This ICCV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Hints of Prompt: Enhancing Visual Representation for Multimodal LLMs in Autonomous Driving

Hao Zhou[1,2*], Zhanning Gao[2], Zhili Chen[3], Maosheng Ye[2], Qifeng Chen[3], Tongyi Cao[2], Honggang Qi[1†]
[1]University of Chinese Academy of Sciences, Beijing, China
[2]DeepRoute.AI, Shenzhen, China
[3]The Hong Kong University of Science and Technology, Hong Kong, China
zhouhao233@mails.ucas.ac.cn, {zhanninggao, maoshengye, tongyicao}@deeproute.ai,
zchenei@connect.ust.hk, cqf@cse.ust.hk, hgqi@ucas.ac.cn

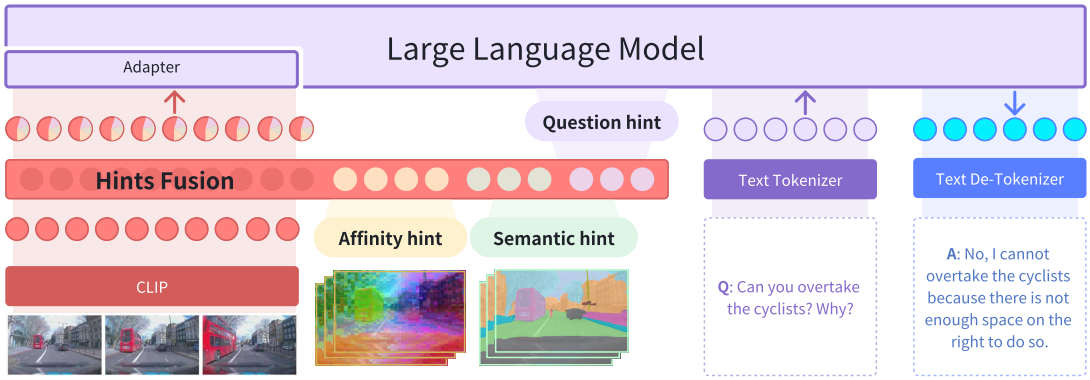


Figure 1. **Overview of the Hints of Prompt (HoP) Framework.** HoP enhances Visual Question Answering (VQA) in autonomous driving by incorporating three hierarchical hints: Affinity, Semantic, and Question. The Affinity hint provides foundational instance-level structures through token-wise connections, aiding in instance boundary and interaction recognition. Building on this, the Semantic hint introduces specific instances along with their category information, adding essential driving-related contexts, such as vehicles and traffic signs. Finally, the Question hint guides the LLM's attention toward image regions pertinent to the question. These hints are fused with visual tokens through a simple Hint Fusion module, aligned via an adapter, and then processed by the LLM to generate answers.

Abstract

In light of the dynamic nature of autonomous driving environments and stringent safety requirements, general MLLMs combined with CLIP alone often struggle to accurately represent driving-specific scenarios, particularly in complex interactions and long-tail cases. To address this, we propose the Hints of Prompt (HoP) framework, which introduces three key enhancements: Affinity hint to emphasize instance-level structure by strengthening token-wise connections, Semantic hint to incorporate high-level information relevant to driving-specific cases, such as complex interactions among vehicles and traffic signs, and Question hint to align visual features with the query context, focusing on question-relevant regions. These hints are fused through a Hint Fusion module, enriching visual representations by capturing driving-related representations with limited domain data, ensuring faster adaptation to driving scenarios. Extensive experiments confirm the effectiveness of the HoP framework, showing that it significantly outperforms previous state-of-the-art methods in all key metrics.

* Work conducted as intern at DeepRoute.AI. † Corresponding author.

1. Introduction

In recent years, there have been growing trends towards formalizing autonomous driving within end-to-end learning-based frameworks [18]. These approaches use extensive human driving data to demonstrate human-like behavior and generalization capabilities, enabling them to handle complex scenarios more effectively. However, the end-to-end pipeline often functions as a *black box*, making it challenging to interpret or explain the planning processes driven by neural networks [7]. This opacity raises significant ethical and legal concerns regarding accountability and trustworthiness in safety-critical applications.

Interpretability has become a crucial research focus in

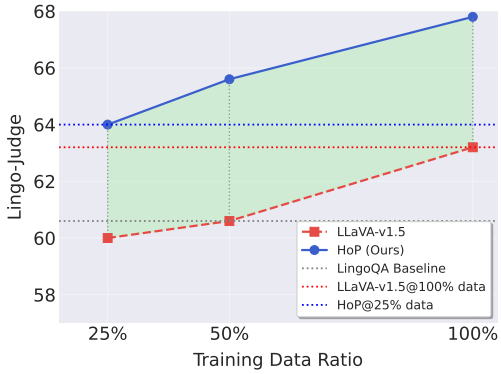


Figure 2. Performance under different training data ratio on the LingoQA dataset. HoP surpasses the full-data performance of LLaVA-v1.5 using only **25%** of the training data.

autonomous driving to address these concerns. Visual Question Answering (VQA) tasks have gained considerable attention as they allow systems to articulate their observations and actions through natural language interactions [41]. The application of Multimodal Large Language Models (MLLMs) [4, 8, 14, 27, 46, 51] has significantly advanced VQA tasks by integrating visual encoders with Large Language Models (LLMs) [10, 12, 48]. Typically, MLLMs comprise a visual encoder, an adapter, and an LLM. For VQA, the visual encoder extracts image features, aligned by the adapter with text tokens. The LLM tokenizes the question and integrates visual tokens to generate an answer.

VQA in autonomous driving requires capturing hierarchical visual features for complex interactions and long-tail scenarios. However, general MLLMs [4, 27, 51] struggle to capture domain-specific representations due to limited visual-language aligned data in the driving domain [38, 43]. Although existing methods [42, 47] attempt to enhance visual representation through multi-feature fusion of visual encoders like CLIP [39], DINOv2 [36], and Pix2Struct [22], their intricate fusion strategies and specialized designs fall short of addressing the unique demands of autonomous driving: *domain-specific adaptability with limited data*.

To enable faster domain adaptation, the visual encoder must quickly capture instance-level information relevant to driving scenarios. However, as shown in Fig. 3, CLIP's visual tokens lack token-wise affinity relationships, which are critical for preserving instance-level structure. Additionally, these general visual encoders underrepresent domain-specific semantics, making them prone to missing small, sparse, yet crucial elements for driving, such as distant vehicles, pedestrians, and traffic signs. Lastly, current MLLMs struggle to adjust visual representations based on specific questions because visual and text tokens are processed separately. This separation limits the model's ability to focus on query-relevant image regions, reducing the effectiveness of context-aware responses in autonomous driving.

Based on these observations, we introduce additional tokens as hints to enhance visual representation. Specifically, we propose three key enhancements: *Affinity hint* to capture instance-level structures by leveraging token-wise affinity relationships from DINOv2 [36]; *Semantic hint* to introduce sparse, driving-specific tokens that incorporate high-level semantic information using queries from DETR-like models; and *Question hint* to correlate visual features according to the query context, ensuring focus on question-relevant regions. We conduct comprehensive studies to design effective and efficient fusion strategies for combining these hint tokens. Fig. 1 illustrates our proposed method, referred to as the Hints of Prompt (HoP) framework. Fig. 2 shows that HoP outperforms the baseline with only **25%** of the training data, demonstrating efficient domain adaptation with limited data in driving scenarios. Furthermore, after analyzing the gains from HoP, we propose an efficient version that directly distills hint tokens from the CLIP backbone, achieving competitive performance more efficiently.

Through extensive experiments on autonomous driving VQA benchmarks, we demonstrate that our method significantly outperforms existing approaches, achieving state-of-the-art results. Our ablation studies further confirm the effectiveness of each hints, as well as the impact of different fusion strategies and LLM scales, illustrating how each component contributes to the overall performance improvement and faster domain adaptation.

To summarize, our contributions are as follows:

- We identify key limitations of current general MLLMs with CLIP alone in autonomous driving: lack of domain priors for domain-specific visual representation.
- We introduce the Hints of Prompt (HoP) framework, a data-efficient solution for enhancing visual representation tailored for autonomous driving scenes.
- Extensive experiments demonstrate significant improvements of our methods, establishing state-of-the-art results on LingoQA [35], DRAMA [33], and BDD-X [21].

2. Related Work

Multimodal Large Language Models. Multimodal Large Language Models (MLLMs) [8, 27, 28, 30, 54, 55] have made significant advancements by integrating Large Language Models (LLMs) [1, 10, 48] with visual encoders, addressing diverse multimodal tasks across various domains.

In autonomous driving, MLLMs address critical challenges in perception, planning, control, question-answering (QA), and generative world models. For example, GPT-Driver [34] reframes motion planning as a language modeling task, enhancing accuracy, while DriveGPT4 [53] combines vision and language input to predict control signals with real-time explanations, improving transparency.

In perception, MLLMs advance sensor data fusion from LiDAR and cameras; models such as Talk2BEV [11] use

Bird's Eye View (BEV) features for complex visual tasks, while LMDrive [40] and DriveVLM [45] employ trajectory-based reasoning to refine scene understanding and decision making. In QA, DriveLM [43] uses graph-based reasoning to enhance real-time planning. For world models, generative approaches like DriveDreamer [52] and GAIA-1 [19] generate realistic driving scenarios for data augmentation and training. These advancements highlight MLLMs' potential to enhance multimodal understanding and optimize autonomous driving systems across diverse applications.

Enhancing Visual Representations in MLLMs. Effective alignment of visual representations with LLM capabilities is essential to improve overall performance. Taking advantage of data, InternVL [8] introduces a large-scale visual encoder to improve granularity of the features, and some works [26, 31, 42, 46, 47] combine multiple visual encoders with a complex fusion strategy to take advantage of complementary strengths. However, effective alignment of visual representations with LLM capabilities remains challenging in autonomous driving.

QA-ViT [17] incorporates textual information into visual features to help MLLMs focus on relevant image regions, enhancing interpretability while maintaining efficiency. Similarly, we utilize the same LLM text embeddings as Question hints, but our fusion strategy is more easily integrated with various visual backbones. While Hint-AD [15] aligns language with perception-planning tokens holistically, and TOKEN [44] leverages enriched representations generated by an end-to-end driving model to mitigate data scarcity, our method uniquely introduces dynamic hint tokens with a lightweight fusion module, enabling context-aware visual-semantic alignment for driving scenarios.

3. Method

3.1. Overview

The Hints of Prompt (HoP) framework (Fig. 1) leverages three distinct token types—Affinity, Semantic, and Question hints—each representing a different level of information. These hints progress from instance-level, token-wise connections in the Affinity hint, to high-level semantic information in the Semantic hint, and finally to question-specific context in the Question hint, gradually refining and aligning the visual representation with the LLM. A simple Hint Fusion module integrates these hints into a cohesive visual representation, which is then processed by an adapter and passed to the LLM for answer generation. By incorporating hints across different information levels, our method enhances context-aware in VQA for autonomous driving and facilitates domain adaptation under data scarcity.

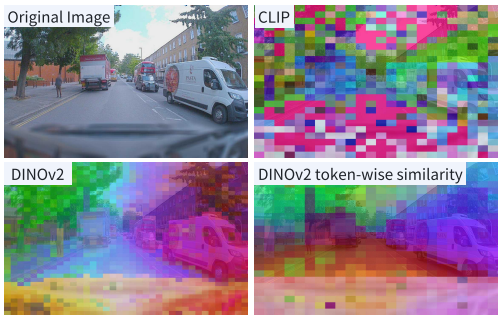


Figure 3. Visualization of token affinity from CLIP [39] and DINOv2 [36]. Similar colors indicate higher affinity scores, with color values derived from a PCA-reduced token vector space. DINOv2 token-wise similarity denotes tokens embedded only from the similarity matrix.

3.2. Hint Tokens

Affinity hint The default practice of MLLMs [27, 28, 31] applies the CLIP [39] encoder for visual feature extraction, where the features are well-aligned with the text embedding space and contain rich semantic knowledge. However, the CLIP visual tokens struggle to understand driving scenes that require spatial reasoning among instances on the road. As shown in Fig. 3, we observe that the CLIP visual tokens drop the affinity information among instances, which might hinder the following LLM from reasoning upon these limited visual tokens. We find that DINOv2 [36] maintains significant affinity among tokens within the same instances, offering a more coherent spatial representation. Therefore, we use DINOv2 tokens as Affinity hint tokens to reinforce the instance-level structure in complex driving scenes.

To verify that the enhancement is due to token-wise affinity rather than specific DINOv2 features, we experiment with similarity information among the DINOv2 visual tokens and observe a notable performance gain. A detailed discussion is provided in Sec. 4.4.

Semantic hint After implementing instance-level enhancement, providing a semantic category specific to the driving domain can further focus visual representations on key instances, such as the cyclist shown in Fig. 4. The Semantic hint aims to incorporate high-level, domain-specific semantic information, such as vehicles, traffic signs, and pedestrians, which is crucial for understanding complex interactions. To achieve this, we employ sparse queries from DETR-like [6] detection or segmentation pipelines as Semantic hints. Specifically, GroundingDINO [29] or Mask2Former [9] will be used for the extraction of Semantic hints. Experimental results (Tab. 5) show that benefiting from Mask2Former's adaptation to driving scenarios, it can align more effectively with the task-specific requirements of VQA in autonomous driving.

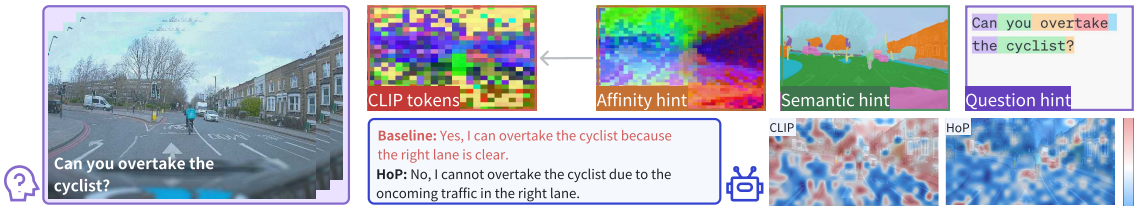


Figure 4. The impact of the three proposed hint types compared to the baseline and attention maps with/without these hints.

Question hint In VQA tasks, different questions often shift attention to specific regions of an image. For example, a question about the weather may focus on broader aspects, such as the sky or overall lighting. In contrast, a question about pedestrians in the path directs attention to finer details and specific areas. Incorporating question-specific information can further refine this focus by enhancing visual representation to capture key instances in the driving environment. Aligning the visual representation with each question's particular demands enables a more targeted, context-aware interpretation, thus enhancing relevance and effectiveness in complex driving scenarios. Similar to QA-ViT [17], we found that using LLM's text tokens as the Question hint effectively achieves context-aware representation. We also explored an alternative approach that utilizes CLIP's text encoder to generate the Question hint. A comparison of these methods is shown in Tab. 6.

As shown in Fig. 4, incorporating the three proposed hints improves the model's response over the baseline. The baseline incorrectly suggests overtaking is possible, overlooking oncoming traffic. As shown by the HoP attention map, the model with hints gains instance-level structure and focuses more accurately on key elements, such as the cyclist and oncoming vehicles. This added context enhances the model's spatial reasoning, resulting in a more precise and relevant response in complex driving scenarios. Detailed quantitative analysis is available in Sec. 4.

3.3. Hints Fusion

We explore several fusion strategies to merge the above hints into the original CLIP [39] visual tokens. Formally, we denote Affinity, Semantic, and Question hints as $\mathbf{H}_\mathrm{A} \in \mathbb{R}^{N\times d}$, $\mathbf{H}_\mathrm{S} \in \mathbb{R}^{M\times d}$, and $\mathbf{H}_\mathrm{Q} \in \mathbb{R}^{K\times d}$, respectively, where N, M, and K are the number of tokens for each hint, and d is the embedding dimension. The original CLIP visual tokens are represented as $\mathbf{P} \in \mathbb{R}^{L\times d}$, and the fused visual tokens are denoted as $\mathbf{P}_\mathrm{f} \in \mathbb{R}^{L\times d}$. The notation $[\cdot]$ indicates the concatenation operation, $\mathtt{CA}(\mathbf{P}, \mathbf{H})$ denotes a Cross-Attention layer with $\mathbf{P}$ as query and $\mathbf{H}$ as key-value, and $\mathtt{SA}(\cdot)$ denotes a Self-Attention layer. As shown in Fig. 5, we present the different fusion strategies as follows.

Concatenation As a baseline, hint tokens are directly concatenated with CLIP tokens, as illustrated in Fig. 5(a):

$$\mathbf{P}_\mathrm{f} = [\mathbf{P}, \mathbf{H}_\mathrm{A}, \mathbf{H}_\mathrm{S}, \mathbf{H}_\mathrm{Q}]. \tag{1}$$

Interestingly, we find that simple concatenation can harm visual representation and degrade performance (see Tab. 7), suggesting that directly injecting multi-level information may confuse the adapter and LLM, hindering relevant information extraction.

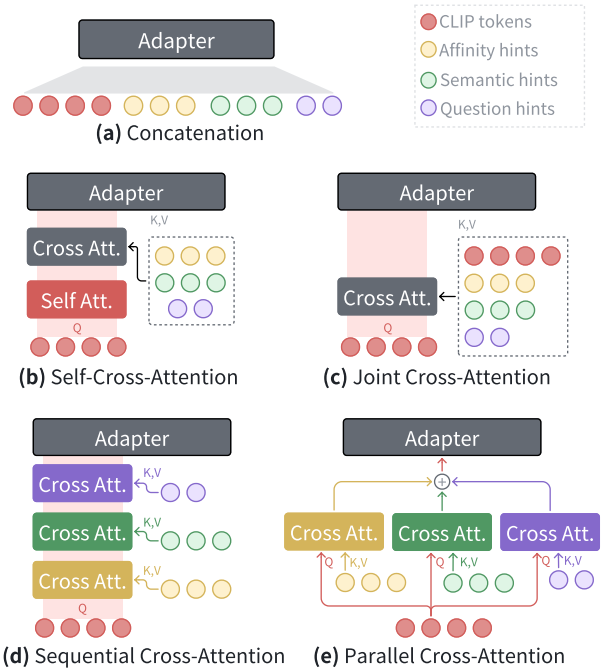


Figure 5. Different fusion strategies for the Hints Fusion module. The joint cross-attention strategy demonstrates the best performance. Residual connections are omitted for simplicity.

Self-Cross-Attention To enable deeper interaction between visual tokens and hint tokens, we propose a Self-Cross-Attention module (Fig. 5(b)) that mimics the standard Transformer [49] layer but omits the feed-forward network (FFN) to accommodate the subsequent adapter module:

$$\mathbf{P}_\mathrm{f} = \mathbf{P} + \mathtt{CA}(\mathtt{SA}(\mathbf{P}), [\mathbf{H}_\mathrm{A}, \mathbf{H}_\mathrm{S}, \mathbf{H}_\mathrm{Q}]). \tag{2}$$

This allows refinement of the visual tokens before incorporating the hint tokens.

Method	LLM	**Lingo-Judge** ↑	BLEU-4 ↑	METEOR ↑	CIDEr ↑
GPT-4V [54] (zero-shot)	-	59.6	6.30	12.4	42.8
LingoQA Baseline [35]	Vicuna-v1.5-7B	60.6	15.0	18.6	65.6
QA-ViT [17]†	Vicuna-v1.5-7B	63.6	14.6	19.0	65.9
VTS [32]	InternLM2-7B	64.2	14.5	**20.5**	56.9
LLaVA-v1.5 [27]†	Vicuna-v1.5-7B	63.2	14.1	19.3	63.7
LLaVA-v1.5 (DINOv2)†	Vicuna-v1.5-7B	62.8	14.7	19.3	<u>68.3</u>
LLaVA-v1.5 (+A-MoF) [47]†	Vicuna-v1.5-7B	64.2	14.5	19.1	64.7
Efficient HoP	Vicuna-v1.5-7B	<u>66.8</u>	<u>15.2</u>	20.0	66.2
HoP	Vicuna-v1.5-7B	**67.8**	**15.8**	<u>20.3</u>	**70.9**

Table 1. **Performance comparison on the LingoQA dataset.** The best-performing method for each metric is highlighted in **bold**, while the second-best method is indicated by an <u>underline</u>. "†" indicates model that are fine-tuned on the LingoQA dataset. *Efficient HoP* represents the lightweight version of our HoP method, while *LLaVA-v1.5* serves as our baseline. *LLaVA-v1.5 (DINOv2)* replaces the CLIP visual encoder with a DINOv2 encoder. The HoP method consistently demonstrates superior performance across all metrics.

Joint Cross-Attention As shown in Fig. 5(c), inspired by Flamingo [2], Joint Cross-Attention omits the self-attention stage and uses the visual representation $\mathbf{P}$ as both query and key-value in a single cross-attention layer:

$$\mathbf{P}_\text{f} = \mathbf{P} + \texttt{CA}(\mathbf{P}, [\mathbf{P}, \mathbf{H}_\text{A}, \mathbf{H}_\text{S}, \mathbf{H}_\text{Q}]). \quad (3)$$

The design simplifies the fusion process without weakening the interaction between visual tokens and hint tokens.

To compare the effects of different query-hint interaction strategies, we further explore two additional approaches: *Parallel Cross-Attention* and *Sequential Cross-Attention*, as shown in Fig. 5(d, e). In these strategies, hint tokens first interact individually with the query and are then fused through either a parallel or cascading way.

Sequential Cross-Attention sequentially applies cross-attention between the visual representation and each hint, progressively enriching the representation:

$$\mathbf{P}_\text{f} = \mathbf{P} + \texttt{CA}(\texttt{CA}(\texttt{CA}(\mathbf{P}, \mathbf{H}_\text{A}), \mathbf{H}_\text{S}), \mathbf{H}_\text{Q}). \quad (4)$$

Parallel Cross-Attention applies cross-attention simultaneously between the visual representation and each hint token, which allows all hint tokens to influence the visual representation in parallel:

$$\mathbf{P}_\text{f} = \mathbf{P} + \texttt{CA}(\mathbf{P}, \mathbf{H}_\text{A}) + \texttt{CA}(\mathbf{P}, \mathbf{H}_\text{S}) + \texttt{CA}(\mathbf{P}, \mathbf{H}_\text{Q}). \quad (5)$$

We empirically find that the proposed Joint Cross-Attention achieves the best performance with high efficiency across diverse VQA tasks in autonomous driving, outperforming the other fusion strategies mentioned above. More details are provided in Sec. 4.4.

3.4. Efficient Hints of Prompt

Although introducing extra hint tokens significantly improves VQA performance, extracting these hints adds a computational burden. To reduce resource overhead, we propose an efficient version of HoP. The main idea is to distill or fine-tune lightweight heads to mimic the behavior of DINOv2 [36] or Mask2Former [9], providing additional instance-level and semantic-level hints to enhance the visual representation while maintaining low latency. More details are provided in Sec. 4.1 and Sec. 4.4.

4. Experiments

We evaluate the Hints of Prompt (HoP) across three datasets: LingoQA [35], DRAMA [33], and BDD-X [21], focusing on its performance in VQA tasks for autonomous driving. Our experiments encompass overall performance improvements, along with ablation studies on the design of each hint, fusion strategies, and efficiency.

4.1. Implementation Details

We implement the Hints of Prompt (HoP) method based on LLaVA-v1.5 [27], freezing the visual encoder and fully fine-tuning the remaining modules.

For the Affinity hint, we use DINOv2-large [36] to extract 576 hint tokens by downsampling the feature map to align with the CLIP tokens (24×24). To verify that the improvement is due to token-wise affinity rather than specific DINOv2 features, we compute a similarity matrix from the 576 DINOv2 tokens and use it as hint tokens, processed through a linear layer. Results are shown in Tab. 4.

For the Semantic hint, we select the top-K queries from Mask2Former [9] or GroundingDINO [29] as hint tokens, based on confidence scores, with each token adding an embedding feature that corresponds to the label index. The Question hint uses the LLM's textual embeddings associated with question tokens.

To create Efficient HoP, we distill and fine-tune two lightweight heads from the CLIP backbone: a 4-layer, 8-heads, 512-dimensional ViT-like [16] decoder for the Affinity hint, distilled via cosine similarity to DINOv2 fea-

Method	BLEU-1↑	BLEU-4↑	METEOR↑	ROUGE↑	CIDEr↑	SPICE↑
LCP [33]	73.9	54.7	39.1	70.0	**3.7**	56.0
VTS [32]	75.3	55.8	40.7	74.7	2.8	58.0
LLaVA-v1.5†	75.8	56.1	41.0	78.0	2.9	58.4
Efficient HoP	76.0	56.2	41.3	78.5	2.7	58.8
HoP	**76.2**	**56.3**	**41.7**	**79.8**	2.8	**59.1**

Table 2. **Performance comparison on the DRAMA dataset.** "†" indicates models that are fine-tuned on the DRAMA dataset. HoP demonstrates enhanced performance over other methods across most metrics, with *LLaVA-v1.5* as the baseline.

Method	Easy			Medium			Hard			All		
	CIDEr↑	B4↑	ROUGE↑	CIDEr↑	B4↑	ROUGE↑	CIDEr↑	B4↑	ROUGE↑	CIDEr↑	B4↑	ROUGE↑
ADAPT [20]	100.93	**20.90**	46.17	62.66	16.44	40.80	52.71	**13.56**	40.49	85.38	17.40	43.04
DriveGPT4 [53]	113.20	20.38	46.46	65.01	16.94	40.51	57.29	12.28	42.07	99.10	**18.32**	44.73
LLaVA-v1.5†	111.84	19.83	47.57	63.76	14.69	40.34	49.80	11.60	41.97	96.51	17.42	45.42
Efficient HoP	118.30	19.95	**47.65**	64.80	17.58	40.73	58.20	12.92	42.58	99.80	17.70	45.65
HoP	**121.01**	20.09	47.60	**65.16**	**18.02**	**40.81**	**60.37**	13.01	**43.52**	**102.20**	17.91	**45.79**

Table 3. **Performance comparison across different splits of the BDD-X testing dataset.** "†" indicates model that are fine-tuned on the BDD-X dataset. The evaluation follows the DriveGPT4 [53] protocol, reporting CIDEr, BLEU-4 (B4), and ROUGE scores across Easy, Medium, Hard, and Overall (All) splits. *LLaVA-v1.5* serves as the baseline.

tures on the LingoQA training dataset; and a ViTDet-like [23] neck with a Mask2Former head for the Semantic hint, trained on the Cityscapes dataset [13]. The training setup follows that of the original Mask2Former.

The whole HoP training is conducted for one epoch using the AdamW optimizer with a cosine learning rate schedule, a warm-up ratio of 0.03, and a base learning rate of 2e-5. All models are trained on 32 NVIDIA A100 GPUs with a batch size of 4 per GPU. We use an NVIDIA RTX 4090 GPU to measure latency in the inference.

4.2. Datasets and Benchmarks

LingoQA [35] is a benchmark dataset for video question answering in autonomous driving, with 28K unique driving scenarios and 419K question-answer pairs. The dataset consists of two subsets: *Action* and *Scenery*, plus a high-quality *Evaluation* subset for benchmarking. The main metric, Lingo-Judge, is an evaluation model aligned with human preferences, while BLEU [37], METEOR [5], and CIDEr [50] provide lexical similarity assessments.

DRAMA [33] dataset comprises 17K video clips of interactive urban traffic scenes in Tokyo, and provides extensive annotations for autonomous driving risk assessment. It includes lots of VQA pairs about 11.8K training samples and 2.5K testing samples. Evaluation metrics include BLEU, METEOR, CIDEr, ROUGE [24], and SPICE [3].

BDD-X [21] is a dataset developed for autonomous driving tasks, focusing on video captioning and rationales, we adopt the VQA pairs and evaluation protocol provided by DriveGPT4 [53], which include 26K training samples and 1.8K testing samples. BLEU, CIDEr, and ROUGE are employed as metrics.

AH types	LJ ↑	B4 ↑	METEOR ↑	CIDEr ↑
Original tokens	**64.6**	14.6	19.0	66.0
Similarity tokens	64.2	**15.5**	**19.3**	**71.1**

Table 4. **Impact of different Affinity hint types on performance.** This ablation examines only the Affinity hint without incorporating other hint types. "Original tokens" refer to DINOv2 tokens, while "Similarity tokens" are derived from the similarity matrix. Abbreviations: AH (Affinity Hint), LJ (Lingo-Judge), B4 (BLEU-4).

4.3. Comparison with Existing Works

Our results on the LingoQA [35], DRAMA [33], and BDD-X [21] datasets demonstrate the effectiveness of the Hints of Prompt (HoP) method in enhancing VQA for autonomous driving. As shown in Tab. 1, HoP significantly outperforms baseline on LingoQA, achieving a top Lingo-Judge score of **67.8** and leading BLEU-4, METEOR, and CIDEr scores.

As shown in Tab. 2, HoP surpasses other methods on the DRAMA dataset across key metrics, including BLEU-1, BLEU-4, METEOR, ROUGE, and SPICE, highlighting its capability to generate detailed and contextually accurate answers in urban traffic scenes.

In Tab. 3, HoP's performance on BDD-X, evaluated using the DriveGPT4 protocol, consistently outperforms other methods across all difficulty levels (Easy, Medium, Hard,

SH types	LJ ↑	B4 ↑	METEOR ↑	CIDEr ↑
GroundingDINO [29]	66.4	15.2	20.1	68.4
Mask2Former [9]	**67.8**	**15.8**	**20.3**	**70.9**

Table 5. **Impact of different Semantic hint types**. Abbreviations: SH (Semantic hint), LJ (Lingo-Judge), B4 (BLEU-4).

QH types	LJ ↑	B4 ↑	METEOR ↑	CIDEr ↑
CLIP	65.2	15.3	20.1	67.0
LLM	**67.8**	**15.8**	**20.3**	**70.9**

Table 6. **Impact of different Question hint types.** "CLIP" refers to hint tokens derived from the CLIP text encoder, "LLM" refers to tokens from the LLM input embedding layer. Abbreviations: QH (Question Hint), LJ (Lingo-Judge), B4 (BLEU-4).

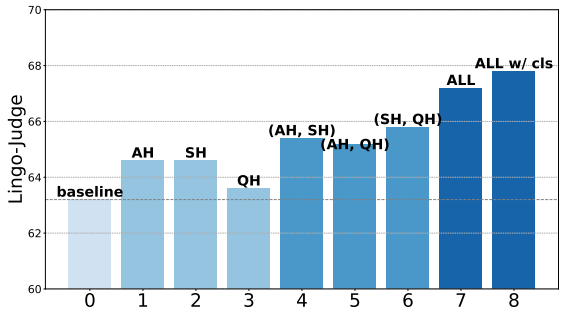


Figure 6. **Impact of each hint type within the HoP.** Abbreviations: AH (Affinity hint), SH (Semantic hint), QH (Question hint). "w/ cls" indicates the addition of class information to the Semantic hint, "ALL" denotes the combination of three hint types.

and All), achieving the highest CIDEr scores and robust results even in challenging driving scenarios.

In addition, the Efficient HoP, a lightweight version of our method, also performs competitively in all datasets, demonstrating the scalability of our approach.

Visualization examples can be found in Fig. 8.

4.4. Ablation Study

To evaluate the contributions of each component within the HoP framework, we conduct comprehensive ablation studies. These include data-efficient domain adaptation of HoP (Fig. 2), analyzing the effectiveness of each hint type independently (Fig. 6), assessing the impact of different hint sources (Tab. 4, Tab. 5, Tab. 6), and hyperparameters (Fig. 7). We also evaluate various fusion strategies for integrating the hints (Tab. 7) and, finally, analyze the efficiency gains of the proposed Efficient HoP (Tab. 8).

Data-efficient domain adaptation of HoP In Fig. 2, HoP trained with only 25% data achieves 64.0 Lingo-Judge score, surpassing the full-data performance of LLaVA-v1.5 (63.2), demonstrates remarkable data efficiency in driving-specific VQA task. More details are in the appendix.

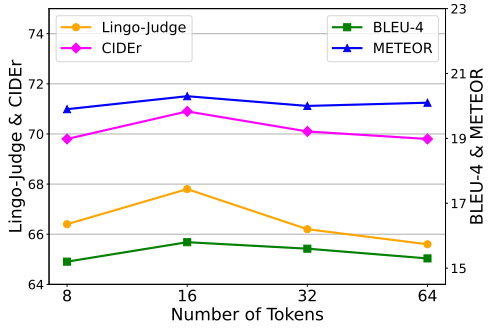


Figure 7. The effect of the number of Semantic hint tokens on performance in the LingoQA [35] dataset.

Combination of Different Hint Types Fig. 6 shows an ablation study on the LingoQA dataset, assessing the impact of each hint type within the HoP framework. Applying each hint independently (IDs 1–3) consistently improves performance over the baseline (ID 0) across all metrics, highlighting each hint's unique contribution to model comprehension. Combining two hints (IDs 4–6) provides additional gains, while using all three hints together (IDs 7–8) achieves the highest scores overall, demonstrating their complementary effects. Notably, adding class information to the Semantic hint (+cls in ID 8) further enhances performance.

Effectiveness of Affinity Hint As shown in Table 4, using DINOv2 [36] token-wise similarity as the source for the Affinity hint provides notable improvements over the baseline, indicating that the core benefit of the Affinity hint lies in capturing token-wise similarity. Although using the full DINOv2 features as the Affinity hint achieves slightly better results, this can be attributed to the additional feature information present in DINOv2. These results confirm the effectiveness of the Affinity hint and underscore the importance of token-wise connections in representing driving scenarios.

Impact of Different Semantic Hint As shown in Tab. 5, Mask2Former [9] achieves superior performance across all metrics, with a Lingo-Judge score of 67.8 compared to 66.4 for GroundingDINO [29]. This boost suggests that Mask2Former's domain adaptation through training on Cityscapes [13] allows it to capture driving-specific semantic details more effectively. In contrast, while GroundingDINO excels in open-set object detection, its lack of domain-specific training limits its alignment with the task-specific requirements of VQA in autonomous driving.

We also analyze the impact of varying the number of Semantic hint tokens on performance in the LingoQA dataset. As shown in Fig. 7, increasing the token count from 8 to 16 improves performance across all metrics. Adding more tokens up to 32 yields similar results, while increasing the count to 64 leads to a decline, likely due to noise introduced by less relevant queries from Mask2Former.

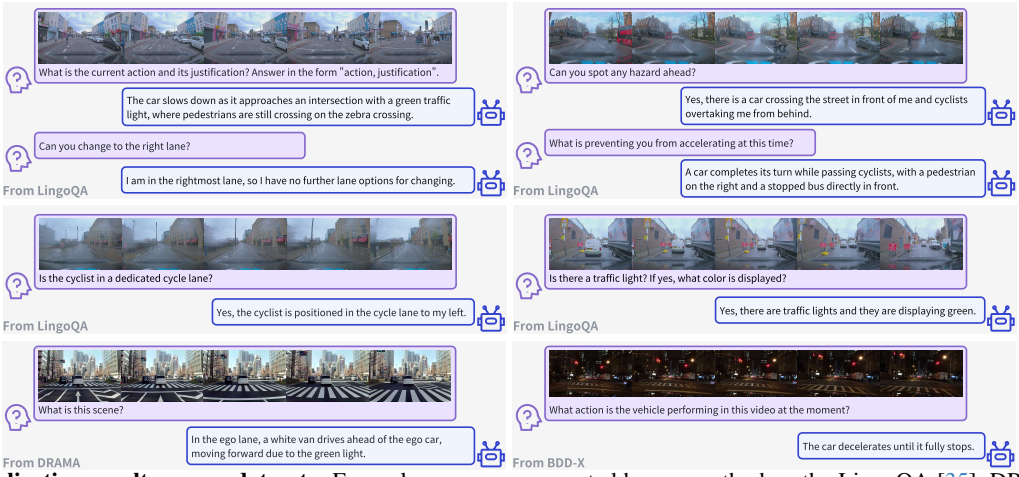


Figure 8. **Visualization results across datasets.** Example responses generated by our method on the LingoQA [35], DRAMA [33], and BDD-X [21] datasets.

Fusion Strategy	LJ ↑	#Params (M)	GFLOPs
Concatenation	58.8	-	-
Self-Cross-Attention	65.2	12.9	5.0
Parallel Cross-Attention	66.6	17.1	8.63
Sequential Cross-Attention	67.0	17.1	8.63
Joint Cross-Attention	**67.8**	**8.7**	**3.8**

Table 7. **Impact of different fusion strategies for integrating hints.** Joint Cross-Attention delivers the best performance with efficient parameter and computational use.

Method	LLM	LJ ↑	#Params (M)	Latency (ms)
VTS	InternLM2-7B	64.2	304.0	740
LLaVA-v1.5	Vicuna-v1.5-7B	63.2	303.5	631
HoP	Vicuna-v1.5-7B	**67.8**	832.0	956
Efficient HoP	Vicuna-v1.5-7B	66.8	385.8	661
	Qwen-v2.5-3B	64.6	385.8	504
	Qwen-v2-0.5B	57.6	385.8	302
with AWQ [25]	Qwen-v2.5-3B	64.2	385.8	**281**

Table 8. **Performance and efficiency of Efficient HoP.** #Params indicate the parameter count for all modules related to *visual representation*. Latency reports the average inference time measured on the LingoQA [35] validation set on an NVIDIA RTX 4090 GPU. AWQ [25] represents the quantized model.

Impact of Different Question Hint Sources We evaluate two sources for extracting question embeddings for the Question hint: CLIP's text model and the LLM embedding layer. As shown in Tab. 6, using the LLM embedding layer yields higher performance across all metrics, achieving a Lingo-Judge score of 67.8 compared to 65.2 with CLIP's text model. This result suggests that the LLM embeddings offer superior alignment for multimodal reasoning tasks.

Hint Fusion Strategy As shown in Tab. 7, the Joint Cross-Attention strategy delivers the best performance, achieving the highest Lingo-Judge (LJ) score of 67.8. This method is not only effective but also highly efficient, requiring only 8.7M parameters and 3.8 GFLOPs—significantly less than the Self-Cross-Attention. In comparison, Sequential and Parallel Cross-Attention strategies, while competitive with LJ scores of 67.0 and 66.6, respectively, require nearly twice the computational resources of Joint Cross-Attention.

Efficient HoP In the original HoP framework, we use DINOv2 [36] and Mask2Former [9] for Affinity and Semantic hints, which boost performance but increase computational costs. To reduce these, we create an Efficient HoP. As shown in Tab. 8, Efficient HoP reduces parameters and latency compared with HoP while achieving comparable performance (Tab. 1, Tab. 2, and Tab. 3). The quantized Efficient HoP with the 3B model maintains this advantage with deployment-ready 281ms latency.

5. Conclusion

In this paper, we introduce the Hints of Prompt (HoP) framework to enhance the visual representation of MLLM in VQA tasks for autonomous driving scenarios. By incorporating Affinity, Semantic, and Question hints, HoP effectively captures instance-level structure, domain-specific elements, and query-specific context, enhancing interpretability and reasoning capabilities, and enabling rapid domain adaptation with limited data. Extensive experiments on the LingoQA, DRAMA, and BDD-X datasets demonstrate that HoP consistently achieves state-of-the-art performance. Moreover, we propose an Efficient HoP variant that reduces computational overhead while maintaining high performance. We believe that HoP's structured integration of multi-level hints offers a promising direction for advancing VQA in complex, safety-critical scenarios.

Acknowledgement

This work was supported by the National Natural Science Foundation of China under Grant No. 62271466 and the National Key Research and Development Program of China Grant No. 2023YFF0716500.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 2

[2] 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. 5

[3] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*, pages 382–398. Springer, 2016. 6

[4] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 1(2):3, 2023. 2

[5] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005. 6

[6] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 3

[7] Li Chen, Penghao Wu, Kashyap Chitta, Bernhard Jaeger, Andreas Geiger, and Hongyang Li. End-to-end autonomous driving: Challenges and frontiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 1

[8] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198, 2024. 2, 3

[9] 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*, pages 1290–1299, 2022. 3, 5, 7, 8

[10] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, 2023. 2

[11] 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. *arXiv preprint arXiv:2310.02251*, 2023. 2

[12] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240): 1–113, 2023. 2

[13] 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. 6, 7

[14] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023. 2

[15] Kairui Ding, Boyuan Chen, Yuchen Su, Huan-ang Gao, Bu Jin, Chonghao Sima, Wuqiang Zhang, Xiaohui Li, Paul Barsch, Hongyang Li, et al. Hint-ad: Holistically aligned interpretability in end-to-end autonomous driving. *arXiv preprint arXiv:2409.06702*, 2024. 3

[16] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 5

[17] Roy Ganz, Yair Kittenplon, Aviad Aberdam, Elad Ben Avraham, Oren Nuriel, Shai Mazor, and Ron Litman. Question aware vision transformer for multimodal reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13861–13871, 2024. 3, 4, 5

[18] Sorin Grigorescu, Bogdan Trasnea, Tiberiu Cocias, and Gigel Macesanu. A survey of deep learning techniques for autonomous driving. *Journal of field robotics*, 37(3):362–386, 2020. 1

[19] Anthony Hu, Lloyd Russell, Hudson Yeo, Zak Murez, George Fedoseev, Alex Kendall, Jamie Shotton, and Gianluca Corrado. Gaia-1: A generative world model for autonomous driving. *arXiv preprint arXiv:2309.17080*, 2023. 3

[20] Bu Jin, Xinyu Liu, Yupeng Zheng, Pengfei Li, Hao Zhao, Tong Zhang, Yuhang Zheng, Guyue Zhou, and Jingjing Liu. Adapt: Action-aware driving caption transformer. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 7554–7561. IEEE, 2023. 6

[21] Jinkyu Kim, Anna Rohrbach, Trevor Darrell, John Canny, and Zeynep Akata. Textual explanations for self-driving vehicles. *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 2, 5, 6, 8

[22] Kenton Lee, Mandar Joshi, Iulia Raluca Turc, Hexiang Hu, Fangyu Liu, Julian Martin Eisenschlos, Urvashi Khandel-

wal, Peter Shaw, Ming-Wei Chang, and Kristina Toutanova. Pix2struct: Screenshot parsing as pretraining for visual language understanding. In *International Conference on Machine Learning*, pages 18893–18912. PMLR, 2023. 2

[23] Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. In *European conference on computer vision*, pages 280–296. Springer, 2022. 6

[24] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004. 6

[25] Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. Awq: Activation-aware weight quantization for on-device llm compression and acceleration. *Proceedings of Machine Learning and Systems*, 6:87–100, 2024. 8

[26] Ziyi Lin, Chris Liu, Renrui Zhang, Peng Gao, Longtian Qiu, Han Xiao, Han Qiu, Chen Lin, Wenqi Shao, Keqin Chen, et al. Sphinx: The joint mixing of weights, tasks, and visual embeddings for multi-modal large language models. *arXiv preprint arXiv:2311.07575*, 2023. 3

[27] 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*, pages 26296–26306, 2024. 2, 3, 5

[28] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 2, 3

[29] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 3, 5, 7

[30] Zuyan Liu, Yuhao Dong, Ziwei Liu, Winston Hu, Jiwen Lu, and Yongming Rao. Oryx mllm: On-demand spatial-temporal understanding at arbitrary resolution. *arXiv preprint arXiv:2409.12961*, 2024. 2

[31] Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Yaofeng Sun, et al. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*, 2024. 3

[32] Yunsheng Ma, Amr Abdelraouf, Rohit Gupta, Ziran Wang, and Kyungtae Han. Video token sparsification for efficient multimodal llms in autonomous driving. *arXiv preprint arXiv:2409.11182*, 2024. 5, 6

[33] Srikanth Malla, Chiho Choi, Isht Dwivedi, Joon Hee Choi, and Jiachen Li. Drama: Joint risk localization and captioning in driving. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1043–1052, 2023. 2, 5, 6, 8

[34] Jiageng Mao, Yuxi Qian, Junjie Ye, Hang Zhao, and Yue Wang. Gpt-driver: Learning to drive with gpt. *arXiv preprint arXiv:2310.01415*, 2023. 2

[35] Ana-Maria Marcu, Long Chen, Jan Hünermann, Alice Karnsund, Benoit Hanotte, Prajwal Chidananda, Saurabh Nair, Vijay Badrinarayanan, Alex Kendall, Jamie Shotton, et al. Lingoqa: Video question answering for autonomous driving. *arXiv preprint arXiv:2312.14115*, 2023. 2, 5, 6, 7, 8

[36] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 2, 3, 5, 7, 8

[37] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002. 6

[38] Tianwen Qian, Jingjing Chen, Linhai Zhuo, Yang Jiao, and Yu-Gang Jiang. Nuscenes-qa: A multi-modal visual question answering benchmark for autonomous driving scenario. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4542–4550, 2024. 2

[39] 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*, pages 8748–8763. PMLR, 2021. 2, 3, 4

[40] 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*, pages 15120–15130, 2024. 3

[41] Himanshu Sharma and Anand Singh Jalal. A survey of methods, datasets and evaluation metrics for visual question answering. *Image and Vision Computing*, 116:104327, 2021. 2

[42] Leyang Shen, Gongwei Chen, Rui Shao, Weili Guan, and Liqiang Nie. Mome: Mixture of multimodal experts for generalist multimodal large language models. *Advances in neural information processing systems*, 37:42048–42070, 2025. 2, 3

[43] Chonghao Sima, Katrin Renz, Kashyap Chitta, Li Chen, Hanxue Zhang, Chengen Xie, Ping Luo, Andreas Geiger, and Hongyang Li. Drivelm: Driving with graph visual question answering. *arXiv preprint arXiv:2312.14150*, 2023. 2, 3

[44] Ran Tian, Boyi Li, Xinshuo Weng, Yuxiao Chen, Edward Schmerling, Yue Wang, Boris Ivanovic, and Marco Pavone. Tokenize the world into object-level knowledge to address long-tail events in autonomous driving. *arXiv preprint arXiv:2407.00959*, 2024. 3

[45] Xiaoyu Tian, Junru Gu, Bailin Li, Yicheng Liu, Chenxu Hu, Yang Wang, Kun Zhan, Peng Jia, Xianpeng Lang, and Hang Zhao. Drivevlm: The convergence of autonomous driving and large vision-language models. *arXiv preprint arXiv:2402.12289*, 2024. 3

[46] Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, et al. Cambrian-

1: A fully open, vision-centric exploration of multimodal llms. *arXiv preprint arXiv:2406.16860*, 2024. 2, 3

[47] Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9568–9578, 2024. 2, 3, 5

[48] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 2

[49] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017. 4

[50] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015. 6

[51] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*, 2023. 2

[52] Xiaofeng Wang, Zheng Zhu, Guan Huang, Xinze Chen, Jiagang Zhu, and Jiwen Lu. Drivedreamer: Towards real-world-driven world models for autonomous driving. *arXiv preprint arXiv:2309.09777*, 2023. 3

[53] 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*, 2024. 2, 6

[54] Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*, 9(1):1, 2023. 2, 5

[55] 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. 2