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

## Supplementary Material

## 1. Additional Experimental Results

### 1.1. Data Efficiency and Lightweight Inference

| Dataset Ratio | Lingo-Judge↑ | | |
|---|---|---|---|
| | LLaVA-v1.5 | HoP (Ours) | Δ |
| 25% | 60.0 | 64.0 | +4.0 |
| 50% | 60.6 | 65.6 | +5.0 |
| 100% | 63.2 | 67.8 | +4.6 |

Table 1. **Data-efficient domain adaptation of HoP.** Δ: the performance gain of HoP over LLaVA-v1.5 at same data ratio.

| LLM | Method | LJ ↑ | Latency (ms) |
|---|---|---|---|
| Qwen-v2-0.5B | Baseline | 54.8 | 295 |
| | + Efficient HoP | 57.6 (+2.8) | 302 |
| Qwen-v2.5-3B | Baseline | 61.2 | 483 |
| | + Efficient HoP | **64.6** **(+3.4)** | 504 |
| | + Efficient HoP (AWQ) | 64.2 (+3.0) | **281** |

Table 2. Performance comparison of Efficient HoP with baseline models. Green numbers indicate LJ (Lingo-Judge) improvements. AWQ indicates the quantized model.

As shown in Tab. 1 and Tab. 2, our HoP framework consistently outperforms LLaVA-v1.5 across all data regimes and model scales. It improves Lingo-Judge scores by up to +5.0 with only 50% training data, highlighting its effectiveness in low-data scenarios. Even with full data, HoP maintains a +4.6 advantage, demonstrating strong visual-language alignment. Furthermore, the Efficient HoP variant achieves 41.8% lower latency via 4-bit AWQ quantization while preserving competitive performance, confirming its scalability and deployment readiness.

### 1.2. Evaluation on Planning Task

To assess the applicability of HoP in real-world autonomous driving scenarios, we evaluate it on the nuScenes open-loop planning benchmark following the OmniDrive setup. As shown in Tab. 3, HoP surpasses all baselines including DriveLM and LLaVA-v1.5 across multiple metrics (lower is better), such as L2 distance, collision rate, and intersection violations. This demonstrates HoP's capability to generate more accurate and safer trajectories by leveraging enriched visual-language representations.

| Method | L2 (m) ↓ | | | | Collision (%) ↓ | | | | Intersection (%) ↓ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1s | 2s | 3s | Avg. | 1s | 2s | 3s | Avg. | 1s | 2s | 3s | Avg. |
| DriveLM† | 1.32 | 2.08 | 3.01 | 2.14 | 0.42 | 2.01 | 4.12 | 2.18 | 0.88 | 3.12 | 6.45 | 3.48 |
| LLaVA-v1.5 | 1.28 | 2.13 | 3.22 | 2.21 | 0.16 | 1.68 | 3.52 | 1.79 | 1.02 | 3.44 | 7.00 | 3.82 |
| HoP | **1.07** | **1.81** | **2.62** | **1.83** | 0.25 | **1.30** | **2.17** | **1.24** | **0.21** | **2.13** | **5.06** | **2.47** |

Table 3. Planning results on nuScenes. †: fair DriveLM reproduction.

### 1.3. Temporal Consistency Analysis

To further evaluate the stability of HoP's predictions across time, we employ the Trajectory Prediction Consistency (TPC) metric introduced in MomAD. TPC measures frame-to-frame deviation between consecutive trajectory predictions. As shown in Tab. 4, HoP achieves lower TPC scores than LLaVA-v1.5, indicating improved temporal coherence, despite operating with frame-wise vision features.

| Method | TPC@1s ↓ | TPC@2s ↓ | TPC@3s ↓ | Avg. ↓ |
|---|---|---|---|---|
| LLaVA-v1.5 | 0.49 | 0.85 | 1.24 | 0.86 |
| HoP | **0.46** | **0.81** | **1.18** | **0.82** |

Table 4. Trajectory prediction consistency (lower is better).

### 1.4. Robustness Under Long-tail Distributions

We further evaluate HoP on CODA-LM, a benchmark designed to measure visual-language reasoning under long-tail distributions. As shown in Tab. 5, HoP achieves the best overall performance and outperforms CODA-VLM in three of four sub-metrics, despite the latter using a stronger backbone (LLaVA-Llama-3-8B-v1.1).

| Method | Final Score ↑ | General ↑ | Region ↑ | Suggestion ↑ |
|---|---|---|---|---|
| LLaVA-v1.5 | 28.17 | 19.30 | 42.06 | 23.16 |
| CODA-VLM | 63.62 | 55.04 | **77.68** | 58.14 |
| HoP | **64.38** | **57.22** | 76.93 | **59.00** |

Table 5. Results on CODA-LM benchmark.

### 1.5. Generalizability to Stronger MLLMs

To validate HoP's compatibility with modern vision-language backbones, we integrate it with Qwen2.5-VL. As reported in Tab. 6, HoP yields a +2.6 gain in Lingo-Judge score over the Qwen2.5-VL base model, establishing a new SOTA on LingoQA and confirming HoP's general applicability across architectures.

| Method | LJ ↑ | BLEU-4 ↑ | METEOR ↑ | CIDEr ↑ |
|---|---|---|---|---|
| Qwen2.5-VL-7B | 68.80 | 13.28 | **19.52** | 60.31 |
| Qwen2.5-VL-HoP | **71.40** | **13.31** | 19.51 | **60.59** |

Table 6. Results on LingoQA with Qwen2.5-VL.


Figure 1. **Visualization results on the LingoQA dataset.** LLaVA-v1.5 serves as our baseline. Abbreviations: Q (Question), GT(Ground Truth).

## 2. More Quantitative Results

We present additional qualitative results of HoP on the LingoQA, DRAMA, and BDD-X datasets, as illustrated in Fig. 1, Fig. 2, and Fig. 3.

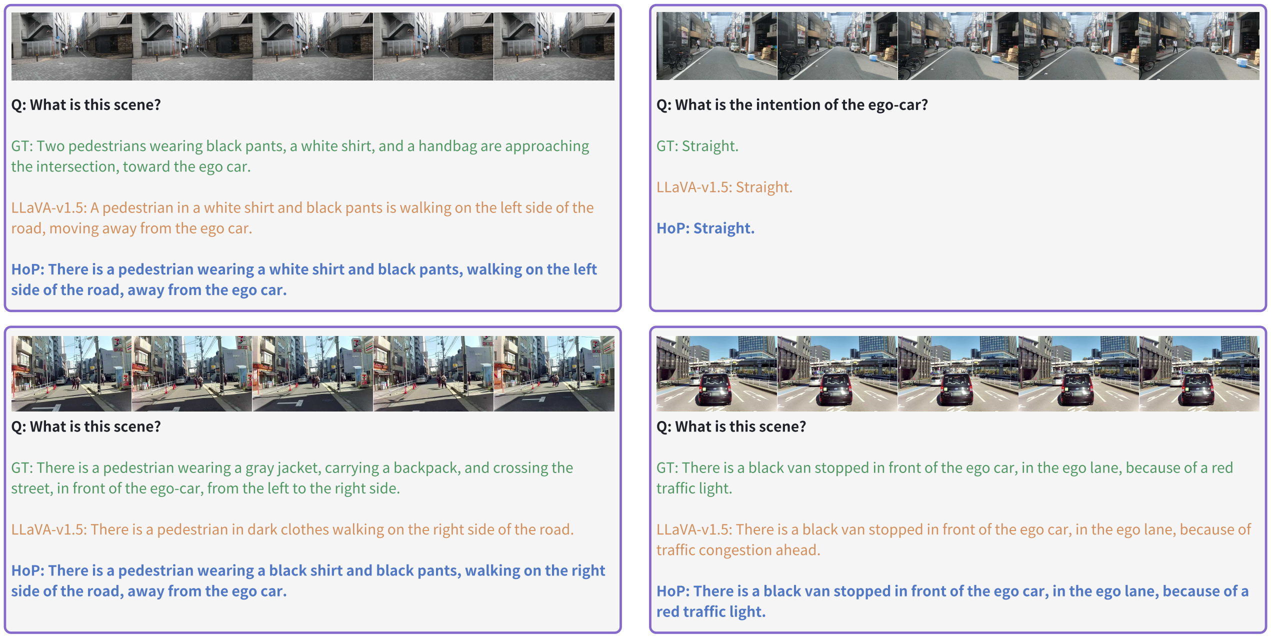


Figure 2. **Visualization results on the DRAMA dataset.** LLaVA-v1.5 serves as our baseline. Abbreviations: Q (Question), GT(Ground Truth).

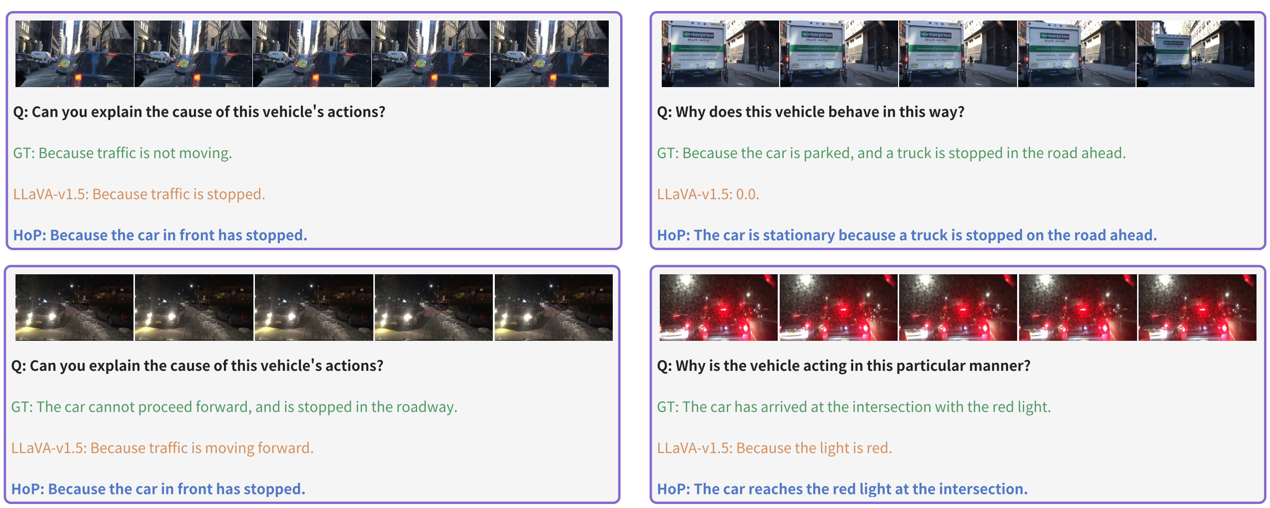


Figure 3. **Visualization results on the BDD-X dataset.** LLaVA-v1.5 serves as our baseline. Abbreviations: Q (Question), GT(Ground Truth).