UMA-AD: A Unified Multimodal Alignment Framework for Interpretable Autonomous Driving

Guangqiang Li[1], Dehui Du[†1], Xing Yu[1]
[1]Software Engineering Institute, East China Normal University, Shanghai, China
51275902059@stu.ecnu.edu.cn, dhdu@sei.ecnu.edu.cn, 51265902017@stu.ecnu.edu.cn

***Abstract*—End-to-end autonomous driving systems have demonstrated exceptional capability in complex environments but suffer from opacity, hindering user trust and safety assurance. While recent interpretability approaches attempt to align linguistic explanations with intermediate perception outputs, they often face two critical limitations: the loss of fine-grained semantics due to Bird's-Eye-View (BEV) compression and the lack of explicit supervision for structured reasoning. To address these challenges, we propose UMA-AD, a novel framework designed to unify multimodal alignment for interpretable autonomous driving. First, we design a Unified Token Aligner module to adapt heterogeneous intermediate outputs from the AD model for the language decoder. Within this module, we integrate a Dual-Branch Video-Enhanced BEV Former to explicitly inject multi-scale video features, effectively compensating for semantic loss in BEV representations. Furthermore, we construct a structured Chain-of-Thought supervision strategy to internalize the hierarchical logic of the "Perception-Prediction-Planning" pipeline, empowering the model to generate explicit reasoning steps, while synchronously predicting low-level control signals to ensure the physical grounding of generated explanations. Extensive experiments on the Nu-X, TOD3Cap, and NuScenes-QA benchmarks validate the effectiveness of our framework in generating logically rigorous linguistic explanations.**

***Index Terms*—autonomous driving, interpretability, chain-of-thought reasoning, multimodal large language models**

I. Introduction

End-to-end autonomous driving frameworks have rapidly emerged as a dominant solution in modern transportation, demonstrating exceptional capability by integrating perception, prediction, and planning into a unified architecture. While this data-driven paradigm significantly improves driving performance [1], [2] in complex urban environments, it simultaneously introduces a severe challenge regarding model transparency. The deep neural networks underpinning these systems often operate as black boxes, rendering the internal reasoning process inaccessible to human users. This opacity makes it difficult for passengers to comprehend the rationale behind vehicle behaviors, which fundamentally undermines user trust and raises critical safety concerns. Therefore, addressing the interpretability of these systems is essential, as providing clear explanations for driving decisions is a prerequisite for the reliable deployment and broader societal acceptance of autonomous vehicles.

To tackle this interpretability bottleneck, natural language, serving as a highly user-friendly communication medium, has been explored to enhance interpretability through tasks such as driving explanation and Visual Question Answering (VQA) [3]. Driven by this goal, research paradigms have evolved from early visual captioning frameworks [4], [5] to sophisticated Multimodal Large Language Models (MLLMs) [6], [7]. Notably, the emerging "aligned paradigm" [8] represents a significant leap by grounding linguistic explanations in intermediate perception outputs, effectively mitigating the hallucination issues common in post-hoc reasoning. However, despite these advancements, current aligned approaches still face two fundamental limitations that hinder precise and transparent explainability. First, regarding physical grounding, these architectures predominantly rely on compressed Bird's-Eye-View (BEV) features [9]. As indicated by recent studies [10], [11], the transformation from perspective view to BEV inevitably discards critical vertical geometric information (e.g., traffic light states), thereby limiting the physical precision of visual perception. Second, regarding reasoning logic, the generation process lacks explicit supervision for structured reasoning. Without strictly adhering to the hierarchical "Perception-Prediction-Planning" logic [12], existing models struggle to articulate the sequential dependencies behind decisions, resulting in explanations that, while plausible, fail to authentically reflect the deep underlying rationale of the driving system, as illustrated in Fig. 1.

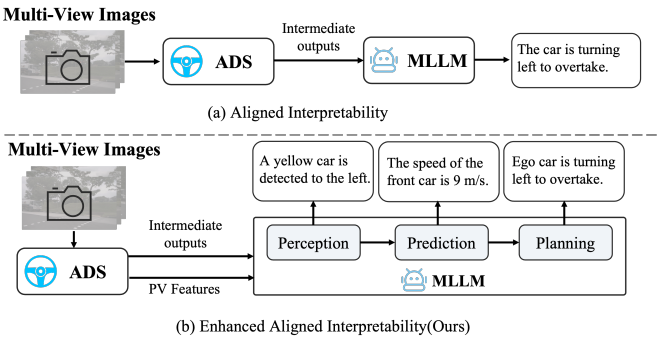


Fig. 1. Comparison of interpretability paradigms in end-to-end autonomous driving. (a) Aligned interpretability generates explanations grounded in intermediate driving outputs. (b) Enhanced aligned interpretability integrates raw PV features for richer semantics and structured Chain-of-Thought reasoning for hierarchical transparency.

To address these limitations, we introduce UMA-AD (Unified Multimodal Alignment for Autonomous Driving). First,

† Corresponding author.

we design a Unified Token Aligner module to effectively align the heterogeneous intermediate outputs of the AD pipeline with the language decoder. Within this module, to address the inherent loss of fine-grained semantics in BEV representations, we integrate a Dual-Branch Video-Enhanced BEV Former, which explicitly injects multi-scale video features to enhance the physical precision of visual perception. Second, regarding logical reasoning, we adapt the DriveLM dataset [12] to construct a structured CoT strategy that strictly adheres to the "Perception-Prediction-Planning" hierarchy, thereby equipping the model with structured reasoning capabilities. Furthermore, the system synchronously predicts low-level control signals to strengthen the physical grounding of the linguistic explanations. We validate the effectiveness of our approach on the Nu-X [8], TOD3Cap [14], and NuScenes-QA [15] benchmarks. The main contributions of this work are briefly summarized as follows:

- We propose UMA-AD, a unified multimodal Alignment framework for interpretable autonomous driving. Capable of processing multi-modal inputs, UMA-AD generates linguistic explanations as well as auxiliary low-level control signals.
- We design a Unified Token Aligner module that adapts intermediate output tokens from the AD model for the language decoder. Furthermore, we integrate a Dual-Branch Video-Enhanced BEV Former within this module to explicitly inject multi-scale video features, compensating for the semantic loss in BEV representations.
- We construct a structured CoT supervision strategy derived from the DriveLM dataset. This strategy enforces the model to internalize the sequential logic of "Perception-Prediction-Planning" during training, thereby empowering it to generate explicit and structured reasoning chains during inference.

II. Related Work

A. Interpretable Autonomous Driving

Explainability in ADS has evolved from visual saliency to natural language generation [3]. Early methods visualized influential regions via attention maps [16] or deconvolutional networks [17]. Moving to semantic insights, BDD-X [4] pioneered linguistic justifications, followed by frameworks integrating object grounding [18], joint control-caption prediction [5], and risk localization [7]. Recent advancements leverage LLMs for end-to-end alignment: DriveGPT4 [6] interprets vehicle actions via video-control history, while Hint-AD [8] aligns 3D tokens into the language space. To evaluate logical depth, DriveLM [12] introduces a graph-structured VQA benchmark, explicitly assessing multi-stage structured reasoning across perception, prediction, and planning.

B. Multimodal Large Language Models

Advancements in MLLMs, exemplified by GPT-4 [19], LLaVA [20], Qwen-VL [21], and InstructBLIP [22], alongside video-centric models [23], have established robust foundations for cross-modal reasoning by aligning visual features with textual embeddings. In autonomous driving, researchers integrate domain-specific modalities—such as LiDAR [24], BEV maps [25], and vectorized elements [26]—directly into the language space to enhance spatial awareness. Crucially, this alignment bridges the gap between abstract geometric data and high-level semantic understanding, allowing models to perceive the underlying structure of dynamic 3D environments with greater fidelity. Furthermore, frameworks like Dolphins [27] and DiLu [28] extend these capabilities to behavioral planning and knowledge-driven reasoning, demonstrating the potential of MLLMs in handling highly complex driving scenarios.

C. Chain-of-Thought Reasoning

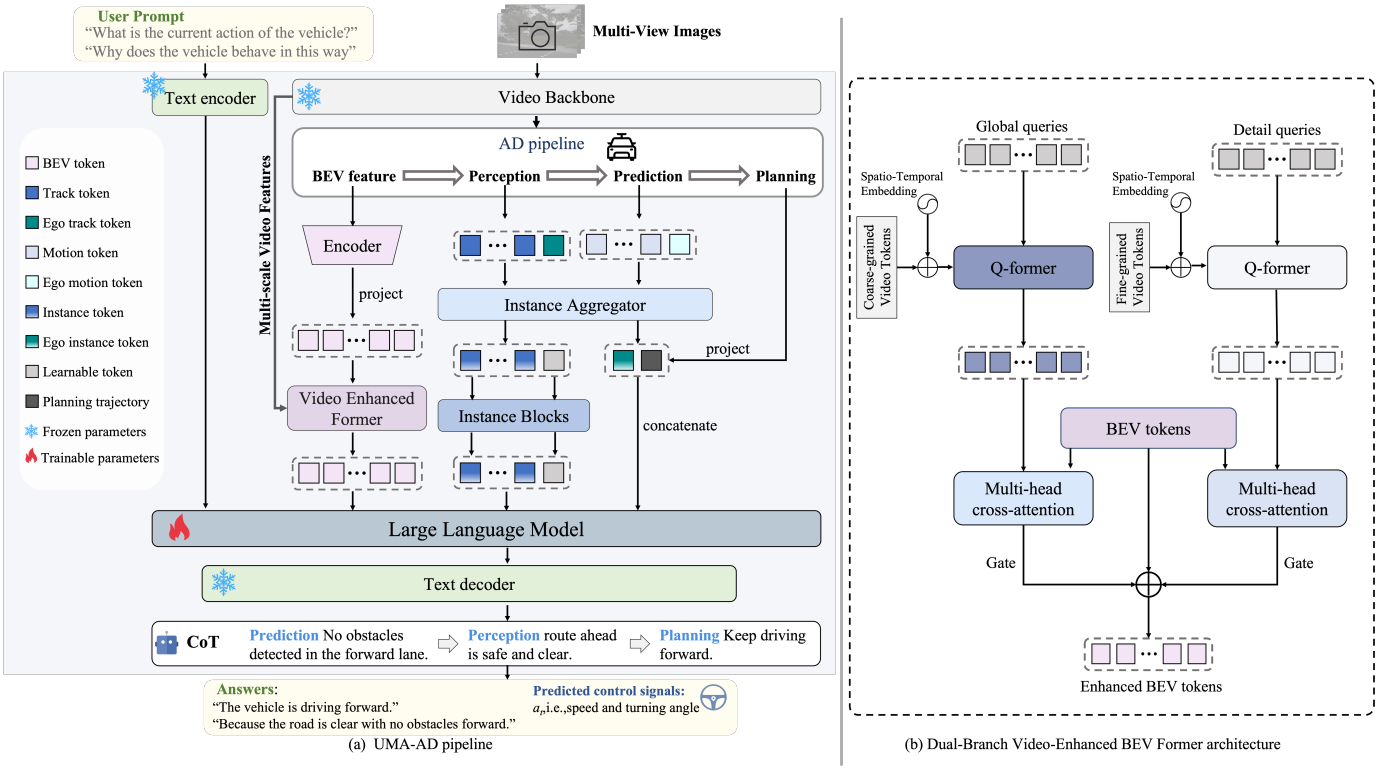
CoT reasoning fundamentally enhances LLMs by decomposing complex tasks into sequential logical steps [29], mimicking human cognition. By explicitly generating intermediate reasoning paths, models can bridge the gap between input questions and final answers more effectively. This paradigm has evolved from zero-shot triggers like "Let's think step by step" [30] to advanced self-consistency strategies [31]. In autonomous driving, CoT is pivotal for interpreting the sequential nature of decision-making. Specifically, it allows systems to not only react to traffic scenarios but also explain the decision-making rationale behind their maneuvers, thereby improving transparency and trust. For instance, DriveVLM [10] employs a CoT pipeline to process spatio-temporal data for long-tail scenarios, while DiLu [28] leverages reasoning modules to apply historical driving knowledge. Similarly, Dolphins [27] utilizes in-context learning to enable conversational reasoning about traffic rules.

III. Methodology

We propose UMA-AD to bridge the gap between autonomous driving systems and linguistic interpretations. An overview of the framework is presented in Fig. 2. The framework comprises an AD backbone, a Unified Token Aligner, and an LLM decoder. Any end-to-end system featuring perception, prediction, and planning can serve as the AD backbone.

A. Overview

As illustrated in Fig. 2, the framework pipeline proceeds as follows. Firstly, we extract hierarchical intermediate queries from a frozen AD backbone, specifically yielding track tokens, motion tokens, and planning tokens, alongside multi-scale raw video features from the image backbone. Secondly, a Unified Token Aligner module adapts these heterogeneous inputs for the language decoder. Within this module, we design an Instance Aggregator to merge instance-level track and motion information, followed by Instance Blocks to unify variable-length tokens into a fixed-length representation. Crucially, we integrate a Dual-Branch Video-Enhanced BEV Former to explicitly inject multi-scale video features, compensating for semantic loss in BEV representations. All processed tokens are concatenated as context prompts (see Sec. III-B). Finally, the



(a) UMA-AD pipeline

(b) Dual-Branch Video-Enhanced BEV Former architecture

Fig. 2. Overview of the UMA-AD framework.(a) UMA-AD pipeline. Taking intermediate output tokens from the AD pipeline as input, the LLMs generates driving explanations and control signals.(b) Detailed architecture of the Dual-Branch Video-Enhanced BEV Former.

aligned context is fed into the LLM. Drawing inspiration from RT-2 [32], we treat control signals as a linguistic modality. Consequently, the model generates a unified output sequence: it explicitly articulates CoT reasoning steps (adhering to the "Perception-Prediction-Planning" hierarchy) to derive the driving explanations, while synchronously predicting control signals (i.e., speed and turning angle) in a fixed format.

B. Unified Token Aligner

The query tokens extracted from the AD pipeline are not directly understandable to the language decoder. Addressing this, we propose the Unified Token Aligner architecture, which aligns heterogeneous inputs into a unified feature space.

Input Denotations. We start by defining the query tokens extracted from the AD pipeline:

- *BEV Tokens:* $F_{bev} \in \mathbb{R}^{H_b \times W_b \times C}$ encode the static environmental context, where H_b, W_b, and C denote the height, width, and channel dimension of the BEV feature map, respectively.
- *Track Tokens:* $\{F^i_{track}\}^{N_{det}}_{i=1} \in \mathbb{R}^D$ contain the position and past trajectory information of each detected agent, where N_{det} represents the number of detected instances and D is the token dimension.
- *Motion Tokens:* $\{F^i_{motion}\}^{N_{det}}_{i=1} \in \mathbb{R}^D$ encode the predicted multimodal future trajectories corresponding to each detected agent in the track tokens.
- *Planning Tokens:* $F_{plan} \in \mathbb{R}^{T_p \times 2}$ represents the future ego-vehicle trajectory planned by the model, where T_p denotes the number of future waypoints.
- *Video Features:* $\mathcal{V} = \{V_l\}^L_{l=1}$ denotes the multi-scale video features extracted from the shared 2D image backbone (e.g., ResNet [33] equipped with FPN). Crucially, these features are obtained prior to the view transformation, preserving fine-grained semantics in the Perspective View (PV). They correspond to T consecutive frames captured by N_{cam} surrounding cameras, where L is the number of feature scales.

Instance Processing. To generate a unified representation for detected instances, we design an Instance Aggregator that fuses the heterogeneous attributes of each instance. Given the i-th instance, we first concatenate its track token F^i_{track} and motion token F^i_{motion}, and adaptively integrate them via a Gated Linear Unit variant:

$$g^i = \sigma(P_{gate}(\text{Concat}(F^i_{track}, F^i_{motion}))) \quad (1)$$

$$F^i_{ins} = P_{proj}(\text{Concat}(F^i_{track}, F^i_{motion})) \odot g^i \quad (2)$$

where P_{proj} projects the concatenated features to the target dimension D_{embed}, and P_{gate} is an MLP with Sigmoid activation $\sigma(\cdot)$ producing the gating weights g^i. Subsequently, to unify the variable number of instances N_{det}, we employ *Instance Blocks*, where a set of learnable queries Q_{ins} aggregate features via Multi-Head Cross-Attention (MHCA):

$$F'_{ins} = \text{MHCA}(Q_{ins}, F_{ins}) \quad (3)$$

This yields a fixed set of instance tokens $F'_{ins} \in \mathbb{R}^{N_{ins} \times D_{embed}}$.

Dual-Branch Video-Enhanced BEV Former. To recover fine-grained textures lost during view transformation, we propose a dual-branch architecture (detailed in Fig. 2(b)), comprising a *Global Branch* (g) and a *Detail Branch* (d). First, the feature encoder E is implemented as a multi-layer convolutional network to extract features from the raw BEV input F_{bev} and down-scale the spatial resolution. Subsequently, an MLP projector P_{bev} maps the channel dimension from C to D_{embed}, yielding:

$$F'_{bev} = P_{bev}(E(F_{bev})) \tag{4}$$

Simultaneously, we process the multi-scale video features $\mathcal{V}$ to serve as inputs for the dual branches. To balance computational efficiency with semantic fidelity, we select the feature map from the coarsest level as the global input V_g to capture holistic context, and the map from the intermediate level (the second finest scale) as the detail input V_d to preserve fine-grained geometric cues. Subsequently, to encode spatio-temporal positional information, learnable time embeddings E_t and camera embeddings E_c are added to these selected features. Semantic content is then extracted via Q-Formers [22]:

$$F_k = \text{Q-Former}(Q_k, V_k + E_t + E_c), \quad k \in \{g, d\} \tag{5}$$

Finally, the extracted visual tokens are injected into the BEV query F'_{bev} via Multi-Head Cross-Attention (MHCA). We denote the attention output for each branch as Δ_k:

$$\Delta_k = \text{MHCA}(F'_{bev}, F_k), \quad k \in \{g, d\} \tag{6}$$

To adaptively balance these contributions, we predict gating weights via an MLP projector P^k_{gate} and perform a gated residual fusion:

$$w_k = \sigma(P^k_{gate}(F'_{bev})), \quad k \in \{g, d\} \tag{7}$$

$$F''_{bev} = \text{LN}(F'_{bev} + w_g \cdot \Delta_g + w_d \cdot \Delta_d) \tag{8}$$

Context Concatenation. The planning tokens F_{plan} are encoded by positional encoding (PE) and a projector P_{plan}. Incorporating the ego-vehicle token $F^{ego}_{instance}$, the final context sequence is:

$$F_{ctx} = \text{Concat}(F''_{bev}, F'_{ins}, F^{ego}_{instance}, P_{plan}(\text{PE}(F_{plan}))) \tag{9}$$

C. Structured CoT Reasoning and Control Signals

Structured CoT Reasoning. To equip the model with capabilities for both precise visual grounding and logical reasoning, we implement a data adaptation strategy on the DriveLM dataset. First, we perform Format Adaptation to address the representation mismatch; specifically, we remap the original abstract object tags (e.g., <c1>) into explicit instance identifiers (e.g., <obj_k>) that align one-to-one with the indices of our model's input track tokens via coordinate matching. This transformation establishes a rigorous correspondence, ensuring that the language decoder can accurately ground textual descriptions to the specific physical entities captured by the perception module. Furthermore, inspired by DriveMLM [13], we implement Decision State Serialization. We restructure the originally discrete QA pairs into a continuous CoT sequence, which is then paired with a standardized reasoning instruction to form a unified instruction-response sample. This formulation serves as a supervisory signal, compelling the model to internalize the sequential order of *Perception* $\rightarrow$ *Prediction* $\rightarrow$ *Planning*. During inference, this enables the model to produce coherent reasoning chains in response to task-specific instructions, bridging the gap between intermediate features and linguistic logic.

Control Signals. We treat control signals and natural language as a unified modality within a shared token space. Specifically, the model generates a control tuple (v_{t+1}, θ_{t+1}) (i.e., velocity and steering angle) for the next timestamp in a fixed format. To provide kinematic context, current vehicle states are integrated into the input prompts, with supervision derived from nuScenes CAN bus records to enhance physical grounding.

IV. Experiments

A. Datasets and Evaluation Metrics

Datasets. We conduct comprehensive evaluations on three benchmarks. Nu-X [8], meticulously constructed upon the widely utilized nuScenes dataset [34], serves as the primary benchmark for driving explanation. By providing large-scale, high-quality human annotations across diverse and complex driving scenarios, it offers a robust foundation for evaluating model generalization. TOD3Cap [14] is employed to assess 3D dense captioning performance, focusing on detailed object-level descriptions. NuScenes-QA [15], also derived from the nuScenes dataset, serves as the benchmark for visual question answering to test spatial reasoning accuracy. Additionally, we utilize the adapted DriveLM [12] dataset solely during the training phase. We leverage the hierarchical "Perception-Prediction-Planning" logic embedded in DriveLM's graph-structured annotations to supervise our CoT reasoning. Crucially, since Nu-X, NuScenes-QA, and DriveLM all originate from nuScenes keyframes, this shared visual foundation ensures consistent data distribution, enabling a strict zero-leakage evaluation of reasoning generalization.

Evaluation Metrics. We evaluate the performance of our method using a diverse set of metrics widely adopted in text generation tasks. These metrics collectively measure various aspects of the generated text, such as lexical overlap, semantic alignment, and fluency. Specifically, for captioning tasks on Nu-X and TOD3Cap, we report BLEU-4 (B4) [35], METEOR (M) [36], ROUGE-L (R) [37], and CIDEr (C) [38]. Following standard protocols [14], we set an IoU threshold of 0.25 for matching object proposals in TOD3Cap. For NuScenes-QA, we report accuracy by directly comparing generated texts with ground truth, categorized into Zero-hop (H0), One-hop (H1), and Overall (All). Furthermore, considering the limitations of n-gram metrics in capturing semantic logic, we employ GPT-4 scoring (G) to score generated explanations (scale 0–10) based on logical coherence and factual accuracy. For

TABLE I
QUANTITATIVE RESULTS ON NU-X, TOD3CAP, AND NUSCENES-QA BENCHMARKS.

Method	Nu-X					TOD3Cap				NuScenes-QA		
	C ↑	B4 ↑	M ↑	R ↑	G ↑	C ↑	B4 ↑	M ↑	R ↑	H0 ↑	H1 ↑	All ↑
GPT-4o	19.0	3.95	10.3	24.9	4.85	160.8	50.4	31.6	43.5	42.0	34.7	37.1
Gemini 1.5 Pro	17.6	3.43	9.3	23.4	4.62	169.7	53.6	33.4	45.9	40.5	32.9	35.4
ADAPT [5]	17.7	2.06	12.8	**27.9**	5.53	-	-	-	-	51.0	44.2	46.4
TOD3Cap [14]	14.5	2.45	10.5	23.0	4.92	120.3	51.5	45.1	70.1	53.0	45.1	49.0
DriveGPT4 [6]	19.8	4.08	11.2	25.4	6.15	-	-	-	-	48.5	42.1	45.3
Hint-UniAD [8]	21.7	**4.20**	12.7	27.0	6.95	342.6	71.9	**48.0**	85.4	56.2	47.5	50.4
Hint-VAD [8]	22.4	4.18	13.2	27.6	7.18	263.7	67.6	47.5	79.4	55.4	48.0	50.5
UMA-UniAD (Ours)	23.4	4.15	13.1	26.9	7.62	**351.2**	**72.5**	47.7	**86.1**	56.4	49.6	52.0
UMA-VAD (Ours)	**24.1**	4.12	**13.6**	27.4	**7.88**	275.4	68.5	47.2	80.8	**56.7**	**50.1**	**52.4**

TABLE II
QUANTITATIVE RESULTS OF CONTROL SIGNALS PREDICTION ON THE NU-X TESTING DATASET.

Method	Speed (m/s)					Turning angle (degree)				
	RMSE ↓	A0.1 ↑	A0.5 ↑	A1.0 ↑	A5.0 ↑	RMSE ↓	A0.1 ↑	A0.5 ↑	A1.0 ↑	A5.0 ↑
DriveGPT4 [6]	1.05	33.15	65.42	75.40	98.10	5.52	60.12	68.55	72.80	95.88
UMA-UniAD (Ours)	**0.82**	**36.85**	**70.12**	**81.50**	**99.25**	**3.85**	**65.40**	**76.25**	**83.20**	**97.50**

control signal prediction, we adopt Root Mean Squared Error (RMSE) to measure the regression error for speed and steering angle. Additionally, we report the Threshold Accuracy ($A\tau$), calculating the percentage of predictions where the error falls within a specific threshold $\tau \in \{0.1, 0.5, 1.0, 5.0\}$.

B. Baselines and Implementation Details

Baselines. We compare UMA-AD against two categories of methods: (1) Specialist Models: ADAPT [5], DriveGPT4 [6], Hint-AD [8], and TOD3Cap [14]; (2) General MLLMs: GPT-4o and Gemini-1.5 Pro, evaluated in a zero-shot setting. For fairness, all vision-based baselines, including general MLLMs, are fed with multi-view video frames to capture temporal dynamics for reasoning.

Training Pipeline. We adopt a progressive two-stage training strategy. *Stage 1: Representation Alignment.* This stage aims to map the AD model's intermediate tokens into the linguistic semantic space. We employ the online alignment tasks [8] (covering counting, localization, motion, and planning alignment) alongside the TOD3 task [14] to enforce robust semantic connections between visual tokens and physical entities. *Stage 2: End-to-End Instruction Tuning.* We perform joint training on nuScenesX [8], NuScenes-QA [15], and the adapted DriveLM. Specifically, nuScenesX provides primary supervision for driving explanations and control signals, while DriveLM offers auxiliary CoT reasoning supervision. This multi-task paradigm leverages reasoning chains to deepen the model's comprehension of the driving environment.

Implementation Details. We implement UMA-AD on two representative backbones to verify universality: UniAD [39] (rasterized) and VAD [40] (vectorized). Following standard settings, the video sequence length is $T = 3$, and we employ Llama-3.2-3B as the decoder. The model is implemented in PyTorch and fine-tuned via LoRA for 10 epochs using AdamW (learning rate $2e^{-4}$, cosine decay). Experiments are conducted on 8 NVIDIA RTX 4090 GPUs with a batch size of 4.

C. Main Results

As shown in Table I, the proposed UMA-AD framework consistently achieves improved performance across key metrics, outperforming both general-purpose MLLMs (e.g., GPT-4o and Gemini 1.5 Pro) and specialist baselines (e.g., DriveGPT4 and Hint-AD), which demonstrates its robustness across different tasks and evaluation dimensions. On Nu-X, the primary benchmark for driving explanation, UMA-VAD obtains the best results with a CIDEr score of 24.1 and a GPT-Score of 7.88, surpassing the previous best method by 1.7 and 0.70 points, respectively. These results indicate that UMA-AD can more effectively transform intermediate driving states into language-interpretable representations, leading to driving explanations that are more accurate and logically coherent.

Beyond driving explanation, the proposed framework also exhibits strong versatility on specialized tasks. On TOD3Cap for fine-grained 3D dense captioning, UMA-UniAD achieves a high CIDEr score of 351.2, demonstrating enhanced capability in describing detailed object attributes and rich scene textures. This further suggests that incorporating video-enhanced features can effectively supplement fine-grained semantics that are missing in BEV representations, thereby improving the model's expressiveness for complex scene details. Meanwhile, UMA-VAD achieves the highest overall accuracy of 52.4% on the NuScenes-QA benchmark, reflecting its stronger spatial understanding and reasoning ability. Taken together, these quantitative results comprehensively validate the robust effectiveness and superiority of UMA-AD across diverse interpretation-oriented autonomous driving tasks.

TABLE III
ABLATION ON DUAL-BRANCH VIDEO-ENHANCED BEV FORMER.

Method	Nu-X					TOD3Cap				NuScenes-QA		
	C ↑	B4 ↑	M ↑	R ↑	G ↑	C ↑	B4 ↑	M ↑	R ↑	H0 ↑	H1 ↑	All ↑
w/o Video Enhancement	21.9	4.05	12.8	26.2	7.15	343.2	72.0	47.1	85.5	55.5	48.2	50.8
w/o Detail Branch	22.4	4.10	12.9	26.5	7.32	345.5	72.1	47.3	85.7	55.9	48.8	51.3
w/o Global Branch	22.9	**4.17**	13.0	26.8	7.51	350.1	72.4	47.6	86.0	**56.5**	49.2	51.6
Full Model (UMA-UniAD)	**23.4**	4.15	**13.1**	**26.9**	**7.62**	**351.2**	**72.5**	**47.7**	**86.1**	56.4	**49.6**	**52.0**

TABLE IV
ABLATION ON UNIFIED TOKEN ALIGNER.

Method	Nu-X					TOD3Cap				NuScenes-QA		
	C ↑	B4 ↑	M ↑	R ↑	G ↑	C ↑	B4 ↑	M ↑	R ↑	H0 ↑	H1 ↑	All ↑
Intermediate Token Ablation												
w/o Track Token	20.8	3.25	11.9	22.8	7.05	148.5	52.5	44.8	72.5	53.2	44.5	47.8
w/o Motion Token	22.8	**4.18**	12.9	26.5	7.48	290.1	68.8	45.5	82.8	55.1	45.8	49.1
w/o Planning Token	21.5	3.65	12.1	24.5	6.55	348.5	72.3	47.5	**86.3**	56.1	48.9	51.2
Token Mixer Design Ablation												
w/o Instance Mix	21.2	3.68	11.6	26.2	6.85	248.2	61.8	46.6	76.8	**56.6**	46.2	48.5
w/o Instance Block	22.4	3.88	12.4	26.6	7.28	285.6	65.5	46.9	83.5	54.8	47.8	50.4
Full Model (UMA-UniAD)	**23.4**	4.15	**13.1**	**26.9**	**7.62**	**351.2**	**72.5**	**47.7**	86.1	56.4	**49.6**	**52.0**

TABLE V
ABLATION ON CONTROL SIGNALS AND COT.

Method	Nu-X					NuScenes-QA		
	C ↑	B4 ↑	M ↑	R ↑	G ↑	H0 ↑	H1 ↑	All ↑
w/o Control Signals	22.9	4.08	12.8	26.6	7.48	56.0	48.5	51.2
w/o CoT Reasoning	22.5	3.95	12.7	26.4	7.35	55.8	49.2	51.0
Full Model (UMA-UniAD)	**23.4**	**4.15**	**13.1**	**26.9**	**7.62**	**56.4**	**49.6**	**52.0**

Table II compares the physical grounding performance on Nu-X. UMA-UniAD significantly outperforms the reproduced DriveGPT4 baseline, notably reducing steering angle RMSE by 30% (3.85° vs. 5.52°). This confirms that the internalized *Planning Tokens* serve as a strong inductive bias, enabling more precise vehicle dynamics prediction, in contrast to pure vision-based baselines.

D. Ablation Studies

Dual-Branch Video-Enhanced BEV Former. As shown in Tab. III, we investigate the impact of the proposed video enhancement module. Compared to the baseline without video enhancement, incorporating video features yields significant improvements across all metrics, validating that temporal visual information effectively compensates for the loss of fine-grained semantics in BEV representations. Note that we do not ablate BEV tokens. Since the AD backbone relies on them for planning, retaining them is necessary to ensure explanations remain faithful to the decision-making process, rather than merely captioning raw video.

Specifically, the model with only the Detail Branch (*w/o Global Branch*) generally outperforms the model with only the Global Branch (*w/o Detail Branch*), particularly in dense captioning tasks (TOD3Cap). This suggests that accurate driving explanations rely heavily on identifying specific visual cues (e.g., traffic lights, turn signals) rather than just global context. Ultimately, the Full Model achieves the best overall performance, demonstrating that the Detail and Global branches capture complementary information to generate comprehensive and logically coherent explanations.

Intermediate Tokens and Mixer Design. We conduct a comprehensive ablation to evaluate the holistic alignment designs, as summarized in Tab. IV. Regarding intermediate tokens, removing Track Tokens significantly impairs dense captioning (TOD3Cap) due to the loss of essential positional and identity information. Planning Tokens provide future trajectory data, proving critical for intent interpretation in Nu-X. Similarly, Motion Tokens are essential for dynamic perception, evidenced by performance declines in NuScenes-QA. Regarding the mixer architecture, the Instance Mixer and Instance Blocks function to enhance feature extraction and adaptation. Removing the mixer leads to a substantial performance drop across tasks, indicating that positional and motion information is not adequately fused. The absence of the Instance Block further degrades high-level semantic alignment, confirming its necessity for consolidating unified representations.

Control Signals and CoT Reasoning. Table V systematically evaluates the contributions of explicit control supervi-

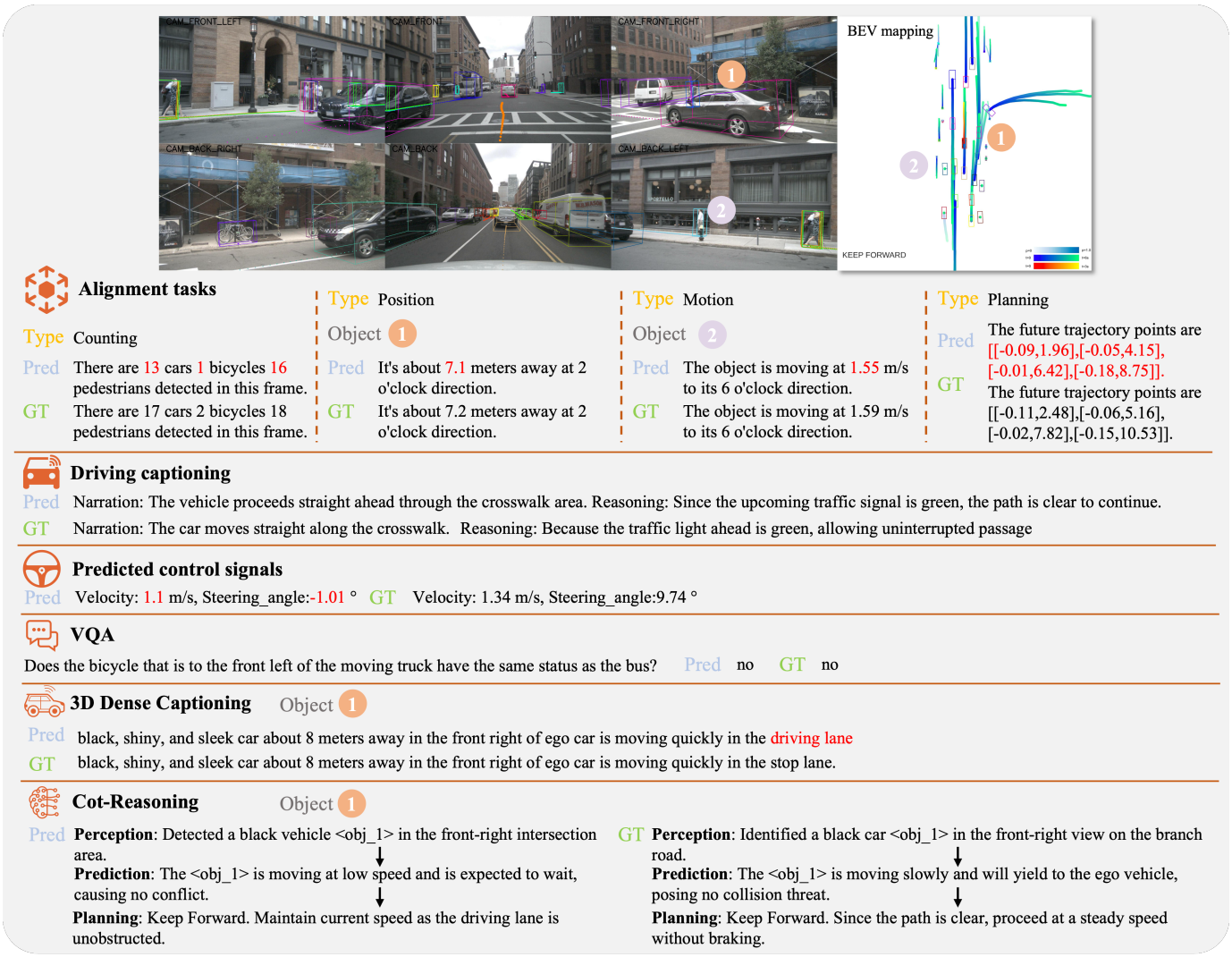


Fig. 3. Qualitative Results. We present examples of the language output generated by UMA-AD across multiple tasks, including driving explanation, 3D dense captioning, VQA, command prediction, CoT-Reasoning, and four categories of alignment tasks. Captions that do not match the ground truth are colored in red.

sion and CoT reasoning. Removing control signals leads to drops in motion-sensitive metrics (e.g., NuScenes-QA H1), indicating that predicting kinematics serves as a vital anchor for recognizing vehicle intentions. Regarding CoT, removing this supervision results in marked declines in logic-oriented metrics (e.g., Nu-X G-Score). This performance gap suggests that without explicit step-by-step guidance, the model struggles to fully internalize the "Perception-Prediction-Planning" sequential logic, which is essential for generating structured and logically consistent explanations.

E. Qualitative Analysis

As illustrated in Fig. 3, our model demonstrates a holistic understanding of complex driving scenarios by seamlessly integrating multi-modal perception with linguistic reasoning. In the Alignment tasks, the model achieves precise numerical predictions for object position (e.g., estimating distance at 7.1m vs. GT 7.2m), validating the efficacy of our Unified Token Aligner in grounding abstract visual features to physical entities. Furthermore, the CoT-Reasoning and Driving Captioning results highlight the model's capability to interpret decision rationale; distinct from shallow description, the model explicitly identifies the potential risk (e.g., the 'black car') and formulates a safe planning decision ('yield to the ego vehicle') based on the predicted motion. This confirms that our framework not only detects atomic details but also provides superior interpretability, offering transparent insights into the decision-making logic of end-to-end driving.

V. Conclusion

We presented UMA-AD, a unified multimodal alignment framework for interpretable autonomous driving. We designed a Unified Token Aligner with a Dual-Branch Video-Enhanced BEV Former to enrich heterogeneous intermediate outputs with fine-grained video semantics. Furthermore, our structured

CoT supervision empowers the model to generate explicit reasoning chains, while auxiliary control signals ensure physical grounding. We believe this work serves as a robust baseline for bridging the critical gap between black-box driving models and human-centric trust. Future work will focus on optimizing inference efficiency for real-time edge deployment and exploring lightweight model alternatives.

REFERENCES

[1] D. A. Pomerleau, "ALVINN: An autonomous land vehicle in a neural network," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, 1988, pp. 305–313.

[2] F. Yu *et al.*, "BDD100K: A diverse driving dataset for heterogeneous multitask learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 2636–2645.

[3] E. Zablocki, H. Ben-Younes, P. Pérez, and M. Cord, "Explainability of deep vision-based autonomous driving systems: Review and challenges," *Int. J. Comput. Vis.*, vol. 130, no. 10, pp. 2425–2452, 2022.

[4] J. Kim, A. Rohrbach, T. Darrell, J. Canny, and Z. Akata, "Textual explanations for self-driving vehicles," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 564–579.

[5] B. Jin *et al.*, "ADAPT: Action-aware driving caption transformer," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, 2023, pp. 7554–7561.

[6] Z. Xu *et al.*, "DriveGPT4: Interpretable end-to-end autonomous driving via large language model," *IEEE Robot. Autom. Lett.*, vol. 9, no. 10, pp. 8186–8193, 2024.

[7] S. Malla, C. Choi, I. Gur, J. Thursby, and A. Roy, "Drama: Joint risk localization and captioning in driving," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, 2023, pp. 1043–1052.

[8] K. Ding *et al.*, "Hint-AD: Holistically aligned interpretability in end-to-end autonomous driving," *arXiv preprint arXiv:2409.06702*, 2024.

[9] Z. Li *et al.*, "BEVFormer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2022, pp. 1–18.

[10] X. Tian *et al.*, "DriveVLM: The convergence of autonomous driving and large vision-language models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2024, pp. 14458–14469.

[11] Y. Wu, L. Wu, Z. Zheng, F. Shi, W. Jin, and J. Lu, "HeightFormer: Explicit height modeling without extra data for camera-only 3D object detection in bird's eye view," *arXiv preprint arXiv:2307.13510*, 2023.

[12] C. Sima *et al.*, "DriveLM: Driving with graph visual question answering," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2024.

[13] W. Wang *et al.*, "DriveMLM: Aligning multi-modal large language models with behavioral planning states for autonomous driving," *arXiv preprint arXiv:2312.09245*, 2023.

[14] B. Jin *et al.*, "TOD3Cap: Towards 3D dense captioning in outdoor scenes," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2024.

[15] T. Qian, J. Chen, L. Zhuo, Y. Jiao, and Y.-G. Jiang, "NuScenes-QA: A multi-modal visual question answering benchmark for autonomous driving scenario," in *Proc. AAAI Conf. Artif. Intell.*, 2024.

[16] J. Kim and J. Canny, "Interpretable learning for self-driving cars by visualizing causal attention," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2017, pp. 2942–2950.

[17] M. Bojarski *et al.*, "VisualBackProp: Efficient visualization of CNNs for autonomous driving," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, 2018, pp. 4701–4708.

[18] T. Deruyttere, P. Vandenabeele, K. Boutens, T. Brox, and M.-F. Moens, "Talk2car: Taking control of your self-driving car," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2019, pp. 2088–2098.

[19] OpenAI, "GPT-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[20] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2024.

[21] J. Bai *et al.*, "Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond," *arXiv preprint arXiv:2308.12966*, 2023.

[22] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2023, pp. 19730–19742.

[23] H. Zhang, X. Li, and L. Bing, "Video-LLaMA: An instruction-tuned audio-visual language model for video understanding," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2023, pp. 543–553.

[24] H. Shao *et al.*, "LMDrive: Closed-loop end-to-end driving with large language models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2024, pp. 15120–15130.

[25] T. Choudhary *et al.*, "Talk2BEV: Language-enhanced bird's-eye view maps for autonomous driving," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, 2024, pp. 16345–16352.

[26] B. Liao *et al.*, "MapTR: Structured modeling and learning for online vectorised HD map construction," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2023.

[27] Y. Ma *et al.*, "Dolphins: Multimodal language model for driving," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2024.

[28] L. Wen *et al.*, "DiLu: A knowledge-driven approach to autonomous driving with large language models," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2024.

[29] J. Wei *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 35, 2022, pp. 24824–24837.

[30] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa, "Large language models are zero-shot reasoners," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 35, 2022, pp. 22199–22213.

[31] X. Wang *et al.*, "Self-consistency improves chain of thought reasoning in language models," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2023.

[32] A. Brohan *et al.*, "Rt-2: Vision-language-action models transfer web knowledge to robotic control," *arXiv preprint arXiv:2307.15818*, 2023.

[33] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 770–778.

[34] H. Caesar *et al.*, "nuScenes: A multimodal dataset for autonomous driving," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 11621–11631.

[35] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: a method for automatic evaluation of machine translation," in *Proc. Annu. Meeting Assoc. Comput. Linguistics (ACL)*, 2002, pp. 311–318.

[36] S. Banerjee and A. Lavie, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," in *Proc. ACL Workshop*, 2005, pp. 65–72.

[37] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*, 2004, pp. 74–81.

[38] R. Vedantam, C. L. Zitnick, and D. Parikh, "CIDEr: Consensus-based image description evaluation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2015, pp. 4566–4575.

[39] Y. Hu *et al.*, "Planning-oriented autonomous driving," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023, pp. 17853–17862.

[40] B. Jiang *et al.*, "VAD: Vectorized scene representation for efficient autonomous driving," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2023, pp. 8340–8350.