.98%
SD-LLaVA	90.97%	90.14%	87.64%	97.96%

reports the objective QA ablation results. CLIP-Only achieves only 29.59% Semantic Accuracy, with particularly weak Elements Accuracy and Counting Accuracy. This indicates that the standard visual pathway alone does not reliably preserve the symbolic facts encoded in BEV semantic maps. Semantic-Only substantially improves factual reasoning, achieving 73.42% Semantic Accuracy and 88.98% Counting Accuracy. This shows that explicit semantic prior tokens provide strong grounding signals for objective BEV questions. The full SD-LLaVA model achieves the best overall result across all four metrics, reaching 90.97% Semantic Accuracy, 90.14% Elements Accuracy, 87.64% Yes/No Accuracy, and 97.96% Counting Accuracy. Compared with CLIP-Only, SD-LLaVA improves Semantic Accuracy by 61.38 percentage points; compared with Semantic-Only, it further improves Semantic Accuracy by 17.55 percentage points. These results show that CLIP visual tokens and semantic prior tokens are complementary for objective BEV reasoning.

2.5.2 SUBJECTIVE QA ABLATION STUDY

Subjective QA evaluates open-ended driving-oriented reasoning. Each BEV image is paired with two subjective questions: one for road-condition description and one for short-horizon driving advice. Unlike objective QA, these questions do not have a single parseable answer. They require the model to generate natural-language responses that are grounded in the BEV scene and useful for driving-oriented interpretation. We use a GPT-based judge Liu et al. (2023b); Gu et al. (2024) to evaluate subjective responses against reference answers. The judge is given the question, the reference response, and the model prediction, and assigns scores along four dimensions: Understanding, Risk Awareness, Safety Actionability, and Clarity. Understanding measures whether the response correctly captures the BEV scene semantics. Risk Awareness measures whether potential hazards are identified. Safety Actionability measures whether the generated advice is concrete and useful. Clarity measures linguistic coherence and readability. The final subjective score is computed as the average over the four dimensions. The complete judge prompt, scoring rubric, decoding configuration, and evaluation template are provided in the appendix. Table 2 reports the subjective QA ablation results. CLIP-Only obtains an average score of 45.30, while Semantic-Only obtains 46.86. The two single-stream variants are close in overall subjective performance, suggesting that neither visual tokens nor semantic prior tokens alone are sufficient for high-quality driving-oriented responses. The full SD-LLaVA model achieves an average score of 69.96, outperforming Semantic-Only by 23.10 points and CLIP-Only by 24.66 points. It also obtains the best score across all four dimensions: 76.12 in Understanding, 65.92 in Risk Awareness, 59.55 in Safety Actionability, and 78.27 in Clarity. These results indicate that open-ended BEV driving reasoning benefits from both structured semantic grounding and visual-language generation ability.

Table 2: Subjective QA ablation study based on LLaVA-v1.6-13B.

Experiment	Understanding	Risk Awareness	Safety Actionability	Clarity	Average
CLIP-Only	47.30	36.33	38.06	59.52	45.30
Semantic-Only	47.50	38.62	40.63	60.70	46.86
SD-LLaVA	76.12	65.92	59.55	78.27	69.96

2.6 Comparison Study

We further apply the SD-LLaVA dual-stream architecture to different LLaVA backbones to examine whether the benefit of the dual-stream design is stable across base models, rather than being tied to a specific backbone size or LLaVA version. For all backbones, we build the full SD-LLaVA architecture and fine-tune it on the proposed BEV QA training set under the same setting, keeping the architecture and training data fixed while varying only the backbone.

2.6.1 Objective QA Comparison Across Model Weights

Table 3 compares objective QA performance across the three backbones. Semantic Accuracy increases consistently from 78.83% on LLaVA-v1.5-7B to 85.46% on LLaVA-v1.5-13B and 90.97% on LLaVA-v1.6-13B, indicating that stronger language backbones translate into better BEV factual reasoning under the same dual-stream design. The LLaVA-v1.6-13B backbone achieves the best result on all four metrics, reaching 90.97% Semantic Accuracy, 90.14% Elements Accuracy, 87.64% Yes/No Accuracy, and 97.96% Counting Accuracy. Across all three backbones, the dual-stream architecture maintains strong objective grounding, showing that its benefit for BEV factual reasoning is not tied to a particular backbone and is strongest with the most capable base model.

Table 3: Objective QA comparison of the SD-LLaVA dual-stream architecture across different LLaVA backbones.

Backbone	Semantic Acc.	Elements Acc.	Yes/No Acc.	Counting Acc.
LLaVA-v1.5-7B	78.83%	75.85%	83.45%	74.08%
LLaVA-v1.5-13B	85.46%	85.54%	78.00%	94.78%
LLaVA-v1.6-13B	90.97%	90.14%	87.64%	97.96%

2.6.2 Subjective QA Comparison Across Model Weights

Table 4 presents subjective QA results across the three backbones. The LLaVA-v1.6-13B backbone performs best by a clear margin, with an average score of 69.96, while the two v1.5 backbones obtain lower and closely spaced averages of 55.88 (7B) and 53.29 (13B). With the LLaVA-v1.6-13B backbone, SD-LLaVA obtains the best score in every subjective dimension: 76.12 in Understanding, 65.92 in Risk Awareness, 59.55 in Safety Actionability, and 78.27 in Clarity. The subjective scores do not increase monotonically with model size alone, as the v1.5-13B backbone is slightly lower than v1.5-7B, but the stronger v1.6 backbone yields a clear improvement in open-ended driving-oriented responses. Overall, these results indicate that the dual-stream design delivers consistent gains in subjective BEV reasoning across backbones, with the strongest backbone giving the best interpretation and driving-oriented generation quality.

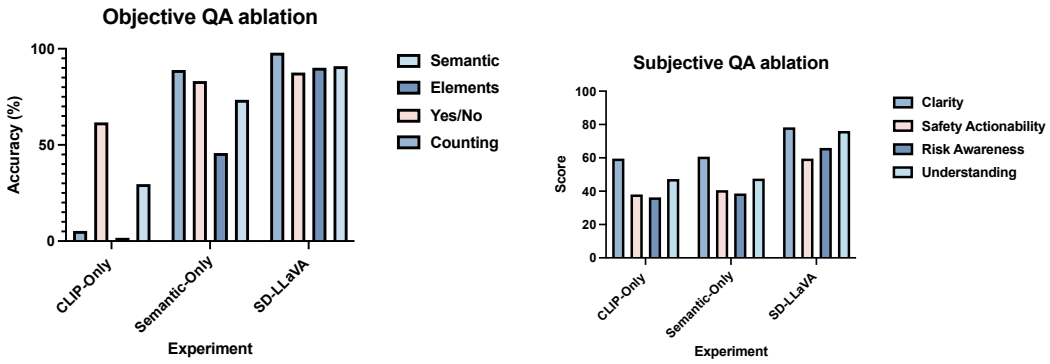
Table 4: Subjective QA comparison of the SD-LLaVA dual-stream architecture across different LLaVA backbones.

Backbone	Understanding	Risk Awareness	Safety Actionability	Clarity	Average
LLaVA-v1.5-7B	59.49	54.69	49.10	60.23	55.88
LLaVA-v1.5-13B	56.58	48.01	46.27	62.29	53.29
LLaVA-v1.6-13B	76.12	65.92	59.55	78.27	69.96

3 EXPERIMENT

4 ANALYSIS AND DISCUSSION

4.1 EFFECTIVENESS OF THE DUAL-STREAM ARCHITECTURE

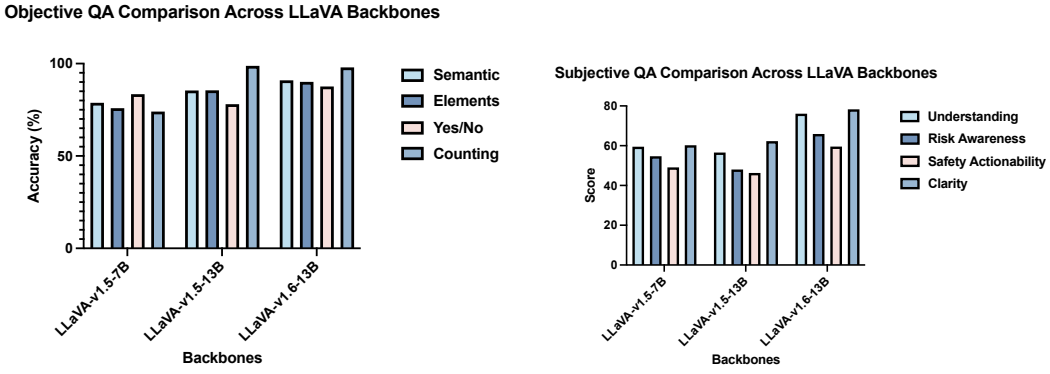


(a) Objective QA ablation. (b) Subjective QA ablation.

Figure 3: Ablation analysis of objective and subjective QA. The results show that SD-LLaVA outperforms both Semantic-Only and CLIP-Only variants, indicating the complementarity between structured semantic priors and visual tokens.

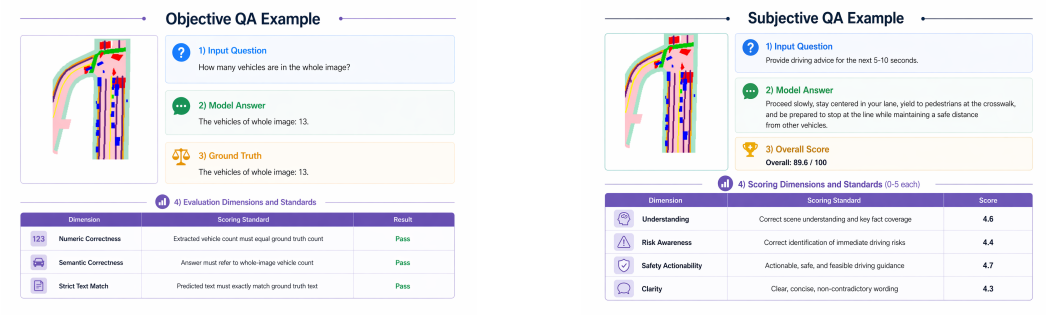
Figure 3 summarizes the ablation results on objective and subjective QA. For objective QA, CLIP-Only performs poorly on element recognition and counting, indicating that standard visual tokens cannot reliably preserve the symbolic semantics encoded in BEV maps. Semantic-Only is substantially stronger, confirming that explicit semantic priors are important for factual grounding. The full SD-LLaVA model achieves the best overall performance across all four objective metrics, showing that visual tokens and semantic prior tokens are complementary rather than redundant. For subjective QA, the advantage of the dual-stream design is also clear. Semantic-Only and CLIP-Only remain close to each other, while SD-LLaVA consistently achieves higher scores across all dimensions. This suggests that open-ended driving reasoning requires both structured semantic grounding and visual-language generation ability.

4.2 EFFECT OF THE BACKBONE ON THE DUAL-STREAM ARCHITECTURE



(a) Objective QA across backbones. (b) Subjective QA across backbones.

Figure 4: Comparison of the SD-LLaVA dual-stream architecture across different LLaVA backbones. The dual-stream design yields strong objective factual grounding and subjective driving-oriented reasoning across all backbones, with the strongest backbone (LLaVA-v1.6-13B) giving the best performance.



(a) Objective QA example.

(b) Subjective QA example.

Figure 5: Qualitative examples of SD-LLaVA on objective and subjective BEV QA. The objective example evaluates constrained factual grounding through vehicle counting, while the subjective example evaluates open-ended driving-oriented reasoning using rubric-based dimensions.

Figure 4 compares the SD-LLaVA dual-stream architecture across three LLaVA backbones. All backbones achieve reasonable performance, and the strongest backbone (LLaVA-v1.6-13B) performs best on both objective and subjective QA. This indicates that the benefit of the dual-stream design is stable across backbones rather than being tied to a specific base model, while a stronger backbone further improves the results. The key factor is the semantic dual-stream representation, which converts BEV semantic structure into learnable prior tokens and fuses them with visual tokens before language decoding.

4.3 Evidence for the BEV QA Formulation

The results also provide preliminary evidence for the usefulness of the proposed BEV QA formulation. Objective QA separates factual reasoning abilities such as element recognition, yes/no judgment, and counting, while subjective QA evaluates road-condition description and short-horizon driving advice. These tasks distinguish CLIP-Only, Semantic-Only, the full SD-LLaVA model, and different backbones, indicating that the formulation has enough discriminative power to separate models of different capability. We note, however, that a systematic validation of the formulation itself, such as zero-shot and no-image baselines to rule out language shortcuts, human verification of the automatically generated labels, and cross-dataset transfer, is left to future work; the present evidence for the formulation is therefore indirect, resting on its ability to discriminate between models rather than on a dedicated dataset study.

4.4 Qualitative Analysis

To further illustrate the behavior of SD-LLaVA under different QA regimes, Figure 5 presents representative objective and subjective BEV QA examples. The objective example asks for the number of vehicles in the whole BEV image. SD-LLaVA produces a parseable numerical answer that matches the ground-truth count, and the evaluation checks numeric correctness, semantic correctness, and strict text matching. This example shows that the semantic prior stream helps preserve explicit BEV facts for constrained factual reasoning.

The subjective example asks for driving advice for the next 5–10 seconds. Unlike objective QA, this task does not have a single strict answer. The model needs to interpret nearby vehicles, pedestrian crossings, stop lines, and lane structure, then generate actionable guidance. The rubric-based evaluation assesses understanding, risk awareness, safety actionability, and clarity.

These examples demonstrate that the proposed task-aware BEV QA formulation supports both constrained factual grounding and open-ended driving-oriented reasoning. This qualitative evidence is consistent with the quantitative results, where the full dual-stream model outperforms single-stream variants across objective and subjective settings.

5 LIMITATIONS

The current method relies on a fixed BEV color palette, so robustness under palette shift, noisy rendering, or different BEV conventions requires further study. The evidence for the proposed BEV QA formulation is currently indirect, resting on its ability to discriminate between models of different capability rather than on a dedicated dataset study; systematic validation, including zero-shot and no-image baselines to rule out language shortcuts, human verification of the automatically generated objective labels, and cross-dataset transfer, is left to future work. In addition, objective QA labels are automatically generated from semantic masks, and subjective QA is evaluated by a GPT-based judge against reference answers rather than by human raters, so the subjective scores should be read as deterministic rubric-based proxy metrics. Finally, the reported results are obtained on a single evaluation split without multi-seed runs or error bars, and future work may also study whether an explicit learned router can further improve controllability.

6 CONCLUSION

We presented SD-LLaVA, a semantic dual-stream framework for BEV-grounded vision-language reasoning. The key idea is to complement conventional CLIP visual tokens with explicit region-aware semantic prior tokens, so that the language model can reason over both continuous visual context and structured BEV evidence. This design addresses the mismatch between natural-image visual encoders and BEV semantic maps, where scene information is encoded through symbolic color labels, spatial layouts, and driving-specific geometric conventions. Experiments on BEVCar-derived BEVSchramm et al. (2024) question answering show that explicit semantic priors are crucial for objective factual grounding, while the full dual-stream model is necessary for stronger open-ended driving-oriented reasoning, and that these gains are stable across different LLaVA backbones. These results indicate that structured BEV priors provide an effective and interpretable interface for adapting LVLMs to autonomous-driving scene representations. Future work will study robustness under different BEV palettes and rendering conventions, cross-dataset generalization, and human-validated evaluation for subjective driving-oriented responses. A systematic validation of the proposed BEV QA formulation, including zero-shot and no-image baselines to rule out language shortcuts and human verification of the automatically generated labels, is another important direction.

REFERENCES

Xiaozhi Chen, Huimin Ma, Ji Wan, Bo Li, and Tian Xia. Multi-view 3d object detection network for autonomous driving. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 1907–1915, 2017.

Tushar Choudhary, Vikrant Dewangan, Shivam Chandhok, Shubham Priyadarshan, Anushka Jain, Arun K Singh, Siddharth Srivastava, Krishna Murthy Jatavallabhula, and K Madhava Krishna. Talk2bev: Language-enhanced bird's-eye view maps for autonomous driving. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 16345–16352. IEEE, 2024.

Xinpeng Ding, Jianhua Han, Hang Xu, Xiaodan Liang, Wei Zhang, and Xiaomeng Li. Holistic autonomous driving understanding by bird's-eye-view injected multi-modal large models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13668–13677, 2024.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, et al. A survey on llm-as-a-judge. *The Innovation*, 2024.

Ayesha Ishaq, Jean Lahoud, Ketan More, Omkar Thawakar, Ritesh Thawkar, Dinura Dissanayake, Noor Ahsan, Yuhao Li, Fahad Shahbaz Khan, Hisham Cholakkal, et al. Drivelmm-o1: A step-by-step reasoning dataset and large multimodal model for driving scenario understanding. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 20501–20508. IEEE, 2025.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024a.

Hongyang Li, Chonghao Sima, Jifeng Dai, Wenhai Wang, Lewei Lu, Huijie Wang, Jia Zeng, Zhiqi Li, Jiazhi Yang, Hanming Deng, et al. Delving into the devils of bird's-eye-view perception: A review, evaluation and recipe. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(4):2151–2170, 2023.

Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Qiao Yu, and Jifeng Dai. Bevformer: learning bird's-eye-view representation from lidar-camera via spatiotemporal transformers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(3):2020–2036, 2024b.

Zongxia Li, Xiyang Wu, Hongyang Du, Fuxiao Liu, Huy Nghiem, and Guangyao Shi. A survey of state of the art large vision language models: Benchmark evaluations and challenges. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 1587–1606, 2025.

Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. A survey on hallucination in large vision-language models. *arXiv preprint arXiv:2402.00253*, 2024a.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023a.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 26296–26306, 2024b.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: Nlg evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 2511–2522, 2023b.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pp. 216–233. Springer, 2024c.

Zhijian Liu, Haotian Tang, Alexander Amini, Xinyu Yang, Huizi Mao, Daniela L Rus, and Song Han. Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. In *2023 IEEE international conference on robotics and automation (ICRA)*, pp. 2774–2781. IEEE, 2023c.

Jiageng Mao, Yuxi Qian, Junjie Ye, Hang Zhao, and Yue Wang. Gpt-driver: Learning to drive with gpt. *arXiv preprint arXiv:2310.01415*, 2023.

Karthik Mohan, Sonam Singh, and Amit Arvind Kale. Bella: End-to-end birds eye view large language assistant for autonomous driving. *arXiv preprint arXiv:2512.06096*, 2025.

Jonah Philion and Sanja Fidler. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *European conference on computer vision*, pp. 194–210. Springer, 2020.

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

Jonas Schramm, Niclas Vödisch, Kürsat Petek, B Ravi Kiran, Senthil Yogamani, Wolfram Burgard, and Abhinav Valada. Bevcar: Camera-radar fusion for bev map and object segmentation. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 1435–1442. IEEE, 2024.

Hao Shao, Yuxuan Hu, Letian Wang, Guanglu Song, Steven L Waslander, Yu Liu, and Hongsheng Li. Lmdrive: Closed-loop end-to-end driving with large language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15120–15130, 2024.

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.

Kexin Tian, Jingrui Mao, Yunlong Zhang, Jiwan Jiang, Yang Zhou, and Zhengzhong Tu. Nuscenes-spatialqa: A spatial understanding and reasoning benchmark for vision-language models in autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4567–4576, 2025.

Qingyao Xu, Siheng Chen, Guang Chen, Yanfeng Wang, and Ya Zhang. Chatbev: A visual language model that understands bev maps. *arXiv preprint arXiv:2503.13938*, 2025.

Zhenhua Xu, Yujia Zhang, Enze Xie, Zhen Zhao, Yong Guo, Kwan-Yee K Wong, Zhenguo Li, and Hengshuang Zhao. Drivegpt4: Interpretable end-to-end autonomous driving via large language model. *IEEE Robotics and Automation Letters*, 9(10):8186–8193, 2024.

Bin Yang, Wenjie Luo, and Raquel Urtasun. Pixor: Real-time 3d object detection from point clouds. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 7652–7660, 2018.