

Less is More: A Simple yet Effective Object-Centric Prompting Strategy for Vision-Language Reasoning in Autonomous Driving

Saiqian Peng[✉], Liping Lu[✉], Bing Shi[✉], and Duanfeng Chu^(✉)[✉]

Wuhan University of Technology, Wuhan, China
{psq,luliping,bingshi,chudf}@whut.edu.cn

Abstract. Large Vision-Language Models (VLMs) show immense potential in autonomous driving but struggle with precise spatial grounding, often hallucinating when forced to regress continuous image coordinates. To overcome this, we introduce a "Less is More" object-centric prompting strategy that fundamentally decouples spatial localization from logical reasoning. By overlaying detected bounding boxes and discrete IDs directly onto multi-view images, the VLM only predicts discrete IDs, which are deterministically mapped back to precise coordinates via an offline lookup table. This model-agnostic approach transforms error-prone regression into reliable discrete matching without requiring structural modifications. Extensive evaluations on the DriveLM-nuScenes dataset demonstrate striking improvements in comprehensive driving reasoning and spatial grounding. Notably, our method boosts the Accuracy and Match scores of the DriveLM-agent by absolute margins of 8.26% and 10.39%, respectively. Furthermore, when applied to stronger models, it achieves impressive increases of 11.85% and 20.02% in Accuracy and Match for Qwen2.5-VL, as well as 8.75% and 19.88% for InternVL-3.5. These consistent gains definitively prove the strong plug-and-play generalizability of our paradigm across varying architectures.

Keywords: Autonomous Driving · Large Vision-Language Models · Visual Question Answering · Visual Prompting

1 Introduction

In recent years, the integration of Large Vision-Language Models has sparked a paradigm shift in the field of autonomous driving. While substantial progress has been achieved in core autonomous driving tasks [12, 45], recent advances in large models have enabled a more unified framework for complex scene understanding, holistic logical reasoning, and human-computer interaction [10, 44]. By taking single or multi-view driving scene images and a textual prompt as inputs, VLMs can perform an array of sophisticated vision-language reasoning

[✉] Corresponding author.

tasks, spanning from foundational perception and future trajectory prediction, to high-level planning and ego-vehicle behavior reasoning [1, 7, 39, 40]. This end-to-end reasoning paradigm holds great promise for building more interpretable and intelligent autonomous systems.

Despite their unprecedented success in high-level semantic reasoning, existing VLMs struggle significantly with the precise spatial grounding required for safety-critical driving applications [44]. A prominent limitation arises when these models are tasked with outputting the spatial coordinates, such as bounding boxes, of specific traffic participants. Standard VLM architectures are fundamentally optimized for discrete token generation and are therefore inherently ill-suited for coordinate regression [6, 15, 32, 50]. Consequently, when forced to predict continuous numerical coordinates, these models frequently suffer from severe hallucinations and catastrophic localization failures. Forcing a language model to act as a spatial regressor essentially conflates the distinct cognitive tasks of logical reasoning and precise localization, ultimately leading to suboptimal performance in both domains.

To overcome this bottleneck, we advocate for a decoupling of object localization and logical reasoning. In this paper, we propose a "Less is More" philosophy and introduce a simple yet highly effective Object-Centric Prompting Strategy (OCPS). As illustrated in Fig. 1, rather than modifying the complex internal architecture of the VLM to accommodate coordinate regression, we reformulate both the input and output data representations to align precisely with the model's natural strengths in discrete classification. Specifically, we first employ an off-the-shelf object detector to extract accurate bounding boxes of key traffic elements. We then overlay these filtered bounding boxes directly onto the raw input images and assign a unique, discrete ID to each object.

During both the fine-tuning and inference stages, the VLM is entirely relieved of the burden of coordinate prediction. Instead, it only needs to predict the discrete ID associated with the target object to answer the prompt. Finally, an offline lookup table seamlessly maps the VLM's predicted ID back to the precise image coordinates obtained from the initial detection step. By transforming a challenging continuous regression problem into a straightforward discrete ID matching task, our method effectively eradicates localization hallucination without requiring any structural modifications to the underlying VLM. This strategy significantly boosts the accuracy of spatial grounding across diverse scene reasoning tasks, proving that in the realm of vision-language reasoning for autonomous driving, a simpler, decoupled approach can yield far superior results.

In summary, our key contributions are listed as follows:

- **A Decoupled Reasoning Paradigm.** We propose a "Less is More" object-centric prompting strategy that fundamentally decouples spatial localization from logical reasoning. By transforming error-prone continuous coordinate regression into reliable discrete ID prediction, we effectively mitigate VLM spatial hallucinations.
- **A Simple, Model-Agnostic Pipeline.** We introduce a lightweight approach that overlays detected bounding boxes and unique IDs directly onto

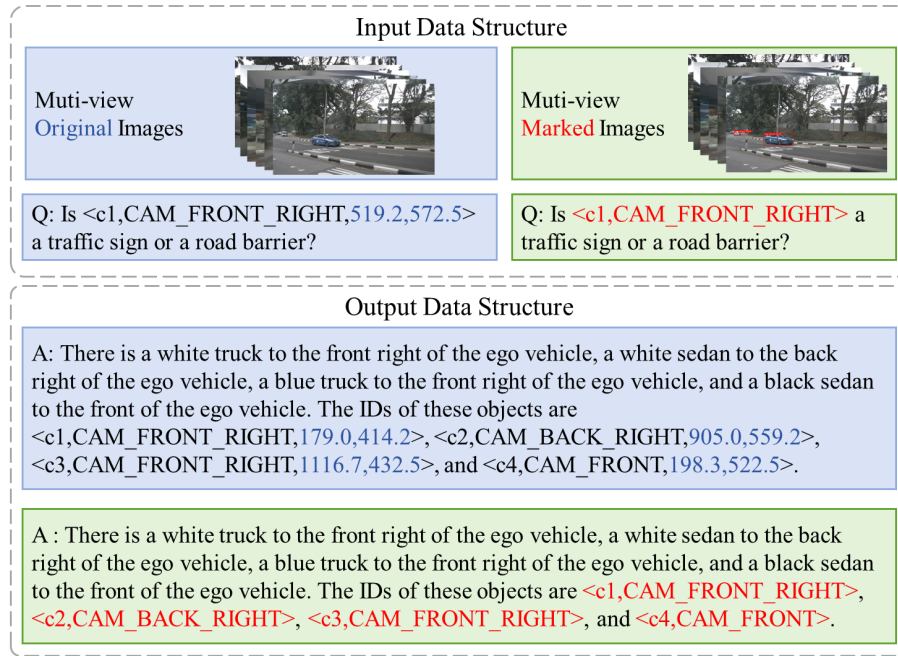


Fig. 1: Comparison of data formats between the traditional paradigm and our Object-Centric Prompting Strategy (OCPS). The traditional paradigm (blue box) takes original images as visual inputs and provides both object IDs and specific coordinates in the textual prompt. Consequently, it forces the model to regress continuous coordinates in the output text. In contrast, our OCPS (green box) utilizes marked images. Its input prompt relies exclusively on discrete IDs, and the resulting output text contains no spatial coordinates, effectively relieving the model from the burden of coordinate regression.

input images. This strategy drastically improves spatial grounding entirely through data representation, requiring absolutely zero structural modifications to the underlying VLMs.

- Extensive Cross-Model Generalizability:** We rigorously evaluate our strategy across a diverse suite of VLMs with varying architectures, including DriveLM-agent [32] (based on LLaMA [37]), Qwen2.5-VL [1], and InternVL-3.5 [40]. Comprehensive comparisons demonstrate that fine-tuning with our object-centric prompting strategy yields striking performance improvements over standard baselines, firmly establishing the robust plug-and-play capability and broad generalizability of our approach.

2 Related Works

2.1 Autonomous Driving Visual Question Answering

Visual Question Answering (VQA) has emerged as a crucial task for evaluating the cognitive and reasoning capabilities of autonomous driving systems, moving beyond the scope of traditional independent perception modules [10]. Built upon foundational multi-modal datasets like nuScenes [2] and Waymo [33], a surge of recent benchmarks has profoundly advanced this field. Specifically, various datasets and frameworks have been proposed to evaluate comprehensive multi-modal understanding, graph-based connectivity, and fine-grained spatial reasoning [17, 19, 28, 32, 35]. Concurrently, further efforts have pushed the boundaries toward step-by-step reasoning, joint importance ranking, and embodied scene understanding [20, 26, 31, 41]. To better ground these language models in physical environments, methods like Talk2BEV [9] introduce language-enhanced Bird’s-eye View representations. Despite these rapid advancements, accurately localizing and referencing specific traffic participants remains a primary bottleneck, as high-level decision-making heavily depends on precise fine-grained spatial grounding [44].

2.2 Spatial Understanding with MLLMs

Following the rapid adaptation of Multimodal Large Language Models (MLLMs) for autonomous driving visual question answering and empirical evaluation [15, 36, 44], capturing complex semantics across multiple views and multi-level features has become essential [6, 50]. To support these domain-specific tasks, general-purpose MLLMs have been explicitly endowed with geometric reasoning [13], broad visual task transferability [21], and fine-grained spatial-temporal comprehension [4, 22]. To further refine precise localization, Qwen2-VL [39] enhances the model’s perception of the visual world at arbitrary resolutions. Moreover, models such as Shikra [5], LLaVA-Grounding [51], and Ferret-v2 [52] have significantly advanced grounded visual chat by explicitly aligning textual tokens with regional coordinates. SpatialRGPT [8] leverages 3D scene graphs and depth information to strengthen the spatial perception and reasoning capabilities of vision-language models. However, representing dense and dynamic driving environments entirely through text-based coordinate regression remains inefficient and error-prone. This limitation naturally motivates the exploration of alternative interaction paradigms, such as visual prompting, to bypass coordinate representations and more intuitively guide the model’s attention to specific on-road objects.

2.3 Visual Prompting

Visual Prompting has emerged as an effective paradigm to fundamentally enhance the image understanding and fine-grained perception of MLLMs [43]. By overlaying explicit visual markers, including masks, points, and bounding boxes onto input images, these techniques guide model attention to specific localized

regions. This approach not only improves spatial grounding but also circumvents the inherent limitations of pure textual coordinate regression [3, 34, 46–48]. Building upon this, subsequent research has achieved precise region and pixel-level comprehension through visual instruction tuning [30, 49, 53] and explicit geometric rendering [24], effectively mitigating object hallucinations during complex reasoning [25, 42]. In the autonomous driving domain, MPDrive [54] utilizes general marker-based prompts for spatial understanding. Diverging from general markers, our work introduces an object-centric representation. By explicitly drawing bounding boxes and assigning unique textual IDs onto the image, we avoid continuous coordinate regression, transforming complex spatial grounding into a highly reliable ID-matching and reasoning process.

3 Method

3.1 Problem Definition

Multimodal Large Language Models have demonstrated remarkable generalization capabilities across various domains. In the context of autonomous driving, these models leverage multi-modal information processing mechanisms to significantly enhance the ego vehicle’s understanding of its surrounding environment.

Formally, we formulate the task as a conditional text generation problem. Given a suite of multi-view images $\mathcal{I} = \{I_1, I_2, \dots, I_N\}$ captured by N surround cameras alongside a driving-related textual query, the objective is to generate a coherent and accurate linguistic response. This process is governed by a parameterized VLM, which synergizes multi-view visual perception with textual instructions. The inference process is formulated in Eq. (1):

$$A = \mathcal{F}_\theta(\mathcal{E}(\mathcal{I}), Q) \quad (1)$$

where each $I_i \in \mathbb{R}^{H \times W \times 3}$ represents the raw visual input from the corresponding camera perspective. The function $\mathcal{E}(\cdot)$ denotes the visual encoder, which projects the multi-view image set $\mathcal{I}$ into a sequence of high-dimensional feature embeddings. Q indicates the tokenized textual query concerning the current driving scenario. Finally, $\mathcal{F}_\theta$ represents the multi-modal large language model parameterized by θ , which reasons over the aligned multi-view visual tokens and textual embeddings to synthesize the final driving response or decision A .

3.2 Overall Architecture

The overall architecture of our Object-Centric Prompting Strategy is depicted in Fig. 2. The architecture comprises three fundamental modules: object-centric prompt generation (Sec. 3.3), visual language model fine-tuning (Sec. 3.4), and coordinate conversion (Sec. 3.5).

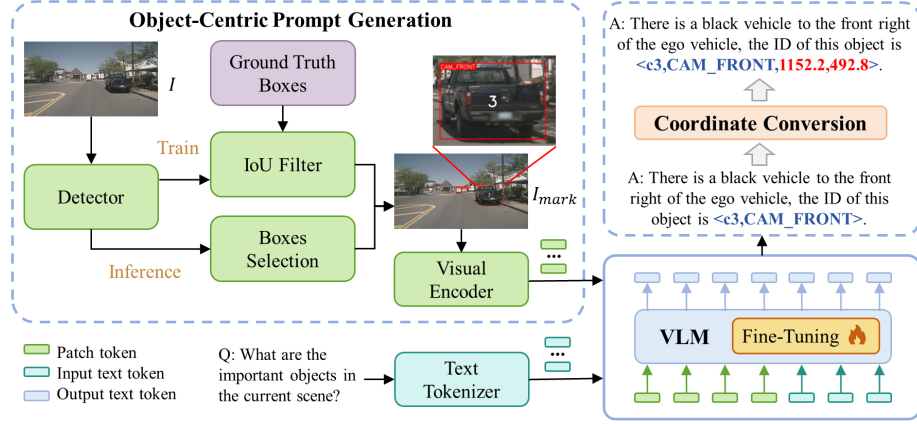


Fig. 2: Overview of the Object-Centric Prompting Strategy. For clarity, the process is illustrated using a single image. During the Object-Centric Prompt Generation stage, the input image I is processed by a detector to generate candidate 2D bounding boxes. We apply IoU matching between detected boxes and ground truths to filter out redundant objects during training, while strictly selecting the highest-confidence objects during inference. The marked image I_{mark} is then generated by drawing the selected bounding boxes on the original image, with the camera perspective marked in the upper-left corner and a unique ID centered within each box. Ultimately, I_{mark} and the query text are processed by the VLM to generate the answer, followed by a coordinate conversion step that maps the predicted IDs back to their specific spatial coordinates.

3.3 Object-Centric Prompt Generation

Given an input image $I \in \mathbb{R}^{H \times W \times 3}$, our goal is to enhance the Vision-Language Model’s spatial grounding capability by embedding explicit visual cues.

Object Proposal Generation. Initially, the input image I is processed by an object detector to extract a set of candidate 2D bounding boxes $B_{det} = \{b_1, b_2, \dots, b_M\}$. This step provides critical object-centric priors, significantly reducing the spatial search space for the VLM.

Filtering Strategy. Raw detections often contain redundant bounding boxes that can introduce noise. To construct high-quality inputs, we apply distinct filtering strategies during training and inference. During training, we introduce a Ground-Truth guided filtering mechanism using Intersection over Union (IoU) matching. Let B_{gt} denote the set of ground-truth boxes. We retain a subset of valid boxes $\hat{B}_{det}$, defined by Eq. (2):

$$\hat{B}_{det} = \{b_i \mid \max_{b_g \in B_{gt}} \text{IoU}(b_i, b_g) > \tau\} \quad (2)$$

where τ is a predefined IoU threshold utilized to filter out redundant objects. Conversely, during inference, we strictly select the objects with the highest detection confidence scores to form the retained subset $\hat{B}_{det}$.

Visual Marker Superimposition. For each retained bounding box $b_j \in \hat{B}_{det}$, we assign a unique numerical identifier ID_j . We then modify the original image I to generate an annotated image, denoted as I_{mark} . Specifically, we draw the selected 2D bounding boxes directly onto the image space, annotating the corresponding camera perspective in the upper-left corner of each box and embedding the unique ID_j precisely at its center.

3.4 Visual Language Model Fine-Tuning

To demonstrate the generalizability of our Object-Centric Prompting Strategy, we integrate and fine-tune our approach across three representative Vision-Language Models: DriveLM-agent [32] (LLaMA-7B [37]), Qwen2.5-VL-7B [1], and InternVL-3.5-4B [40].

Feature Encoding. We utilize the pre-trained vision encoders inherent to each baseline architecture: CLIP ViT-L/14 [11, 29] for the DriveLM-agent, the standard ViT for Qwen2.5-VL, and InternViT [7] for InternVL-3.5. The extracted multi-view visual features are concatenated and mapped into the LLM embedding space via a learnable projection layer. Concurrently, the textual prompts incorporating our discrete object IDs are processed using each model’s standard tokenizer.

Parameter-Efficient Fine-Tuning. To align the visual and textual representations, we employ architecture-specific fine-tuning techniques: LLaMA-Adapter v2 [14] is utilized for the DriveLM-agent, while LoRA [18] is applied to both Qwen2.5-VL and InternVL-3.5.

Optimization. All models are optimized using a standard auto-regressive cross-entropy loss, applied exclusively to the ground-truth answer tokens (ignoring padding and visual prompt indices), as shown in Eq. (3):

$$\mathcal{L}_{CE} = -\frac{1}{T} \sum_{t=1}^T \log P_{\theta}(x_t | x_{<t}, \mathbf{v}) \quad (3)$$

where T is the target sequence length and P_{θ} represents the conditional probability predicted by the model at step t . By formulating the objective to strictly predict discrete IDs instead of regressing continuous spatial coordinates, this optimization process seamlessly aligns with the inherent strengths of large language models.

3.5 Inference and Offline Coordinate Recovery.

To recover precise spatial localization from the generated textual response, we perform a deterministic ID-to-coordinate mapping during the post-processing stage. During the initial object detection phase, we maintain a cached coordinate dictionary $\mathcal{D}$ that stores the spatial metadata of all detected objects. Specifically, each object is indexed by a composite key comprising its unique ID and its

associated camera view, mapping directly to its 2D center coordinates (x, y) , as shown in Eq. (4):

$$\mathcal{D}(id, v) \mapsto (x, y) \quad (4)$$

When the multimodal large language model outputs a token sequence containing an object reference in the format $\langle id, v \rangle$ (e.g., $\langle c3, CAM_FRONT \rangle$), we parse this textual tag and query the cached dictionary $\mathcal{D}$. This direct lookup operation seamlessly transforms the discrete ID representation back into a continuous spatial format $\langle id, v, x, y \rangle$. This deterministic mapping strategy explicitly bypasses the need for the language model to regress precise numerical coordinates, thereby ensuring highly stable and accurate spatial localization in the final output.

4 Experiments

In this section, we systematically evaluate our object-centric prompting strategy. We first detail the experimental setup, followed by a comprehensive analysis of quantitative and qualitative results against baselines, and conclude with extensive ablation studies.

4.1 Experimental Setup

Dataset. We evaluate the effectiveness of our proposed method on the DriveLM-nuScenes dataset [32]. Since the ground truth annotations for the official test set are not publicly available and the evaluation server is no longer maintained, we follow the protocol established by [15] to repartition the official training data. Specifically, we split the data into a new training set and test set at a 9:1 ratio. This yields 26,458 QA pairs for training and 2,992 QA pairs for testing. Each QA sample comprises a question, a corresponding answer, and six multi-view images.

Evaluation Metrics. To comprehensively evaluate the model’s performance on the autonomous driving visual question answering task, we employ a multi-dimensional suite of evaluation metrics. Specifically, these metrics are designed to rigorously assess the model’s capabilities across four critical dimensions: complex scene understanding, high-level logical reasoning, precise spatial grounding, and natural language generation.

Accuracy Score. We employ the Accuracy score to assess the model’s proficiency in answering four-way multiple-choice and binary yes/no questions. A response is considered correct only if it achieves an exact string match with the ground truth annotation. This strict evaluation criterion accurately quantifies the model’s precise decision-making capabilities.

Chat Score. To assess the semantic equivalence between the generated answers and the ground truth, we introduce Llama 3 [16] as an evaluator. Unlike naive string matching, this metric can identify answers that are semantically

correct despite different wordings, thereby providing a more comprehensive evaluation of the model’s comprehension.

Match Score. For perception-oriented queries requiring precise object localization, we utilize the Match metric. To account for variations in the quantity and ordering of generated objects, this metric relies on a distance-based optimal matching strategy based on L_1 distances. A true positive is recorded only if the distance between the predicted coordinate and its nearest ground truth is less than 16 pixels. The metric then computes the F1 score to harmonize precision and recall, ultimately formulating the final Match score as a weighted average of the F1 score and the Chat semantic score.

Language Score. We evaluate the linguistic quality of the generated responses using standard natural language processing metrics. Specifically, we utilize BLEU-1/2/3/4 [27] to measure n-gram precision, ROUGE-L [23] to assess structural recall via the longest common subsequence, and CIDEr [38] to calculate the overall semantic consensus.

Final Score. To obtain a holistic evaluation of the model’s performance, we calculate a final overall score via a weighted sum of the four aforementioned metrics. The calculation is formulated in Eq. (5):

$$S_{\text{final}} = w_a \cdot S_{\text{Accuracy}} + w_c \cdot S_{\text{Chat}} + w_m \cdot S_{\text{Match}} + w_l \cdot S_{\text{Language}} \quad (5)$$

where w_a , w_c , w_m , and w_l represent the weights for the Accuracy score, Chat score, Match score and Language score, respectively. Following the standard evaluation protocol, these weights are predefined as $w_c = 0.4$ and $w_a = w_m = w_l = 0.2$.

Implementation Details. Our experiments involve fine-tuning DriveLM-agent, Qwen2.5-VL, and InternVL-3.5. We limit the maximum text sequence to 512 tokens and standardize the input visual resolution to 224×224 for all models. For the DriveLM-agent, an additional 336×336 resolution variant is evaluated to analyze visual granularity. Multi-view features from the six cameras are concatenated sequentially. All models are optimized using AdamW (learning rate: 1×10^{-4} , weight decay: 0.02) on two NVIDIA GeForce RTX 3090 GPUs. The training process spans 4 epochs for the DriveLM-agent and 3 epochs for both Qwen2.5-VL and InternVL-3.5. Due to memory limitations, we set the per-device batch size to 1 and utilize gradient accumulation to achieve a larger effective batch size.

4.2 Quantitative Results

We systematically evaluate the effectiveness of our proposed Object-Centric Prompting Strategy through extensive quantitative comparisons on the DriveLM-nuScenes dataset. Tab. 1 summarizes the performance of three representative large vision-language models: DriveLM-agent (based on LLaMA-7B), Qwen2.5-VL, and InternVL-3.5. By comparing each baseline against its counterpart equipped with our method, we demonstrate that our approach consistently

yields substantial improvements across different architectures, thereby confirming its strong plug-and-play generalizability.

Table 1: Performance evaluation on the DriveLM-nuScenes dataset. Quantitative comparison of different large vision-language models with and without our proposed method. The selected base models are built upon LLaMA-7B (DriveLM-agent), Qwen2.5-VL-7B-Instruct (Qwen2.5-VL), and InternVL3.5-4B (InternVL3.5), respectively. The best results within each baseline group are highlighted in **bold**.

Method	Accuracy	Chat	Match	Language						Final
				BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	CIDEr	
DriveLM-agent [32]	0.6024	72.71	32.29	0.6993	0.6340	0.5706	0.5089	0.6861	0.0352	0.5621
DriveLM-agent+Ours	0.6850	72.63	42.68	0.7105	0.6460	0.5838	0.5228	0.7213	0.0501	0.6023
Qwen2.5-VL [1]	0.7009	69.73	34.54	0.7555	0.6916	0.6315	0.5731	0.7197	0.1549	0.5814
Qwen2.5-VL+Ours	0.8194	69.46	54.56	0.7958	0.7380	0.6824	0.6270	0.7778	0.2910	0.6520
InternVL-3.5 [40]	0.7388	71.65	37.88	0.7449	0.6834	0.6253	0.5688	0.7217	0.1598	0.6030
InternVL-3.5+Ours	0.8263	71.19	57.76	0.8257	0.7715	0.7186	0.6650	0.8117	0.4568	0.6724

Improvements in Logical Reasoning. The primary goal of autonomous driving VLMs is to make sound decisions based on complex scenarios. As shown in Tab. 1, our method significantly enhances the logical reasoning capabilities of the base models, reflected by the remarkable surge in the **Accuracy** metric. Specifically, our method boosts the Accuracy of the baseline DriveLM-agent from 0.6024 to 0.6850 (+8.26%). When applied to stronger architectures, the gains are even more pronounced, achieving absolute increases of 11.85% and 8.75% for Qwen2.5-VL and InternVL-3.5, respectively. This consistent improvement validates our core motivation: by decoupling spatial localization from the text generation process, the VLM is no longer burdened by the difficult task of continuous coordinate regression. Consequently, the model can allocate its full representational capacity to high-level logical reasoning and decision-making.

Breakthroughs in Spatial Grounding. Precise spatial grounding has historically been a bottleneck for VLMs. Our discrete ID-based formulation fundamentally mitigates the hallucination issues associated with coordinate regression, leading to striking improvements in the **Match** metric. The integration of our method improves the Match score of the baseline DriveLM-agent by an absolute margin of 10.39%. More impressively, for Qwen2.5-VL and InternVL-3.5, the Match scores leap by 20.02% (from 34.54 to 54.56) and 19.88% (from 37.88 to 57.76), respectively. These breakthroughs demonstrate that mapping predicted discrete IDs to precise coordinates via an offline lookup table is a far more reliable spatial grounding paradigm for LLM-based architectures.

Language Quality and Overall Performance. In addition to reasoning and grounding, our method notably improves the linguistic quality of the generated responses. Across all three base models, standard language metrics (BLEU-1 to BLEU-4 and ROUGE-L) exhibit consistent gains. Most notably, the CIDEr score, which heavily penalizes generic responses and rewards highly

descriptive captions, experiences a massive surge. For instance, the CIDEr score of InternVL-3.5 jumps from 0.1598 to 0.4568. While we observe a negligible drop in the Chat metric, this minor fluctuation represents a common and acceptable trade-off when enforcing stricter, object-centric output formats. Ultimately, the comprehensive Final score, which aggregates all aspects of performance, confirms the absolute superiority of our proposed method across all tested baselines.

4.3 Qualitative Results

To provide a more intuitive comparison of the different approaches, we present qualitative visualizations in Fig. 3. Given the six multi-view images and the query prompt, "What object should the ego vehicle notice first when the ego vehicle is getting to the next possible location?", the baseline DriveLM-agent responds: "Firstly, notice that $\langle c3, CAM_FRONT, 105.5, 505.0 \rangle$. The object is stationary, so the ego vehicle should continue ahead at the same speed." In contrast, our method generates: "Firstly, notice $\langle c3, CAM_FRONT, 582.6, 541.4 \rangle$. The object is going ahead, so the ego vehicle should continue ahead at the same speed." It is evident that our approach yields an accurate and reliable response. While the baseline DriveLM-agent produces a logically plausible reasoning structure, its visual grounding coordinates deviate completely from the ground truth, exhibiting a clear hallucination phenomenon.

Fig. 4 provides a qualitative multi-choice VQA example for ego-vehicle behavior prediction. Given six multi-view images, models must comprehensively understand the environment to deduce the vehicle’s driving status. As shown, although DriveLM-agent, Qwen2.5-VL, and InternVL-3.5 possess strong general visual reasoning capabilities, they output erroneous predictions when facing dense spatial driving scenes. Conversely, when augmented with our object-centric method, all three models are consistently corrected to the ground truth. This demonstrates that our approach significantly enhances the models’ ability to capture critical spatial cues for accurate behavioral reasoning across different architectures.

4.4 Ablation Studies

To validate the effectiveness of the proposed modules in our architecture, we conduct comprehensive ablation experiments on the DriveLM-nuScenes dataset. As shown in Tab. 2, we establish the vanilla DriveLM-agent as our baseline and incrementally integrate Object-Centric Prompt Generation (OCPG) and Coordinate Conversion (CC) to analyze their individual contributions.

Effect of OCPG. The OCPG module shifts the model from continuous coordinate regression to discrete object-ID reasoning. As shown in Tab. 2, this significantly boosts Accuracy ($0.6024 \rightarrow 0.6850$) and Chat score ($72.71 \rightarrow 75.96$), demonstrating that abstracting spatial details allows the model to better focus on semantic understanding and logical deduction. However, relying solely on OCPG decreases the Match score and language metrics. This decline occurs

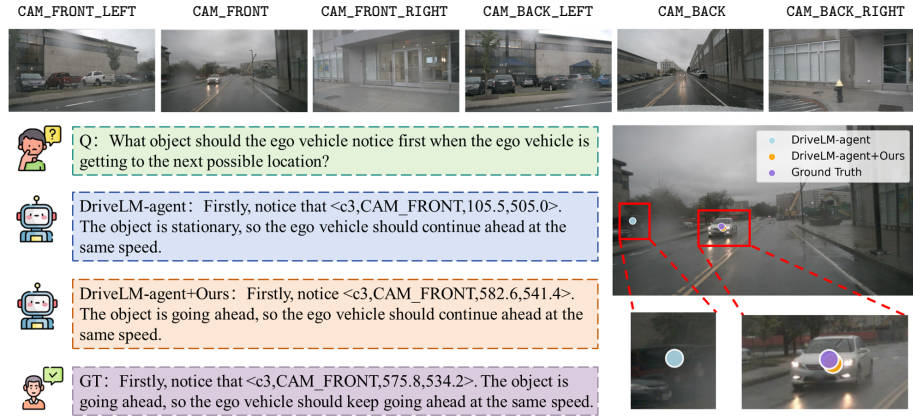


Fig. 3: Comparison of the responses from the DriveLM-agent with and without the Object-Centric Prompting Strategy. The blue, orange, and purple dots indicate the objects focused on by the DriveLM-agent, our method, and the ground truth, respectively.

because OCPG-only outputs generate discrete IDs without specific coordinates, leading to a structural mismatch with the ground truth.

Effect of CC. To resolve the structural mismatch caused by OCPG-only outputs, we integrate the CC module. Our complete architecture maintains the high Accuracy (0.6850) while substantially improving the Match score (28.27 $\rightarrow$ 42.68). Furthermore, it restores and exceeds the baseline’s language metrics. This demonstrates that mapping discrete IDs back to explicit spatial coordinates successfully bridges semantic reasoning and physical localization, culminating in the best Final score (0.6023).

Table 2: Ablation study of proposed components on the DriveLM-nuScenes dataset. OCPG and CC denote Object-Centric Prompt Generation and Coordinate Conversion, respectively. The best results are highlighted in **bold**.

Method	Component		Language									
	OCPG	CC	Accuracy	Chat	Match	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	CIDEr	Final
DriveLM-agent [32]			0.6024	72.71	32.29	0.6993	0.6340	0.5706	0.5089	0.6861	0.0352	0.5621
DriveLM-agent+OCPG	✓		0.6850	75.96	28.27	0.6619	0.5866	0.5238	0.4706	0.6871	0.0224	0.5807
Ours	✓	✓	0.6850	72.63	42.68	0.7105	0.6460	0.5838	0.5228	0.7213	0.0501	0.6023

Tab. 3 further ablates the impact of visual encoder resolutions and IoU thresholds. The higher-resolution CLIP ViT-L/14₃₃₆ consistently outperforms the standard CLIP ViT-L/14 in grounding tasks, at an IoU of 0.3, it increases the Match score to 46.12 and the Final score to 0.6094, confirming that enhanced visual features improve spatial awareness. Furthermore, varying the IoU thresholds reveals that relaxing spatial constraints naturally boosts the Match

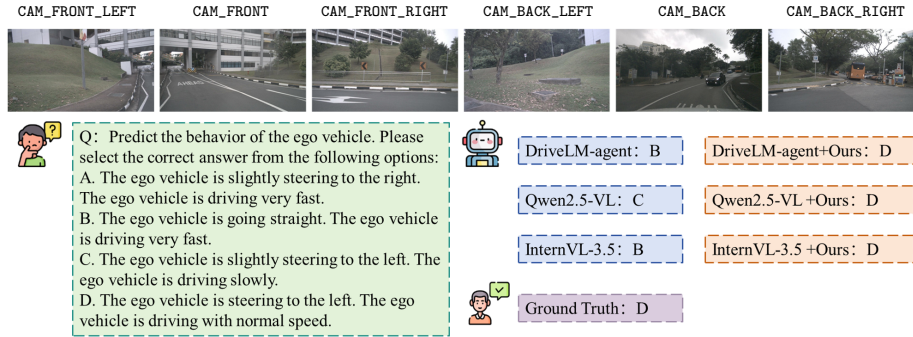


Fig. 4: Effectiveness of our proposed method across different MLLM base-lines. In this scenario, baseline models including DriveLM-agent, Qwen2.5-VL, and InternVL-3.5 fail to accurately predict the ego vehicle’s behavior. By integrating our method, all models are successfully guided to yield the correct prediction.

Table 3: Ablation study on visual encoder resolutions and IoU thresholds. All experiments are based on our complete architecture built upon the DriveLM-agent baseline. We compare the standard CLIP ViT-L/14 with the higher-resolution CLIP ViT-L/14₃₃₆ across varying IoU matching thresholds. The Ground Truth (GT) setting is included to establish the theoretical performance upper bound with perfect spatial localization.

Method	Visual Encoder	IoU	Accuracy	Chat	Match	Language					Final	
						BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L		CIDEr
Ours-0.7	CLIP ViT-L/14 GT	0.7	0.6720	73.78	37.72	0.7067	0.6412	0.5777	0.5155	0.6972	0.0535	0.5925
Ours-0.5		0.5	0.6922	72.80	40.28	0.7129	0.6480	0.5855	0.5242	0.7213	0.0490	0.5998
Ours-0.3		0.3	0.6850	72.63	42.68	0.7105	0.6460	0.5838	0.5228	0.7213	0.0501	0.6023
GT		-	0.7447	69.96	53.05	0.6890	0.6339	0.5812	0.5321	0.7042	0.8194	0.6279
Ours-0.7	CLIP ViT-L/14 ₃₃₆ GT	0.7	0.6713	73.69	39.22	0.7397	0.6707	0.6037	0.5377	0.7181	0.0505	0.5982
Ours-0.5		0.5	0.6850	72.46	42.91	0.7286	0.6601	0.5942	0.5296	0.7068	0.0502	0.6020
Ours-0.3		0.3	0.6900	71.99	46.12	0.7369	0.6689	0.6035	0.5396	0.7256	0.0497	0.6094
GT		-	0.7637	70.82	61.35	0.7280	0.6685	0.6101	0.5557	0.7263	0.8029	0.6552

score but introduces visual noise that slightly degrades Chat score. However, the higher-resolution CLIP ViT-L/14₃₃₆ exhibits stronger robustness against this noise, sustaining Accuracy improvements even under looser matching criteria.

Finally, the Ground Truth setting establishes a theoretical upper bound by providing perfect object localizations. For CLIP ViT-L/14₃₃₆, the GT inputs substantially boost Accuracy (0.7637), the Match score (61.35), and CIDEr (0.8029). This significant margin highlights precise preliminary object detection as a critical bottleneck for further enhancing the model’s comprehensive reasoning and description capabilities.

While the Ground Truth experiments indicate that localization quality is a major bottleneck of the proposed framework, we further validate this observation through a perception failure analysis. Specifically, we perform inference

without providing object detection bounding boxes during evaluation. As shown in Tab. 4, removing bounding box information causes a notable performance decline across all metrics, with Accuracy decreasing from 0.6850 to 0.5745, Match dropping from 42.68 to 27.36, and the Final score falling from 0.6023 to 0.5578. These findings further confirm that reliable object localization plays a crucial role in our framework, as explicit spatial cues are necessary for accurate visual grounding and subsequent reasoning.

Table 4: Perception failure analysis via removing bounding box information. Performance comparison of our framework with and without object detection bounding boxes during evaluation.

Method	Bounding Box	Accuracy	Match	Final
DriveLM-agent	-	0.6024	32.29	0.5621
DriveLM-agent+Ours	w	0.6850	42.68	0.6023
DriveLM-agent+Ours	w/o	0.5745	27.36	0.5578

5 Conclusion

In this paper, we propose a "Less is More" object-centric prompting strategy to tackle the logical reasoning and spatial grounding challenges of Vision-Language Models in autonomous driving. By overlaying discrete IDs directly onto multi-view images and utilizing offline lookup tables, we fundamentally decouple high-level logical reasoning from low-level spatial localization, effectively transforming error-prone coordinate regression into reliable discrete matching. Our extensive evaluations on the DriveLM-nuScenes dataset demonstrate the strong plug-and-play capability of this model-agnostic approach without requiring any structural modifications. Specifically, our method boosts the Accuracy and Match scores of the DriveLM-agent baseline by absolute margins of 8.26% and 10.39%, respectively. Furthermore, these consistent gains extend to stronger architectures, achieving impressive increases of 11.85% and 20.02% for Qwen2.5-VL, as well as 8.75% and 19.88% for InternVL-3.5. While we have currently validated the effectiveness of our approach on a single dataset, in future work, we plan to conduct extensive experiments across multiple diverse autonomous driving benchmarks to further verify the broad generalizability and robustness of our paradigm.

Acknowledgements

This work is supported in part by the National Natural Science Foundation of China (52472438), the Hubei Provincial Key R&D Program of (2024BAB033), the Wuhan Municipal key R&D Program (20250512030403), and the Science and Technology Project of the Department of Transport of Hubei Province (2025-69-3-5).

References

1. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2.5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
2. Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nusenes: A multimodal dataset for autonomous driving. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 11618–11628 (2020). <https://doi.org/10.1109/CVPR42600.2020.01164>
3. Cai, M., Liu, H., Mustikovela, S.K., Meyer, G.P., Chai, Y., Park, D., Lee, Y.J.: Vip-llava: Making large multimodal models understand arbitrary visual prompts. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 12914–12923 (2024). <https://doi.org/10.1109/CVPR52733.2024.01227>
4. Chen, B., Xu, Z., Kirmani, S., Ichter, B., Sadigh, D., Guibas, L., Xia, F.: Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 14455–14465 (2024). <https://doi.org/10.1109/CVPR52733.2024.01370>
5. Chen, K., Zhang, Z., Zeng, W., Zhang, R., Zhu, F., Zhao, R.: Shikra: Unleashing multimodal llm’s referential dialogue magic. arXiv preprint arXiv:2306.15195 (2023)
6. Chen, L., Hassani, H., Nikan, S.: Ts-vlm: Text-guided softsort pooling for vision-language models in multi-view driving reasoning. arXiv preprint arXiv:2505.12670 (2025)
7. Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., Li, B., Luo, P., Lu, T., Qiao, Y., Dai, J.: Intern vl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 24185–24198 (2024). <https://doi.org/10.1109/CVPR52733.2024.02283>
8. Cheng, A.C., Yin, H., Fu, Y., Guo, Q., Yang, R., Kautz, J., Wang, X., Liu, S.: Spatialrgpt: Grounded spatial reasoning in vision-language models. *Advances in Neural Information Processing Systems* **37**, 135062–135093 (2024)
9. Choudhary, T., Dewangan, V., Chandhok, S., Priyadarshan, S., Jain, A., Singh, A.K., Srivastava, S., Jatavallabhula, K.M., Krishna, K.M.: Talk2bev: Language-enhanced bird’s-eye view maps for autonomous driving. In: 2024 IEEE International Conference on Robotics and Automation (ICRA). pp. 16345–16352 (2024). <https://doi.org/10.1109/ICRA57147.2024.10611485>
10. Cui, C., Ma, Y., Cao, X., Ye, W., Zhou, Y., Liang, K., Chen, J., Lu, J., Yang, Z., Liao, K.D., et al.: A survey on multimodal large language models for autonomous driving. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. pp. 958–979 (2024)
11. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: *Int. Conf. Learn. Represent.* (2021)
12. Feng, R., Xi, N., Chu, D., Wang, R., Deng, Z., Wang, A., Lu, L., Wang, J., Huang, Y.: Artemis: Autoregressive end-to-end trajectory planning with mixture of experts for autonomous driving. *IEEE Robotics and Automation Letters* **11**(1), 226–233 (2025)
13. Gao, J., Pi, R., Zhang, J., Ye, J., Zhong, W., Wang, Y., HONG, L., Han, J., Xu, H., Li, Z., Kong, L.: G-LLaVA: Solving geometric problem with multi-modal large language model. In: *Int. Conf. Learn. Represent.* (2025)

14. Gao, P., Han, J., Zhang, R., Lin, Z., Geng, S., Zhou, A., Zhang, W., Lu, P., He, C., Yue, X., Li, H., Qiao, Y.: Llama-adapter v2: Parameter-efficient visual instruction model. arXiv preprint arXiv:2304.15010 (2023)
15. Gopalkrishnan, A., Greer, R., Trivedi, M.: Multi-frame, lightweight & efficient vision-language models for question answering in autonomous driving. In: First Vision and Language for Autonomous Driving and Robotics Workshop (2024)
16. Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., et al.: The llama 3 herd of models. arXiv preprint arXiv:2407.21783 (2024)
17. Guo, X., Zhang, R., Duan, Y., He, Y., Nie, D., Huang, W., Zhang, C., Liu, S., Zhao, H., Chen, L.: SURDS: Benchmarking spatial understanding and reasoning in driving scenarios with vision language models. In: Adv. Neural Inform. Process. Syst. (2025)
18. Hu, E.J., yelong shen, Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. In: Int. Conf. Learn. Represent. (2022)
19. Inoue, Y., Yada, Y., Tanahashi, K., Yamaguchi, Y.: Nuscenes-mqa: Integrated evaluation of captions and qa for autonomous driving datasets using markup annotations. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops. pp. 930–938 (January 2024). <https://doi.org/10.1109/WACVW60836.2024.00104>
20. Ishaq, A., Lahoud, J., More, K., Thawakar, O., Thawkar, R., Dissanayake, D., Ahsan, N., Li, Y., Khan, F.S., Cholakkal, H., 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). <https://doi.org/10.1109/IROS60139.2025.11247360>
21. Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Zhang, P., Li, Y., Liu, Z., Li, C.: LLaVA-onevision: Easy visual task transfer. Trans. Mach. Learn. Res. (2025)
22. Li, H., Chen, J., Wei, Z., Huang, S., Hui, T., Gao, J., Wei, X., Liu, S.: Llava-st: A multimodal large language model for fine-grained spatial-temporal understanding. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 8592–8603 (2025). <https://doi.org/10.1109/CVPR52734.2025.00803>
23. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Text summarization branches out. pp. 74–81 (2004)
24. Lin, W., Wei, X., An, R., Gao, P., Zou, B., Luo, Y., Huang, S., Zhang, S., Li, H.: Draw-and-understand: Leveraging visual prompts to enable MLLMs to comprehend what you want. In: Int. Conf. Learn. Represent. (2025)
25. Lu, Y., Cao, J., Wu, Y., Li, B., Tang, L., Ji, Y., Wu, C., Wu, J., Zhu, W.: Rsvp: Reasoning segmentation via visual prompting and multi-modal chain-of-thought. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 14699–14716 (2025). <https://doi.org/10.18653/v1/2025.acl-long.715>
26. Nie, M., Peng, R., Wang, C., Cai, X., Han, J., Xu, H., Zhang, L.: Reason2drive: Towards interpretable and chain-based reasoning for autonomous driving. In: Eur. Conf. Comput. Vis. pp. 292–308. Springer (2024). https://doi.org/10.1007/978-3-031-73347-5_17
27. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of

- the Association for Computational Linguistics. pp. 311–318 (2002). <https://doi.org/10.3115/1073083.1073135>
28. Qian, T., Chen, J., Zhuo, L., Jiao, Y., Jiang, Y.G.: Nuscenes-qa: A multi-modal visual question answering benchmark for autonomous driving scenario. In: AAAI. vol. 38, pp. 4542–4550 (2024). <https://doi.org/10.1609/aaai.v38i5.28253>
 29. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: Int. Conf. Mach. Learn. pp. 8748–8763. PMLR (2021)
 30. Rasheed, H., Maaz, M., Shaji, S., Shaker, A., Khan, S., Cholakkal, H., Anwer, R.M., Xing, E., Yang, M.H., Khan, F.S.: Glamm: Pixel grounding large multi-modal model. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 13009–13018 (2024). <https://doi.org/10.1109/CVPR52733.2024.01236>
 31. Sachdeva, E., Agarwal, N., Chundi, S., Roelofs, S., Li, J., Kochenderfer, M., Choi, C., Dariush, B.: Rank2tell: A multimodal driving dataset for joint importance ranking and reasoning. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 7513–7522 (2024). <https://doi.org/10.1109/WACV57701.2024.00734>
 32. Sima, C., Renz, K., Chitta, K., Chen, L., Zhang, H., Xie, C., Beißwenger, J., Luo, P., Geiger, A., Li, H.: Drivelm: Driving with graph visual question answering. In: Eur. Conf. Comput. Vis. pp. 256–274. Springer (2024). https://doi.org/10.1007/978-3-031-72943-0_15
 33. Sun, P., Kretschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., Guo, J., Zhou, Y., Chai, Y., Caine, B., Vasudevan, V., Han, W., Ngiam, J., Zhao, H., Timofeev, A., Ettinger, S., Krivokon, M., Gao, A., Joshi, A., Zhang, Y., Shlens, J., Chen, Z., Anguelov, D.: Scalability in perception for autonomous driving: Waymo open dataset. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 2443–2451 (2020). <https://doi.org/10.1109/CVPR42600.2020.00252>
 34. Tang, W., Sun, Y., Gu, Q., Li, Z.: Visual position prompt for mllm based visual grounding. IEEE Trans. Multimedia pp. 1–16 (2026). <https://doi.org/10.48550/arXiv.2503.15426>
 35. Tian, K., Mao, J., Zhang, Y., Jiang, J., Zhou, Y., Tu, Z.: Nuscenes-spatialqa: A spatial understanding and reasoning benchmark for vision-language models in autonomous driving. In: Int. Conf. Comput. Vis. pp. 4567–4576 (2025)
 36. Tian, X., Gu, J., Li, B., Liu, Y., Wang, Y., Zhao, Z., Zhan, K., Jia, P., Lang, X., Zhao, H.: DriveVLM: The convergence of autonomous driving and large vision-language models. In: 8th Annual Conference on Robot Learning (2024)
 37. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., Lample, G.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023)
 38. Vedantam, R., Lawrence Zitnick, C., Parikh, D.: Cider: Consensus-based image description evaluation. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 4566–4575 (2015). <https://doi.org/10.1109/CVPR.2015.7299087>
 39. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al.: Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. arXiv preprint arXiv:2409.12191 (2024)
 40. Wang, W., Gao, Z., Gu, L., Pu, H., Cui, L., Wei, X., Liu, Z., Jing, L., Ye, S., Shao, J., et al.: Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. arXiv preprint arXiv:2508.18265 (2025)

41. Wang, W., Duan, C., Peng, Z., Liu, Y., Zhou, B.: Embodied scene understanding for vision language models via metavqa. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 22453–22464 (2025). <https://doi.org/10.1109/CVPR52734.2025.02091>
42. Woo, S., Zhou, K., Zhou, Y., Wang, S., Guan, S., Ding, H., Cheong, L.L.: Black-box visual prompt engineering for mitigating object hallucination in large vision language models. In: Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers). pp. 529–538 (2025). <https://doi.org/10.18653/v1/2025.naacl-short.45>
43. Wu, J., Zhang, Z., Xia, Y., Li, X., Xia, Z., Chang, A., Yu, T., Kim, S., Rossi, R.A., Zhang, R., et al.: Visual prompting in multimodal large language models: A survey. arXiv preprint arXiv:2409.15310 (2024)
44. Xie, S., Kong, L., Dong, Y., Sima, C., Zhang, W., Chen, Q.A., Liu, Z., Pan, L.: Are vlms ready for autonomous driving? an empirical study from the reliability, data and metric perspectives. In: Int. Conf. Comput. Vis. pp. 6585–6597 (2025)
45. Xin, G., Chu, D., Lu, L., Deng, Z., Lu, Y., Wu, X.: Multi-agent trajectory prediction with difficulty-guided feature enhancement network. IEEE Robotics and Automation Letters **10**(3), 2582–2589 (2025)
46. Yang, J., Zhang, H., Li, F., Zou, X., Li, C., Gao, J.: Set-of-mark prompting unleashes extraordinary visual grounding in gpt-4v. arXiv preprint arXiv:2310.11441 (2023)
47. Yang, L., Li, X., Wang, Y., Wang, X., Yang, J.: Fine-grained visual text prompting. IEEE Trans. Pattern Anal. Mach. Intell. **47**(3), 1594–1609 (2025). <https://doi.org/10.1109/TPAMI.2024.3504568>
48. Yu, R., Yu, W., Wang, X.: Attention prompting on image for large vision-language models. In: Eur. Conf. Comput. Vis. pp. 251–268. Springer (2024). https://doi.org/10.1007/978-3-031-73404-5_15
49. Yuan, Y., Li, W., Liu, J., Tang, D., Luo, X., Qin, C., Zhang, L., Zhu, J.: Osprey: Pixel understanding with visual instruction tuning. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 28202–28211 (2024). <https://doi.org/10.1109/CVPR52733.2024.02664>
50. Zhang, E., Dai, X., Huang, M., Lv, Y., Miao, Q.: Minidrive: More efficient vision-language models with multi-level 2d features as text tokens for autonomous driving. In: Chinese Conference on Pattern Recognition and Computer Vision (PRCV). pp. 255–269. Springer (2025). https://doi.org/10.1007/978-981-95-5740-0_18
51. Zhang, H., Li, H., Li, F., Ren, T., Zou, X., Liu, S., Huang, S., Gao, J., Leizhang, Li, C., et al.: Llava-grounding: Grounded visual chat with large multimodal models. In: Eur. Conf. Comput. Vis. pp. 19–35. Springer (2024). https://doi.org/10.1007/978-3-031-72775-7_2
52. Zhang, H., You, H., Dufter, P., Zhang, B., Chen, C., Chen, H.Y., Fu, T.J., Wang, W.Y., Chang, S.F., Gan, Z., Yang, Y.: Ferret-v2: An improved baseline for referring and grounding with large language models. In: First Conference on Language Modeling (2024)
53. Zhang, S., Sun, P., Chen, S., Xiao, M., Shao, W., Zhang, W., Liu, Y., Chen, K., Luo, P.: Gpt4roi: Instruction tuning large language model on region-of-interest. In: Eur. Conf. Comput. Vis. pp. 52–70. Springer (2025)
54. Zhang, Z., Li, X., Xu, Z., Peng, W., Zhou, Z., Shi, M., Huang, S.: Mpdriive: Improving spatial understanding with marker-based prompt learning for autonomous driving. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 12089–12099 (2025). <https://doi.org/10.1109/CVPR52734.2025.01129>