

UniDriveVLA: Unifying Understanding, Perception, and Action Planning for Autonomous Driving

Yongkang Li, Yu Zhu, Lijun Zhou, Sixu Yan, Bencheng Liao, Tianyi Yan, Kaixin Xiong, Zhenxin Zhu, Hongwei Xie, Haiyang Sun, Bing Wang, Guang Chen, Wenyu Liu *Senior Member, IEEE*, Hangjun Ye, and Xinggang Wang *Senior Member, IEEE*

Abstract—Vision-Language-Action (VLA) models have emerged in autonomous driving, promising to improve the cognitive capabilities of driving systems with rich world knowledge. However, adapting pretrained Vision-Language Models (VLMs) for driving presents an optimization dilemma: strengthening spatial perception can compromise the semantic reasoning inherited from pretraining. We argue that this difficulty arises from two sources: interference among understanding, perception, and action objectives under shared transformations, and driving-domain adaptation that alters pretrained transformations supporting semantic reasoning. To address them, we propose UniDriveVLA, a Unified Driving Vision-Language-Action model that unifies understanding, perception, and action planning for autonomous driving. We build UniDriveVLA on a Mixture-of-Transformers (MoT) architecture with dedicated experts for task-specific transformations and introduce Masked Joint Attention to control their information exchange. We further propose reasoning-preserving spectral regularization (ReSPECT) to discourage low-rank updates from concentrating along the dominant singular subspaces of pretrained weights. Within this framework, sparse perception extracts structured spatial states, which the Action Expert combines with semantic context to predict continuous trajectories. Extensive experiments demonstrate strong performance across understanding, spatial perception, and action planning. Closed-loop evaluations on NeuroNCAP and Fail2Drive further demonstrate safety-critical robustness and long-tail generalization, respectively. Finally, UniDriveVLA-Flash enables real-time end-to-end inference at 10.4 Hz.

Index Terms—Autonomous driving, Vision-Language-Action models, spatial perception, semantic reasoning.

I. INTRODUCTION

Autonomous driving has made significant progress in perception [1], [2], [3], prediction [4], [5], and planning [6], [7]. Against this backdrop, Vision-Language-Action (VLA) models [8], [9], [10], [11], [12] have emerged as a promising direction, leveraging rich world knowledge and reasoning capabilities to enhance the cognitive capabilities of driving systems. Most driving VLAs, however, are built upon 2D Vision-Language Models (VLMs) [13], [14], [15] pretrained on internet-scale image-text data. Although these models provide strong semantic understanding, they are not designed for explicit spatial perception, limiting their ability to capture

the geometric structure required for driving. Recent works [12], [16], [17], [18], [19], [20] therefore augment VLA systems with 3D spatial representations, either by aligning structured 2D/3D features with language representations or by introducing spatial tokens into shared decoders. Yet a trade-off remains, as illustrated in Fig. 1: directly adopting 2D VLMs preserves native semantic reasoning but yields limited spatial perception, whereas 3D enhancement improves spatial perception but can compromise pretrained reasoning capabilities.

This trade-off arises not merely from introducing 3D spatial representations, but from how heterogeneous driving capabilities are jointly optimized during VLM adaptation. Existing methods introduce spatial information either by aligning structured 2D/3D features with language representations [18], [19] or by incorporating spatial tokens into a shared decoder [16], [17], [21]. Despite these different designs, understanding, perception, and action remain largely coupled through shared transformations. This coupling can induce interference among heterogeneous objectives during optimization and weaken the semantic reasoning inherited from the pretrained VLM. Accordingly, a unified driving VLA requires task-specific transformations, controlled information exchange, and regularization to preserve pretrained semantic capabilities during driving adaptation.

To address these challenges, we propose **UniDriveVLA**, a Vision-Language-Action model that unifies understanding, perception, and action planning for autonomous driving. To reduce interference among heterogeneous objectives, UniDriveVLA adopts a Mixture-of-Transformers (MoT) architecture with dedicated experts for understanding, perception, and action planning. Masked Joint Attention regulates information exchange across these experts, allowing semantic, spatial, and action representations to interact without reverting to shared task transformations. Beyond task specialization, we introduce reasoning-preserving spectral regularization (ReSPECT) to preserve pretrained semantic capabilities by penalizing low-rank updates that overlap with the dominant subspaces of pretrained transformations. To provide driving-specific spatial awareness, the Perception Expert extracts task-relevant geometry from multi-scale VLM features through sparse perception. The Action Expert then integrates semantic context from Understanding with spatial evidence from Perception to support continuous trajectory planning. For real-time end-to-end inference, we further develop UniDriveVLA-Flash, an efficient variant of the unified framework.

Extensive experiments demonstrate strong planning per-

Yongkang Li, Yu Zhu, Sixu Yan, Bencheng Liao, Wenyu Liu, and Xinggang Wang are with Huazhong University of Science and Technology.

Lijun Zhou, Tianyi Yan, Kaixin Xiong, Zhenxin Zhu, Hongwei Xie, Bing Wang, Guang Chen, Hangjun Ye, and Haiyang Sun are with Xiaomi EV.

Tianyi Yan is also with SKL-IOTSC, University of Macau.

Corresponding author: Xinggang Wang xgwang@hust.edu.cn.

Project page: <https://xiaomi-research.github.io/unidrivevla/>.

Code: <https://github.com/xiaomi-research/unidrivevla>.

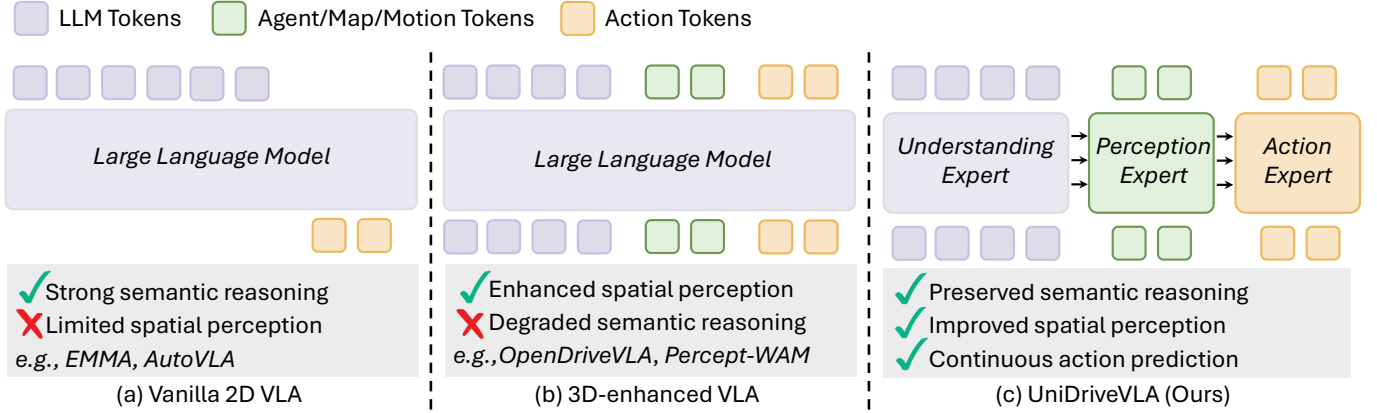


Fig. 1: **Comparison of VLA paradigms for autonomous driving.** (a) Vanilla 2D VLA provides strong semantic reasoning but limited spatial perception. (b) 3D-enhanced VLA improves spatial perception but may degrade semantic reasoning. (c) UniDriveVLA decouples understanding, perception, and action with the Mixture-of-Transformers architecture, achieving both.

formance of UniDriveVLA in both open-loop and closed-loop evaluations. Beyond planning, UniDriveVLA achieves competitive spatial perception and strong driving-oriented understanding while retaining general multimodal capabilities after driving adaptation. Further evaluations on NeuroNCAP and Fail2Drive demonstrate safety-critical robustness and generalization under long-tail scenario shifts, respectively.

The main contributions are summarized as follows:

- We propose UniDriveVLA, a unified Vision-Language-Action model that integrates understanding, perception, and action planning through a Mixture-of-Transformers architecture, with dedicated experts handling task-specific transformations and Masked Joint Attention enabling controlled cross-expert interaction.
- We introduce ReSPECT, a reasoning-preserving spectral regularization method that penalizes low-rank adaptation updates overlapping with the dominant subspaces of pretrained transformations to preserve pretrained semantic capabilities during driving adaptation.
- Extensive experiments show that UniDriveVLA retains semantic understanding while achieving competitive spatial perception. The complete model achieves leading planning results across several matched open-loop and closed-loop evaluations. Further evaluations on NeuroNCAP and Fail2Drive demonstrate robust closed-loop behavior under safety-critical interactions and strong generalization to long-tail scenario shifts, respectively.

II. RELATED WORK

A. Vision-Language-Action Models for Autonomous Driving

Recent studies [8], [9], [11], [12], [18], [19], [22], [23] have introduced Vision-Language Models (VLMs) into autonomous driving to leverage their pretrained world knowledge and reasoning capabilities for handling semantically complex and long-tail driving scenarios. Dual-system approaches, such as DriveVLM [24] and Senna [25], adopt a fast-slow paradigm that combines VLMs with end-to-end (E2E) driving models [6], [7], [26], where the VLM predicts low-frequency trajectories or provides high-level instructions to guide the E2E model. In contrast to these dual-system frameworks, recent methods have

shifted toward unified single-system architectures [9], [10], [11], [12], [18], [27], [28], [29], [30], enabling a single model to perform trajectory planning directly. To meet the inference-rate demands of autonomous driving, recent methods pair VLMs with dedicated action decoders [9], [12], [23], [30], [31], [32] for efficient continuous trajectory prediction, while others explore Diffusion Large Language Models (DLLMs) [33], [34], [35], [36] as alternatives to autoregressive reasoning. Recent works [8], [9], [10], [37], [38], [39], [40], [41], [42] have also explored reinforcement learning (RL) for fine-tuning VLA models to further improve driving performance and policy alignment.

Despite these advances, reliable planning requires not only semantic reasoning but also accurate spatial perception of the driving scene. Many existing VLA approaches [8], [10], [11] build on pretrained VLMs through task-specific fine-tuning, yet remain limited in driving-specific spatial perception. To improve spatial perception in VLA models, some approaches incorporate bird’s-eye-view (BEV) encoders [18] or 3D Q-Formers [12], [19], [43] to extract 3D features that are subsequently fed into large language models (LLMs). Alternatively, other methods [16], [17], [21], [44] enhance LLMs with spatial tokens or auxiliary occupancy supervision and jointly optimize language and spatial-perception objectives. Although these strategies enrich VLA models with 3D spatial information, optimizing spatial perception and semantic reasoning through shared parameters may create interference between the two objectives, potentially compromising the semantic capabilities inherited from pretraining. In contrast, our work focuses on reconciling spatial perception and semantic reasoning within a unified driving VLA framework.

B. Mixture-of-Transformers

The Mixture-of-Transformers (MoT) [45] paradigm was introduced as a sparse multimodal architecture that assigns modality-specific Transformer parameters while preserving cross-modal interaction through global self-attention. Subsequent works [46], [47], [48] extended this design to multimodal understanding and generation, using modality-specific parameters to preserve pretrained language capabilities and global

self-attention to transfer language knowledge into the generative pathway, thereby improving generation performance. Around the same time, π_0 [49] introduced a related mixture-based Transformer architecture for robotics, coupling a pretrained VLM with a flow-matching Action Expert for continuous action generation. Leveraging MoT's combination of expert specialization and cross-modal interaction, subsequent works have adopted similar architectures in VLA models for robotics [50], [51], [52], [53], [54], [55] and autonomous driving [8], [23], [29], [56], [57], [58]. However, several driving VLAs [8], [23], [29] adopt a π_0 -style design, while their spatial perception remains limited relative to the demands of reliable planning. Our work instead augments MoT with a sparse perception expert, preserving semantic reasoning while strengthening the spatial perception required for reliable planning.

C. Sparse Perception for Autonomous Driving

Traditional autonomous driving perception heavily relies on dense 3D representations [59] or BEV grid constructions [1], [2], [60], which introduce computational redundancy. To circumvent this limitation, sparse perception models have emerged as an efficient alternative. These approaches adopt a query-driven paradigm [61], using a set of sparse 3D queries to aggregate multi-view image features. Pioneering works such as DETR3D [62] and PETR [63] established sparse object-query formulations for 3D detection, bypassing dense view transformation. Subsequent advancements, including Sparse4D [64] and SparseBEV [65], further bolstered spatial perception by integrating temporal modeling and adaptive spatio-temporal sampling mechanisms. More recently, this sparse paradigm has been extended to end-to-end autonomous driving systems. Methods such as SparseDrive [26] and SparseAD [66] represent dynamic driving scenes using unified sparse queries and integrate detection, tracking, and planning within a single framework. Nevertheless, these sparse driving systems do not integrate their spatial representations with pretrained semantic reasoning within a unified VLA framework.

III. METHODOLOGY

The overall architecture of UniDriveVLA is illustrated in Fig. 2. In Section III-A, we first define the problem of jointly modeling semantic understanding, spatial perception, and action planning. Section III-B then presents expert specialization with controlled interaction through the MoT architecture and Masked Joint Attention, followed by ReSPECT in Section III-C for regulating adaptation of the Understanding Expert. Section III-D further describes sparse spatial perception and continuous action generation. Finally, Section III-E presents UniDriveVLA-Flash, which combines architectural and runtime optimizations for real-time end-to-end inference, as illustrated in Fig. 3.

A. Problem Formulation

Given multi-view camera observations I_{cam} , the historical ego trajectory I_{hist} , task instruction I_{inst} , and navigation command I_{nav} , UniDriveVLA jointly models semantic understanding, spatial perception, and action planning for safe and reliable driving. We denote the semantic output as Y_{und} , the structured spatial output as Y_{per} , and the future ego trajectory as

$T_{\text{traj}} = \{(x_k, y_k)\}_{k=1}^K$, where K denotes the planning horizon. A unified driving VLA can therefore be formulated as

$$(Y_{\text{und}}, Y_{\text{per}}, T_{\text{traj}}) = \Phi_{\Theta}(I_{\text{cam}}, I_{\text{hist}}, I_{\text{inst}}, I_{\text{nav}}), \quad (1)$$

where Φ_{Θ} denotes the unified model parameterized by Θ .

The three capabilities are learned from heterogeneous supervision. Semantic understanding is optimized by $\mathcal{L}_{\text{und}}$, spatial perception by $\mathcal{L}_{\text{per}}$, and action planning by $\mathcal{L}_{\text{act}}$. Their joint objective is

$$\mathcal{L}_{\text{joint}} = \lambda_{\text{und}}\mathcal{L}_{\text{und}} + \lambda_{\text{per}}\mathcal{L}_{\text{per}} + \lambda_{\text{act}}\mathcal{L}_{\text{act}}, \quad (2)$$

where λ_{und} , λ_{per} , and λ_{act} balance the three objectives.

Two potential sources of interference arise during training. First, heterogeneous objectives can produce conflicting updates when they jointly optimize shared transformations. Second, even after task-specific transformations are separated, driving-domain updates within the Understanding Expert can perturb the pretrained transformations that support semantic reasoning. The former motivates expert specialization, whereas the latter motivates ReSPECT.

B. Expert Specialization with Controlled Interaction

Semantic understanding, spatial perception, and action planning require different representations and supervision, while reliable driving depends on their cooperation. As shown in Fig. 2, UniDriveVLA builds upon the Mixture-of-Transformers (MoT) paradigm [45] and assigns each capability a dedicated Transformer expert. Let $\mathbf{H}_l^g$ denote the hidden states at layer l for $g \in \mathcal{G} = \{\text{U}, \text{P}, \text{A}\}$, corresponding to Understanding, Perception, and Action. Each expert applies its own attention transformations:

$$\begin{aligned} \bar{\mathbf{H}}_l^g &= \text{LN}_{l,\text{in}}^g(\mathbf{H}_l^g), \\ [\mathbf{Q}_l^g \mathbf{K}_l^g \mathbf{V}_l^g] &= \bar{\mathbf{H}}_l^g \begin{bmatrix} \mathbf{W}_{Q,l}^g & \mathbf{W}_{K,l}^g & \mathbf{W}_{V,l}^g \end{bmatrix}. \end{aligned} \quad (3)$$

Head reshaping, query-key normalization, and rotary position encoding are omitted for clarity. The output projection and feed-forward network are likewise expert-specific. This parameterization removes direct sharing of Transformer transformations among the three capabilities but does not specify which streams may exchange information. UniDriveVLA therefore implements Masked Joint Attention by imposing a directed visibility mask on MoT joint attention for the Understanding, Perception, and Action streams. For $\mathbf{R} \in \{\mathbf{Q}, \mathbf{K}, \mathbf{V}\}$, the expert-specific representations are concatenated along the token dimension, after which joint attention is computed as

$$\begin{aligned} \mathbf{R}_l &= [\mathbf{R}_l^{\text{U}}; \mathbf{R}_l^{\text{P}}; \mathbf{R}_l^{\text{A}}], \quad \mathbf{R} \in \{\mathbf{Q}, \mathbf{K}, \mathbf{V}\}, \\ \mathbf{Z}_l &= \text{Softmax}\left(\frac{\mathbf{Q}_l \mathbf{K}_l^{\top}}{\sqrt{d_k}} + \mathbf{M}_l\right) \mathbf{V}_l. \end{aligned} \quad (4)$$

Following the order of Understanding, Perception, and Action, the expert-level visibility relation is

$$\mathcal{V} = \begin{bmatrix} \text{U} \leftarrow \text{U} & \times & \times \\ \text{P} \leftarrow \text{U} & \text{P} \leftarrow \text{P} & \times \\ \text{A} \leftarrow \text{U} & \text{A} \leftarrow \text{P} & \text{A} \leftarrow \text{A} \end{bmatrix}. \quad (5)$$

Rows identify query experts and columns identify key-value experts. The notation $g \leftarrow h$ indicates that expert g can attend to

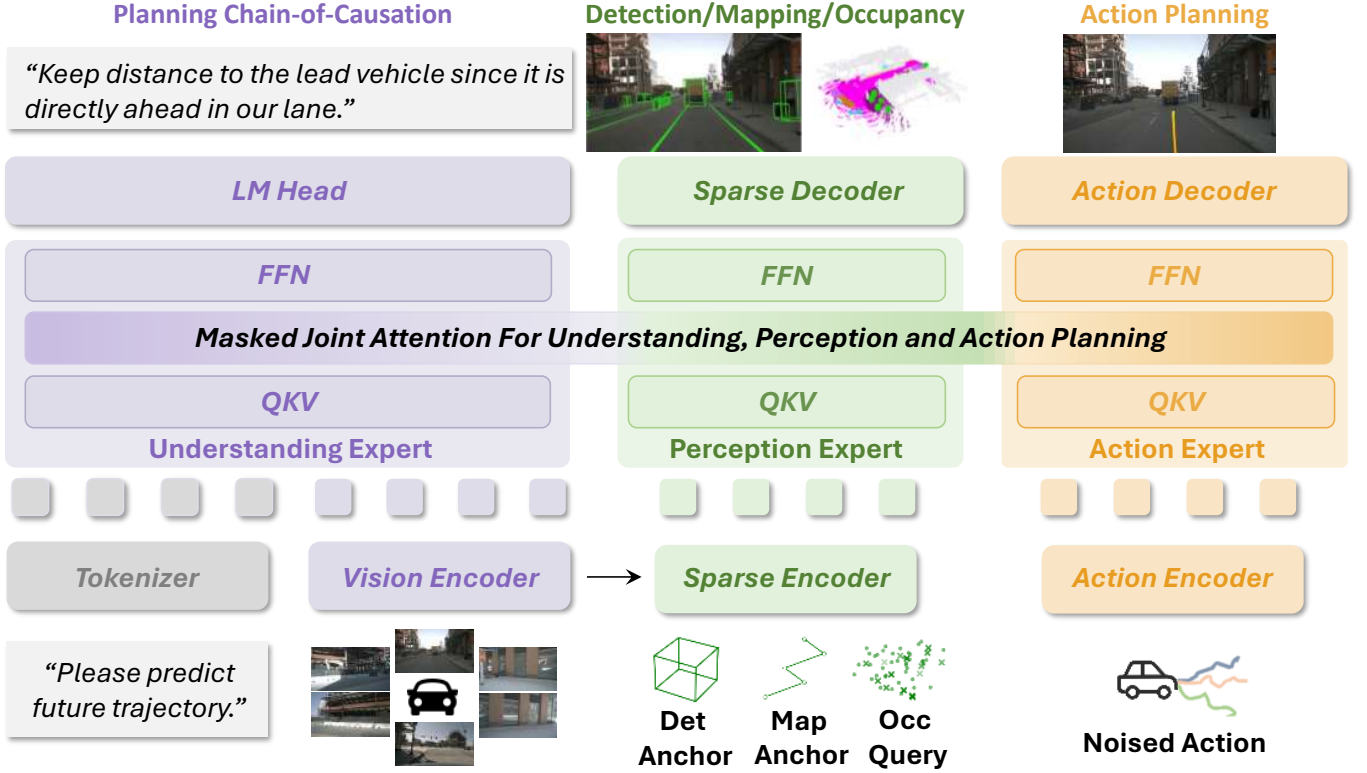


Fig. 2: **Architecture overview of UniDriveVLA.** UniDriveVLA realizes expert specialization with controlled interaction by assigning understanding, perception, and action to dedicated Transformer experts and coordinating their token streams through Masked Joint Attention. The three pathways produce a Chain-of-Causation, structured perception outputs, and a future trajectory.

expert h , whereas $\times$ denotes blocked interaction. The additive mask M_l in Eq. (4) instantiates this relation at the token level: blocked entries are assigned $-\infty$, while permitted entries retain the corresponding causal and padding constraints. The Understanding Expert first generates the Chain-of-Causation reasoning sequence without accessing subsequent Perception or Action states. Its hidden states then condition Perception, while Action integrates both semantic and spatial evidence.

After joint attention, Z_l is partitioned at the original stream boundaries as $Z_l = [Z_l^U; Z_l^P; Z_l^A]$. Each segment is then updated by the output projection, residual pathway, normalization, and feed-forward network of its corresponding expert. This design disentangles parameter specialization from information access: each capability retains its own transformations, while information follows the directed hierarchy $U \rightarrow P$ and $\{U, P\} \rightarrow A$. It reduces direct competition among heterogeneous objectives.

C. ReSPECT: Reasoning-Preserving Spectral Regularization

Expert specialization mitigates competition across task pathways, but the transformations of the pretrained VLM are still altered by driving-specific updates. ReSPECT retains the standard low-rank parameterization of LoRA [67] and introduces a normalized bilateral spectral regularizer that discourages update energy from concentrating in the dominant input and output singular subspaces of the pretrained

transformation [68], [69]. For a frozen pretrained linear weight W_0 , the adapted transformation is

$$W = W_0 + \Delta W, \quad \Delta W = \frac{\alpha}{r} BA, \quad (6)$$

where r and α denote the LoRA rank and scaling factor, respectively, and only the low-rank factors A and B are optimized.

To measure the components of ΔW associated with the dominant singular directions of the pretrained transformation, we decompose $W_0 = U \Sigma V^T$ and retain its leading k left and right singular vectors U_k and V_k . These vectors span the dominant output and input subspaces of W_0 , respectively. The corresponding update components are

$$\begin{aligned} \Delta W V_k &= W V_k - W_0 V_k, \\ U_k^T \Delta W &= U_k^T W - U_k^T W_0. \end{aligned} \quad (7)$$

The term $\Delta W V_k$ measures the update response to the dominant input directions, whereas $U_k^T \Delta W$ measures the update projected onto the dominant output subspace. Their squared Frobenius norms quantify the update energy associated with these two pretrained subspaces.

For each adapted layer l , ReSPECT penalizes these subspace-associated energies relative to the total update energy:

$$\mathcal{R}_l^{\text{ReSPECT}} = \frac{\|\Delta W_l V_{k,l}\|_F^2 + \|U_{k,l}^T \Delta W_l\|_F^2}{\|\Delta W_l\|_F^2 + \varepsilon}, \quad (8)$$

where $\varepsilon > 0$ is a small constant for numerical stability. The denominator normalizes the sum of the two subspace-associated

energies by the total update energy. Apart from ε , the ratio is invariant to rescaling $\Delta \mathbf{W}_l$. ReSPECT therefore penalizes the relative concentration of update energy in the dominant pretrained subspaces rather than the overall update magnitude.

The layer-wise penalties are incorporated into the unified training objective as

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{joint}} + \sum_{m \in \mathcal{P}} \beta_m \sum_{l \in \mathcal{S}_m} \mathcal{R}_l^{\text{ReSPECT}}, \quad (9)$$

where $\mathcal{P}$ denotes the selected pretrained components, $\mathcal{S}_m$ contains their adapted linear layers, and $\beta_m \geq 0$ controls the corresponding regularization strength. The singular subspaces are computed once from the frozen pretrained weights and remain fixed throughout adaptation. Unlike norm regularization, which penalizes update magnitude, ReSPECT discourages low-rank updates from concentrating their energy in the dominant input and output singular subspaces. It is designed to retain pretrained semantic capabilities without removing the capacity for driving-specific learning. The regulated transformations produce Understanding states that subsequently provide semantic context for the Perception and Action Experts.

D. Sparse Spatial Perception and Action Planning

To avoid the representation mismatch caused by directly integrating explicit 3D features into a pretrained VLM [18], [19], UniDriveVLA constructs query-based sparse states from multi-scale 2D VLM features within the Perception Expert [26], [70]. Through Masked Joint Attention, the Understanding states provide semantic context to the sparse Perception states, which the Action Expert then uses to condition trajectory generation.

Let $\mathcal{G}_{\text{sp}} = \{\text{det}, \text{map}, \text{ego}, \text{mot}\}$ denote the sparse perception tasks. For $g \in \mathcal{G}_{\text{sp}}$, let $\mathbf{q}_{g,i}$ denote the i -th task-specific query with geometric reference $\mathbf{a}_{g,i}$. Its spatial representation is

$$\mathbf{s}_{g,i} = \bar{\mathbf{q}}_{g,i} + \mathcal{A}_g(\mathcal{F}; \Pi(\mathbf{a}_{g,i})), \quad (10)$$

where Π projects each geometric reference onto the image planes and $\mathcal{A}_g$ aggregates the corresponding evidence from the multi-scale VLM features $\mathcal{F}$. Before feature aggregation, sparse queries interact with their temporal states and with queries from other perception tasks, yielding $\bar{\mathbf{q}}_{g,i}$. Occupancy is excluded from $\mathcal{G}_{\text{sp}}$ because its latent queries do not carry explicit geometric references. Between refinement stages, the resulting sparse states and occupancy states jointly form $\mathbf{H}^{\text{P}}$ and are contextualized by $\mathbf{H}^{\text{U}}$ through Masked Joint Attention. The contextualized sparse states retain their geometric references and continue through task-specific refinement to produce structured predictions without constructing a dense BEV grid, while the occupancy states are decoded using the pretrained Occ-VAE from DOME [71].

For action planning, the Action Expert integrates semantic context from the Understanding states $\mathbf{H}^{\text{U}}$ and spatial evidence from the Perception states $\mathbf{H}^{\text{P}}$ into the action stream through Masked Joint Attention, producing action states $\mathbf{H}^{\text{A}}$ for trajectory generation. The action stream contains noised trajectory tokens obtained by interpolating between normalized future ego displacements and Gaussian noise under standard conditional flow matching [72]. The action stream also contains a status

token encoding navigation and ego status and an ego-history token encoding recent motion.

Following [73], we additionally supervise the reconstructed ego trajectory in metric coordinates:

$$\mathcal{L}_{\text{act}} = \mathcal{L}_{\text{FM}} + \lambda_{\text{wp}} \left\| \hat{T}_{\text{traj}} - T_{\text{traj}} \right\|_2^2. \quad (11)$$

Here, $\mathcal{L}_{\text{FM}}$ applies Min-SNR weighting to the velocity-matching objective, balancing supervision across noise levels. The coefficient λ_{wp} weights waypoint supervision. The metric trajectory $\hat{T}_{\text{traj}}$ is reconstructed by recovering clean displacements from the predicted flow, then denormalizing and accumulating them.

The capability-level objective in Eq. (2) is instantiated with Chain-of-Causation (CoC) supervision [8] for Understanding, structured detection, mapping, ego-status, motion, and occupancy supervision for Perception, and Eq. (11) for Action. These objectives supervise each expert in its native output space, while Masked Joint Attention governs information exchange among them.

E. UniDriveVLA-Flash: Efficient End-to-End Inference

UniDriveVLA incurs inference cost from visual-prefix attention and iterative action sampling. UniDriveVLA-Flash preserves expert specialization and reduces costs through visual token compression, perception-conditioned modulation, and learnable action queries.

UniDriveVLA-Flash compresses visual tokens for the Understanding Expert with VisionZip [74], while retaining pre-merger features for the Perception Expert. For perception-conditioned adaptive normalization (AdaNorm), the first-stage Perception states $\mathbf{H}_{\text{pre}}^{\text{P}}$ are pooled by group, concatenated, and projected into γ_{P} and β_{P} :

$$\tilde{\mathbf{H}}_0^{\text{A}} = \mathbf{H}_0^{\text{A}} + \gamma_{\text{P}} \odot \text{LN}(\mathbf{H}_0^{\text{A}}) + \beta_{\text{P}}. \quad (12)$$

The broadcast parameters modulate the initial Action tokens $\mathbf{H}_0^{\text{A}}$, transferring spatial evidence without token-level Perception-to-Action attention.

Learnable action queries replace the noised trajectory tokens used in flow matching [72], eliminating iterative ODE integration at inference. The Action Expert $\mathcal{F}_{\text{A}}$ conditions the modulated Action tokens $\tilde{\mathbf{H}}_0^{\text{A}}$ on the Understanding states $\mathbf{H}^{\text{U}}$, and the action decoder D_{act} predicts normalized displacements in one pass:

$$\mathbf{H}^{\text{A}} = \mathcal{F}_{\text{A}}(\tilde{\mathbf{H}}_0^{\text{A}}; \mathbf{H}^{\text{U}}), \quad \widehat{\Delta \mathbf{T}} = D_{\text{act}}(\mathbf{H}^{\text{A}}). \quad (13)$$

The predicted displacements are denormalized and accumulated into $\hat{T}_{\text{traj}}$. Runtime acceleration further uses TF32, SDPA, `torch.compile`, CUDA Graphs, and TensorRT.

IV. EXPERIMENTS

Our experiments examine whether UniDriveVLA improves spatial perception without compromising semantic reasoning. We then assess whether this balance is reflected in accurate and reliable planning across diverse driving conditions. Controlled ablations investigate how the proposed design choices affect spatial perception, semantic reasoning, and planning.

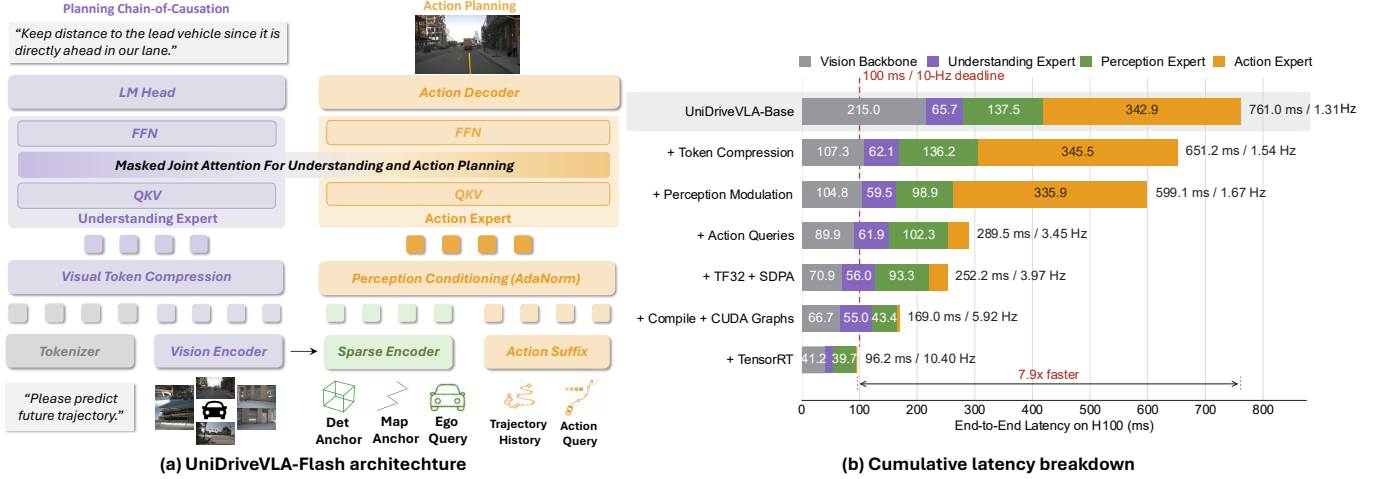


Fig. 3: **Architecture and efficiency of UniDriveVLA-Flash.** (a) A streamlined inference pathway combines visual token compression, perception-conditioned action modulation, and learnable action queries. (b) Cumulative latency profiling shows the latency reductions from successive architectural and runtime optimizations.

Qualitative analyses contrast successful driving behaviors with residual failure modes, clarifying the scope and limitations of the model.

A. Experimental Setup

1) *Implementation Details:* We instantiate 2B UniDriveVLA-Base and 8B UniDriveVLA-Large from Qwen3-VL [13], retaining its vision encoder, projector, and language model as the Understanding Expert and adding Perception and Action Experts. Unless otherwise specified, results use UniDriveVLA-Base. We additionally evaluate UniDriveVLA-Large on nuScenes and understanding benchmarks. For each input view resized to 960×544 , the vision encoder provides 34×60 pre-merger features to the Perception Expert, while the projector produces a 17×30 token grid for the Understanding Expert. The Perception Expert uses 900 detection queries and 100 map queries with K-means anchors, one ego query with a fixed ego anchor, six motion queries per detected agent with K-means trajectory anchors selected by agent class, and 625 randomly initialized learnable occupancy queries. The Action Expert receives noisy trajectory tokens together with a token that combines navigation and ego status. UniDriveVLA predicts six ego waypoints at intervals of 0.5 s over a 3 s planning horizon, while the motion head predicts six trajectory modes over 6 s on nuScenes and 3 s on Bench2Drive.

Following ReCogDrive [9], Stage 1 fine-tunes the Understanding Expert for three epochs at 4×10^{-5} on a 3:7 mixture of driving-specific data and general-domain data from FineVision [84]. Driving-specific CoC annotations are generated using Alpaymayo 1.5 [8]. Stage 2 jointly trains the Perception and Action Experts while adapting the Understanding Expert’s vision encoder, projector, and language-model LoRA parameters. This stage runs for 20 epochs on nuScenes and 10 on Bench2Drive using AdamW [85] with cosine annealing, a 2×10^{-4} base learning rate for the added experts, and a $0.5 \times$ learning-rate multiplier for the trainable Understanding-Expert parameters. Beginning in Stage 2, the Action Expert is optimized with flow matching, where

normalized ego displacements are linearly interpolated with standard Gaussian noise at times sampled from a rescaled Beta(1.5, 1.0) distribution over $[0.001, 1]$. Following Hybrid-Planner [73], the planning loss combines Min-SNR-weighted flow-velocity MSE ($\gamma = 5$) with waypoint MSE weighted by 0.5, restricting gradients through cumulative integration to the three most recent displacement steps. For Understanding-Expert adaptation, UniDriveVLA-Base and UniDriveVLA-Large use LoRA [67] with $(r, \alpha) = (64, 128)$ and $(128, 256)$, respectively, and a spectral rank of $k = 16$. ReSPECT uses $(\beta_{\text{vis}}, \beta_{\text{LM}}) = (0.010, 0.003)$ for Base and $(0.005, 0.003)$ for Large, while the projector retains standard LoRA. Stage 3 freezes the Understanding Expert and fine-tunes the Perception and Action Experts at 1×10^{-4} for 10 epochs on nuScenes and 5 on Bench2Drive, with motion-forecasting supervision enabled for the Perception Expert. Across all three stages, training is distributed across 16 NVIDIA H20 GPUs using DeepSpeed with BF16 precision. Unless otherwise specified, all results are reported using the Stage-2 checkpoint, while Tab. III uses the complete Stage-3 checkpoint.

At inference, ten Euler steps integrate the Action Expert’s vector field, after which the predicted displacements are denormalized and accumulated to recover the ego trajectory. In Bench2Drive, the evaluation agent tracks the predicted trajectory at 20 Hz in synchronous CARLA using separate lateral and longitudinal proportional-integral-derivative (PID) controllers. For NeuroNCAP, the predicted trajectories are executed by the benchmark-provided linear-quadratic regulator (LQR).

2) Evaluation Benchmarks:

Understanding. FidelityDrivingBench assesses scene description (SD), traffic question answering (T-QA), noteworthy-object perception and reasoning (NoPR), and knowledge retention rate (KRR), while LingoQA [94] evaluates driving-oriented visual question answering. MMStar [95], MMMU [96], RealWorldQA (RWQA), AI2D, MME [97], VLMsAreBlind (VAB) [98], and ChartQA [99] measure retained general multimodal capability.

TABLE I: Closed-loop and multi-ability results on Bench2Drive [75]. Eff., Comf., SR, and DS denote Efficiency, Comfort, Success Rate, and Driving Score, respectively. * denotes privileged feature distillation, and $^{\circ}$ marks the vision-only TFv6 setting. Bold indicates the best result in each column.

Method	Training Data	Closed-loop Metrics $\uparrow$				Multi-Ability Test (%) $\uparrow$					
		Eff.	Comf.	SR	DS	Merge	Overtake	Brake	Give Way	Sign	Mean
UniAD-Base [6]	Think2Drive	129.21	43.58	16.36	45.81	14.10	17.78	21.67	10.00	14.21	15.55
VAD [7]	Think2Drive	157.94	46.01	15.00	42.35	8.11	24.44	18.64	20.00	19.15	18.07
MomAD (SD) [76]	Think2Drive	174.91	51.20	18.11	47.91	13.21	21.02	18.01	20.00	21.07	18.66
ThinkTwice* [77]	Think2Drive	69.33	16.22	31.23	62.44	27.38	18.42	35.82	50.00	54.23	37.17
DriveAdapter* [78]	Think2Drive	70.22	16.01	33.08	64.22	28.82	26.38	48.76	50.00	56.43	42.08
DriveTransformer-L [79]	Think2Drive	100.64	20.78	35.01	63.46	17.57	35.00	48.36	40.00	52.10	38.60
ReCogDrive [9]	Think2Drive	138.18	17.45	45.45	71.36	29.73	20.00	69.09	20.00	71.34	42.03
DriveMoE [23]	Think2Drive	175.96	15.31	48.64	74.22	34.67	40.00	65.45	40.00	59.44	47.91
ORION [12]	Think2Drive	151.48	17.38	54.62	77.74	25.00	71.11	78.33	30.00	69.15	54.72
AutoVLA [10]	PDM-Lite	146.93	39.33	57.73	78.84	–	–	–	–	–	–
SimLingo [30]	PDM-Lite	259.23	33.67	67.27	85.07	54.01	57.04	88.33	53.33	82.45	67.03
R2SE [80]	PDM-Lite	243.89	23.26	69.54	86.28	53.33	61.25	90.00	50.00	84.21	67.76
BridgeDrive [81]	PDM-Lite	236.49	20.98	74.99	87.99	69.92	66.67	90.00	50.00	89.47	73.15
LinkVLA [82]	PDM-Lite	255.84	34.62	74.55	91.01	60.00	80.00	93.33	50.00	83.68	73.40
TFv6 $^{\circ}$ [83]	LEAD	–	–	79.50	91.60	–	–	–	–	–	–
UniDriveVLA-Base	Think2Drive	198.86	11.78	51.82	78.37	38.75	80.00	50.00	30.00	58.95	51.53
UniDriveVLA-Base	LEAD	200.27	26.64	87.16	94.50	80.77	84.44	88.33	50.00	91.40	78.99

Perception. The nuScenes validation set [100] evaluates 3D detection, online mapping, and motion forecasting.

Planning. We evaluate open-loop planning on nuScenes under the ST-P3 [86] and UniAD [6] conventions and closed-loop driving on Bench2Drive [75] across five ability categories. NeuroNCAP [101] tests three safety-critical collision configurations, while Fail2Drive [102] evaluates closed-loop generalization on 100 matched route pairs covering in-distribution and long-tail scenarios.

3) *Evaluation Metrics:* FidelityDrivingBench [103] reports SD, T-QA, NoPR, and KRR. LingoQA reports Lingo-Judge, and general benchmarks use official aggregate scores.

On nuScenes, detection uses mean average precision (mAP) and the nuScenes detection score (NDS), while mapping uses class-wise average precision (AP) and mAP averaged over 0.5, 1.0, and 1.5 m Chamfer thresholds. Motion forecasting reports minimum average displacement error (minADE) and minimum final displacement error (minFDE) across six 12-step modes spanning 6 s at 2 Hz. Planning reports L2 error and collision rate at 1, 2, and 3 s and their means. ST-P3 averages 0.5-s steps through each horizon, whereas UniAD evaluates the corresponding endpoint.

Bench2Drive reports Success Rate (SR), Driving Score (DS), Efficiency, Comfort, and five ability-specific success rates. Its principal metrics are

$$SR = 100 \frac{N_{\text{success}}}{N}, \quad DS = \frac{1}{N} \sum_{i=1}^N RC_i \prod_j p_{i,j}. \quad (14)$$

Here, N is the route count, N_{success} counts infraction-free on-time completions, RC_i is route-completion percentage, and $p_{i,j}$ penalizes infraction j on route i .

NeuroNCAP reports collision rate and the following five-star score:

$$S_{\text{NCAP}} = \begin{cases} 5, & \text{if no collision,} \\ 4 \max\left(0, 1 - \frac{v_{\text{imp}}}{v_{\text{ref}}}\right), & \text{otherwise.} \end{cases} \quad (15)$$

Here, v_{imp} is relative impact speed and v_{ref} its no-intervention reference.

Fail2Drive reports Driving Score (DS), Success Rate (SR), and their harmonic mean (HM) on paired in-distribution and generalization splits:

$$HM = \frac{2 DS SR}{DS + SR}. \quad (16)$$

The paired routes preserve the driving context while introducing a targeted distribution shift, allowing the corresponding performance degradation to be measured directly.

B. Quantitative Results

a) *Experiments on the Bench2Drive Benchmark:* Tab. I summarizes closed-loop performance and fine-grained driving abilities on Bench2Drive. Under Think2Drive supervision, UniDriveVLA achieves the highest Efficiency and Driving Score of 198.86 and 78.37, respectively, while ranking first in merging and overtaking with 38.75% and 80.00%. Training on the higher-quality LEAD expert data further raises the Success Rate to 87.16% and the Driving Score to 94.50, outperforming TFv6 by 7.66 and 2.90 points, respectively. The mean multi-ability score concurrently increases from 51.53% to 78.99%, showing that UniDriveVLA effectively translates stronger expert supervision into more reliable planning and interactive driving.

TABLE II: Open-loop planning results on nuScenes under the ST-P3 [86] and UniAD [6] evaluation protocols. We report L2 error and collision rate at 1, 2, and 3 s and their averages, with and without ego-state input. Bold metric values indicate the best result within each setting. [‡] denotes SparseDrive results re-evaluated under the GPT-Driver protocol [87].

Method	ST-P3 metrics								UniAD metrics								LLM
	L2 (m) ↓				Collision (%) ↓				L2 (m) ↓				Collision (%) ↓				
	1s	2s	3s	Avg.	1s	2s	3s	Avg.	1s	2s	3s	Avg.	1s	2s	3s	Avg.	
Methods with Ego-State Input																	
ST-P3 [86]	1.33	2.11	2.90	2.11	0.23	0.62	1.27	0.71	-	-	-	-	-	-	-	-	-
VAD [7]	0.17	0.34	0.60	0.37	0.04	0.27	0.67	0.33	-	-	-	-	-	-	-	-	-
UniAD [6]	-	-	-	-	-	-	-	-	0.20	0.42	0.75	0.46	0.02	0.25	0.84	0.37	-
RDA-Driver [88]	0.17	0.37	0.69	0.40	0.01	0.05	0.26	0.10	0.23	0.73	1.54	0.80	0.00	0.13	0.83	0.32	LLaVA-7B
BEV-Planner [89]	0.16	0.32	0.57	0.35	0.00	0.29	0.73	0.34	-	-	-	-	-	-	-	-	-
HPP [90]	0.26	0.37	0.59	0.40	0.02	0.05	0.11	0.06	0.30	0.61	1.15	0.72	0.03	0.07	0.35	0.15	-
OmniDrive [19]	0.14	0.29	0.55	0.33	0.00	0.13	0.78	0.30	-	-	-	-	-	-	-	-	LLaVA-7B
ORION [12]	0.17	0.31	0.55	0.34	0.05	0.25	0.80	0.37	-	-	-	-	-	-	-	-	LLaVA-7B
FSDrive [22]	0.14	0.25	0.46	0.28	0.03	0.06	0.21	0.10	0.18	0.39	0.77	0.45	0.00	0.06	0.42	0.16	Qwen2-VL-3B
AutoVLA [10]	0.25	0.46	0.73	0.48	0.07	0.07	0.26	0.13	0.33	0.81	1.45	0.86	0.08	0.11	0.85	0.35	Qwen2.5-VL-3B
OpenDriveVLA [18]	0.14	0.30	0.55	0.33	0.02	0.07	0.22	0.10	0.19	0.58	1.24	0.67	0.02	0.18	0.70	0.30	Qwen2.5-VL-3B
UniDriveVLA-Base	0.19	0.32	0.53	0.35	0.03	0.04	0.14	0.07	0.23	0.56	1.11	0.63	0.00	0.07	0.40	0.16	Qwen3-VL-2B
UniDriveVLA-Large	0.19	0.32	0.53	0.35	0.03	0.05	0.10	0.06	0.24	0.55	1.10	0.63	0.00	0.05	0.28	0.11	Qwen3-VL-8B
Methods without Ego-State Input																	
VAD [7]	0.41	0.70	1.05	0.72	0.03	0.19	0.43	0.21	-	-	-	-	-	-	-	-	-
UniAD [6]	0.45	0.70	1.04	0.73	0.62	0.58	0.63	0.61	0.59	1.01	1.48	1.03	0.16	0.51	1.64	0.77	-
ELM [91]	-	-	-	-	-	-	-	-	0.34	1.23	2.57	1.38	0.12	0.50	2.36	0.99	BLIP2-2.7B
OccWorld [92]	0.39	0.73	1.18	0.77	0.11	0.19	0.67	0.32	0.52	1.27	2.41	1.40	0.12	0.40	2.08	0.87	GPT3-like
BEV-Planner [89]	0.30	0.52	0.83	0.55	0.10	0.37	1.30	0.59	-	-	-	-	-	-	-	-	-
HPP [90]	0.41	0.61	0.86	0.63	0.03	0.08	0.24	0.12	0.48	0.91	1.54	0.97	0.03	0.17	0.68	0.29	-
SparseDrive [‡] [26]	0.28	0.53	0.84	0.55	0.03	0.05	0.15	0.08	0.38	0.92	1.66	0.99	0.02	0.08	0.53	0.21	-
OmniDrive [19]	0.40	0.80	1.32	0.84	0.04	0.46	2.32	0.94	-	-	-	-	-	-	-	-	LLaVA-7B
FSDrive [22]	0.28	0.52	0.80	0.53	0.06	0.13	0.32	0.17	0.40	0.89	1.60	0.96	0.07	0.12	1.02	0.40	Qwen2-VL-3B
UniDriveVLA-Base	0.26	0.48	0.77	0.50	0.02	0.07	0.23	0.11	0.34	0.84	1.54	0.91	0.00	0.17	0.63	0.27	Qwen3-VL-2B
UniDriveVLA-Large	0.26	0.47	0.75	0.49	0.04	0.10	0.25	0.13	0.34	0.82	1.48	0.88	0.05	0.22	0.67	0.31	Qwen3-VL-8B

TABLE III: Comparison of detection, mapping, and motion prediction performance on the nuScenes validation dataset.

Method	Detection		Map			Motion		
	mAP↑	NDS↑	AP _{ped} ↑	AP _{divider} ↑	AP _{boundary} ↑	mAP↑	minADE(m) ↓	minFDE(m) ↓
UniAD [6]	0.380	0.498	-	-	-	-	0.710	1.020
VAD [7]	0.276	0.397	0.406	0.515	0.506	0.476	-	-
SparseDrive-S [26]	0.418	0.525	0.499	0.570	0.584	0.551	0.620	0.990
EgoFSD-S (PV) [93]	0.410	0.528	0.549	0.557	0.573	0.560	-	-
HiP-AD [70]	0.424	0.535	-	-	-	0.571	0.610	-
UniDriveVLA-Base	0.405	0.466	0.361	0.458	0.480	0.433	0.801	1.300
UniDriveVLA-Large	0.414	0.483	0.382	0.432	0.472	0.429	0.842	1.415

b) *Planning Results on the nuScenes Benchmark:* Tab. II presents a comprehensive comparison of end-to-end trajectory planning performance on nuScenes under the ST-P3 and UniAD evaluation protocols. Overall, UniDriveVLA-Large attains the lowest trajectory errors without ego-state input and matches or improves the best prior collision rates with ego-state input. In particular, under the more challenging setting without ego-state input, UniDriveVLA-Large attains the lowest average L2 errors of 0.49 m and 0.88 m under ST-P3 and UniAD, outperforming FSDrive by 0.04 m and 0.08 m, respectively. In the setting with ego-state input, UniDriveVLA-Large matches the best ST-P3 average collision rate of 0.06% and reduces the best prior UniAD average collision rate from 0.15% to 0.11%. The gains are most pronounced in trajectory accuracy without explicit ego-state input.

c) *Perception Results on the nuScenes Benchmark:* Beyond planning, we evaluate the spatial perception capabilities

of UniDriveVLA on the nuScenes validation set. As shown in Tab. III, UniDriveVLA-Large achieves a detection mAP of 0.414 and an NDS of 0.483, outperforming VAD by 0.138 and 0.086, respectively. Across its two model scales, UniDriveVLA further supports mapping and motion prediction, with UniDriveVLA-Base obtaining a map mAP of 0.433 and motion-prediction errors of 0.801 m minADE and 1.300 m minFDE. These results demonstrate that UniDriveVLA provides explicit spatial modeling across detection, mapping, and motion prediction while jointly supporting semantic understanding and action planning.

d) *Driving-Oriented Understanding:* As shown in Tab. IV, UniDriveVLA remains competitive on FidelityDrivingBench, with particularly strong T-QA and NoPR results, and achieves the highest Lingo-Judge score of 75.2, surpassing the closest competitor by 6.0 points. These results show that UniDriveVLA maintains driving-oriented semantic competence while jointly

TABLE IV: Driving-oriented understanding on FidelityDrivingBench and LingoQA. FidelityDrivingBench baseline results are taken from [103].

Method	FidelityDrivingBench				LingoQA
	SD↑	T-QA↑	NoPR↑	KRR↑	L-Judge↑
InternVL3 (8B) [14]	66.8	42.4	55.2	–	–
Qwen2.5-VL (7B) [104]	65.1	31.0	47.7	–	–
LLaVA-OV (7B)	63.7	35.9	43.3	–	–
ReCogDrive (8B) [9]	65.0	45.5	41.3	74.0%	67.2
ImpromptuVLA (7B) [105]	60.8	46.3	32.2	67.5%	–
RoboTron-Drive [106]	61.9	28.2	30.7	–	69.2
WiseAD [27]	47.8	19.8	5.2	30.0%	60.4
Mini-Intern-BDDX	51.1	33.8	19.8	53.5%	–
UniDriveVLA (8B)	57.9	42.3	32.3	48.3%	75.2

TABLE VI: Closed-loop safety evaluation under the original NeuroNCAP protocol [101]. We report NeuroNCAP Score (NS) and Collision Rate (CR, %).

Method	Stationary		Frontal		Side		Average	
	NS↑	CR↓	NS↑	CR↓	NS↑	CR↓	NS↑	CR↓
UniAD [6]	0.84	87.8	0.10	98.4	1.26	79.6	0.73	88.6
ImpromptuVLA [105]	2.55	54.8	1.86	72.8	1.78	67.6	2.06	65.1
Reasoning-VLA7B+ [110]	2.06	57.9	2.33	57.4	2.17	64.0	2.19	59.8
VaVAM [56]	3.13	47.2	2.67	50.0	2.07	60.8	2.62	52.7
PrismADDiffusionDrive [111]	4.31	21.8	2.28	75.2	3.25	35.2	3.28	44.1
ColaVLA [112]	3.54	32.3	3.16	51.6	3.75	26.4	3.48	36.8
ImagiDrive-S _{InternVL2.5} [113]	4.15	33.8	2.45	74.0	3.88	26.8	3.49	44.9
UniDriveVLA-Base	4.89	2.2	4.59	9.8	5.00	0.0	4.83	4.0

learning spatial perception and action planning.

e) General Visual Capabilities: A key challenge in driving-oriented adaptation is preserving the general multimodal capability of the foundation model. As shown in Tab. V, UniDriveVLA achieves the strongest results among the evaluated VLA models on MMMU (47.2), VLMsAreBlind (29.6), and ChartQA (65.7), while ranking second on the remaining four benchmarks. Although performance decreases relative to the Qwen3-VL-8B base model after driving adaptation, this broad performance profile indicates that UniDriveVLA retains meaningful general multimodal competence.

C. Robustness and Generalization

To further assess the robustness and generalization of UniDriveVLA in realistic closed-loop driving scenarios, we evaluate its safety-critical performance on NeuroNCAP and its response to long-tail scenario shifts on Fail2Drive.

a) Safety-Critical Robustness: As shown in Tab. VI, UniDriveVLA achieves the highest NeuroNCAP Score and the lowest collision rate in all three scenario categories, yielding averages of 4.83 and 4.0%, respectively. Compared with the prior best average score of 3.49 from ImagiDrive-S and the prior lowest average collision rate of 36.8% from ColaVLA, UniDriveVLA improves the score by 38.4% and reduces the collision rate by 89.1%, bringing the latter down by nearly an order of magnitude. Its collision rate is 2.2% for stationary encounters and 9.8% for frontal encounters, while no collision is recorded in side encounters. The consistent advantage across all three interaction types demonstrates robust closed-loop behavior in safety-critical scenarios.

TABLE V: General multimodal understanding of general-purpose VLMs and VLA models. Gray rows denote general-purpose VLMs, and bold marks the best VLA result in each column.

Model	General Multimodal Benchmarks						
	MMStar↑	MMMU↑	RWQA↑	AI2D↑	MME↑	VAB↑	ChartQA↑
GPT-5 Nano [107]	41.3	57.6	60.7	65.7	–	40.2	48.6
Qwen2.5-VL (7B) [104]	63.7	51.1	69.0	82.5	2317	40.3	82.5
Qwen3-VL (8B) [108]	63.0	52.8	69.0	83.2	2364	61.9	83.8
InternVL3 (8B) [14]	69.1	54.8	67.8	83.9	2393	42.0	86.6
ReCogDrive (8B) [9]	53.7	46.2	31.5	75.9	2156	11.9	56.9
ChatVLA (2B) [109]	44.6	37.7	57.9	66.5	1860	27.4	65.1
$\pi_{0.5}$ (2B) [50]	24.7	24.3	40.0	31.7	1264	8.9	6.7
Alpamayo (8B) [8]	10.2	24.6	3.8	0.0	0	0.3	0.0
UniDriveVLA (8B)	49.4	47.2	53.7	71.3	1876	29.6	65.7

TABLE VII: Closed-loop evaluation on the 100 matched route pairs of Fail2Drive [102]. We report Driving Score (DS), Success Rate (SR), and their harmonic mean (HM).

Method	In-Distribution			Generalization		
	DS↑	SR (%)↑	HM↑	DS↑	SR (%)↑	HM↑
TCP [114]	24.7	39.1	30.3	24.5	31.4	27.5
UniAD [6]	47.5	36.3	41.2	44.0	27.6	33.9
ORION [12]	53.0	52.0	52.5	51.2	46.0	48.5
HiP-AD [70]	74.1	70.7	72.4	67.1	56.7	61.5
SimLingo [30]	82.6	79.3	80.9	71.7	55.0	62.2
AlignDrive [115]	76.0	80.0	78.0	71.4	62.0	66.4
TFv5 [116]	83.3	78.5	80.8	75.4	61.1	67.5
UniDriveVLA-Base	85.80	85.0	85.40	76.37	68.0	71.94

b) Generalization under Long-Tail Scenario Shifts: As shown in Tab. VII, UniDriveVLA obtains DS, SR, and HM scores of 85.80, 85.0, and 85.40 on the in-distribution split and 76.37, 68.0, and 71.94 on the generalization split. Under the long-tail scenario shifts, it surpasses TFv5 by 0.97, 6.9, and 4.44 points in DS, SR, and HM, respectively. These results show that UniDriveVLA maintains strong closed-loop driving performance across both evaluation splits.

D. Ablation and Mechanism Analysis

a) Effect of MoT Decoupling: Tab. VIII compares the MoT-based architecture with a shared-weight decoder under the same training protocol. The MoT-based configuration raises LingoQA from 60.0 to 74.2, detection NDS from 0.437 to 0.466, and map mAP from 0.407 to 0.433, demonstrating concurrent gains in semantic understanding and spatial perception. It also improves closed-loop safety, increasing the NeuroNCAP Score from 3.45 to 4.83 and reducing the collision rate from 31.1% to 4.0%. Together, these results show that the MoT-based design achieves a stronger cross-task balance than the shared-weight baseline under the evaluated protocol.

b) Ablation of Perception Components: Tab. IX examines how progressively incorporating outputs from the Perception Expert affects closed-loop safety, with open-loop L2 error included for reference. The ego-status component improves open-loop accuracy but degrades closed-loop safety, showing that open-loop trajectory error alone does not fully characterize closed-loop safety. Adding detection reverses this trend, increasing the NeuroNCAP Score from 3.60 to 3.94 and reducing the

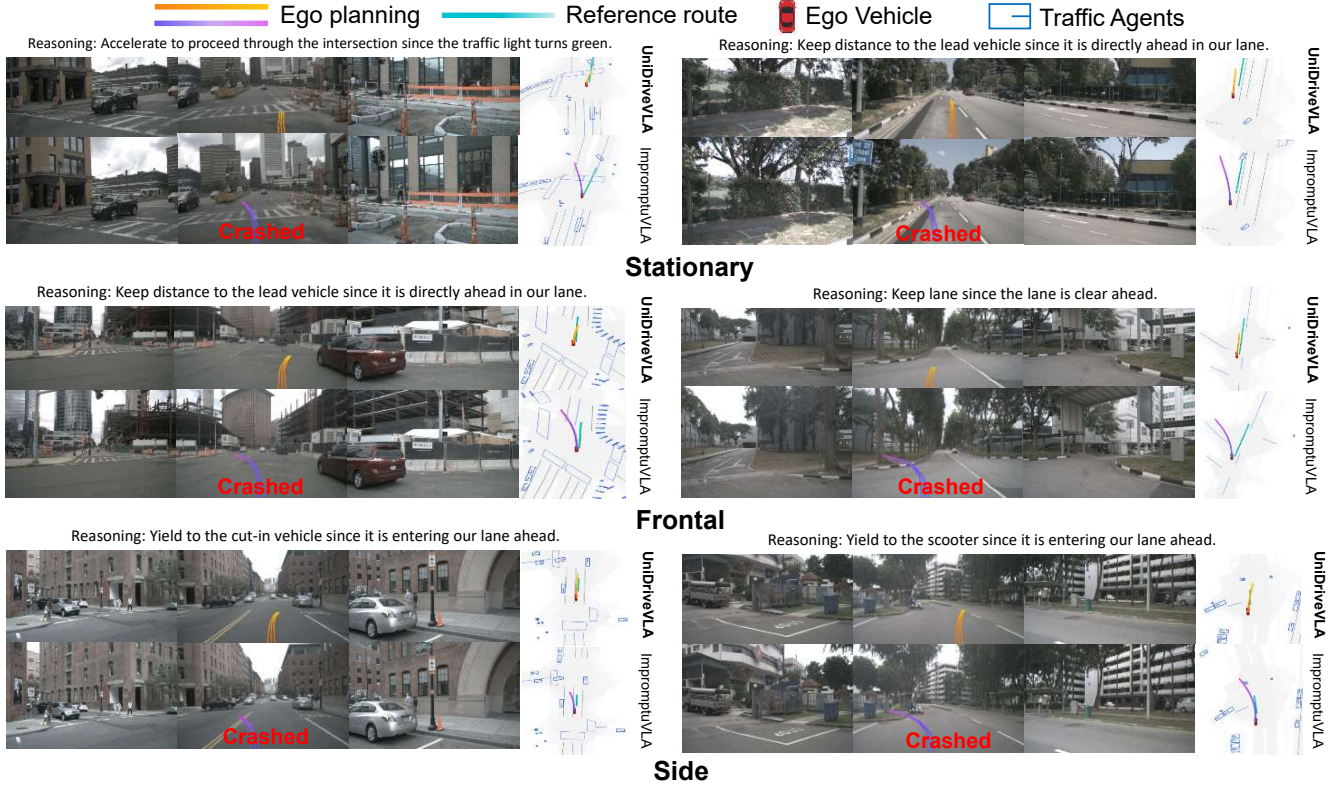


Fig. 4: **Qualitative closed-loop comparison on NeuroNCAP.** Each group shows UniDriveVLA (top) and ImpromptuVLA (bottom) in stationary, frontal, or side collision configurations. Colored curves visualize ego plans, cyan curves show reference routes, and red annotations mark collisions in the displayed ImpromptuVLA rollouts.

TABLE VIII: Comparison between a shared-weight decoder and Mixture-of-Transformers (MoT) in UniDriveVLA-Base across understanding, perception, and planning. General VQA averages seven multimodal benchmarks. LingoQA measures driving-oriented understanding, perception metrics are reported on nuScenes, and planning metrics on NeuroNCAP.

Architecture	Understanding		Perception		Planning	
	General VQA (%) $\uparrow$	LingoQA (%) $\uparrow$	Det (NDS) $\uparrow$	Map (mAP) $\uparrow$	NS $\uparrow$	CR (%) $\downarrow$
Shared-Weight Decoder	49.18	60.0	0.437	0.407	3.45	31.1
Mixture-of-Transformers	50.62	74.2	0.466	0.433	4.83	4.0

TABLE IX: Cumulative ablation of perception components in UniDriveVLA-Base. We report L2 error on nuScenes and NeuroNCAP Score (NS) and Collision Rate (CR, %).

Base	Components					Open-loop L ₂ (m) $\downarrow$	Closed-loop	
	Ego	Det	Map	Occ	Motion		NS $\uparrow$	CR $\downarrow$
✓						0.59	3.86	25.1
✓	✓					0.56	3.60	28.1
✓	✓	✓				0.49	3.94	22.7
✓	✓	✓	✓			0.51	4.04	19.8
✓	✓	✓	✓	✓		0.50	4.83	4.0
✓	✓	✓	✓	✓	✓	0.52	4.81	4.5

collision rate from 28.1% to 22.7%, while achieving the lowest L2 error of 0.49. Mapping further raises the score to 4.04 and lowers the collision rate to 19.8%. Occupancy provides the largest incremental gain, reaching a score of 4.83 and a collision rate of 4.0% while retaining a near-best L2 error of 0.50. Adding motion forecasting yields similar but slightly lower

TABLE X: Component-wise ReSPECT coefficient ablation on UniDriveVLA-Base ($r = 64$, $\alpha = 128$). We report LingoJudge, nuScenes perception, and NeuroNCAP safety.

Method	Coefficient		Understanding	Perception		Planning	
	β_{Vis}	β_{LM}	L-Judge $\uparrow$	Det NDS $\uparrow$	Map mAP $\uparrow$	NS $\uparrow$	CR (%) $\downarrow$
LoRA	—	—	66.8	0.461	0.424	4.09	18.8
ReSPECT	0.010	0.001	69.6	0.470	0.457	4.08	18.6
ReSPECT	0.010	0.003	74.2	0.466	0.433	4.83	4.0
ReSPECT	0.030	0.003	70.4	0.479	0.465	4.47	11.3

closed-loop safety metrics, indicating no further gain under this setting. Together, these results show that detection, mapping, and occupancy progressively strengthen closed-loop safety, with occupancy contributing the most substantial improvement.

c) *Effect of ReSPECT*: Tab. X examines how the component-wise ReSPECT coefficients affect semantic under-

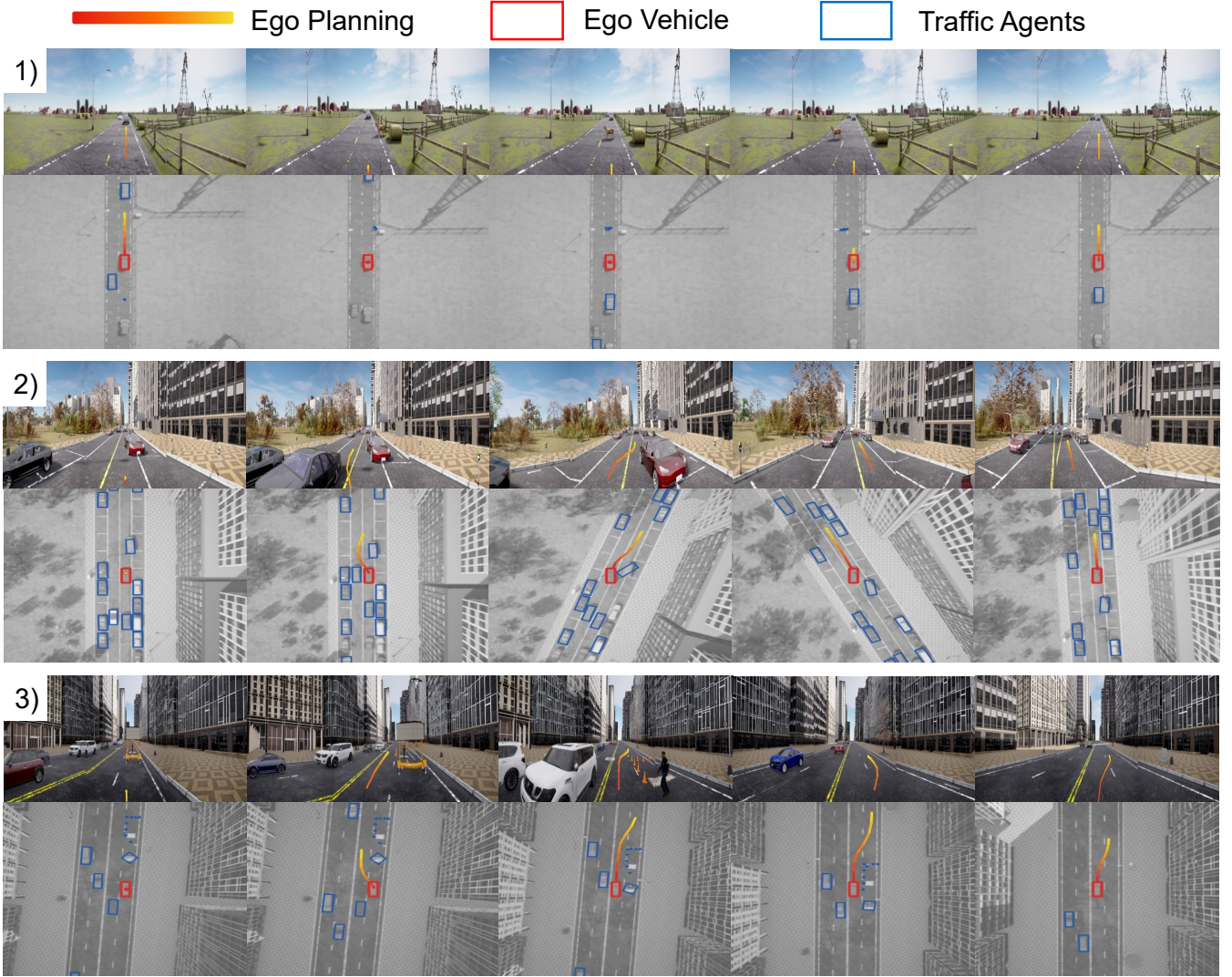


Fig. 5: **Qualitative closed-loop planning on Fail2Drive.** Camera observations and bird’s-eye-view predictions are shown under (1) animal, (2) bad parking, and (3) construction pedestrian. Red-to-yellow curves indicate ego plans, while blue boxes denote surrounding traffic agents.

standing, spatial perception, and closed-loop safety. Relative to standard LoRA, $(\beta_{\text{vis}}, \beta_{\text{LM}}) = (0.010, 0.003)$ raises Lingo-Judge from 66.8 to 74.2 and the NeuroNCAP Score from 4.09 to 4.83, while reducing the collision rate from 18.8% to 4.0% and improving both perception metrics. With the counterpart coefficient fixed, the higher tested β_{LM} yields a higher Lingo-Judge score, whereas the higher tested β_{vis} yields higher perception scores. Among these settings, the selected coefficients achieve the best Lingo-Judge and closed-loop safety while maintaining perception above standard LoRA.

E. Efficiency and Deployment

Despite their strong multimodal capabilities, the substantial inference cost of large-scale VLA models remains a major obstacle to real-time deployment in autonomous driving. We therefore reduce computation in visual processing and action decoding and accelerate the resulting computation graph at runtime. At the model level, we apply VisionZip [74] to select

TABLE XI: Inference efficiency and planning performance of UniDriveVLA-Flash on NVIDIA H100. Each row includes all preceding optimizations. ST-P3 reports average L2 error (m) / collision rate (%) without ego-status inputs. [†] denotes the complete UniDriveVLA-Flash configuration.

Configuration	Inference			Planning
	Latency↓	Hz↑	Speedup↑	ST-P3↓
UniDriveVLA-Base	0.761 s	1.31	1.00×	0.50 / 0.11
<i>Model-level</i>				
+ Visual token compression	0.651 s	1.54	1.17×	0.50 / 0.13
+ Perception modulation	0.599 s	1.67	1.27×	0.58 / 0.15
+ Learnable action queries	0.290 s	3.45	2.63×	0.55 / 0.15
<i>Runtime-level</i>				
+ TF32 + SDPA	0.252 s	3.97	3.02×	0.54 / 0.14
+ Torch Compile + CUDA Graphs	0.169 s	5.92	4.50×	0.55 / 0.13
+ TensorRT [†]	0.096 s	10.40	7.91×	0.55 / 0.13

and merge multi-view visual tokens before they enter the

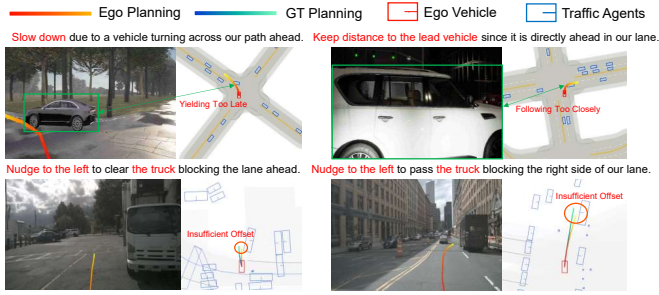


Fig. 6: **Failure cases of UniDriveVLA.** The model correctly identifies critical traffic participants and produces appropriate Chain-of-Causation (CoC) reasoning, but the intended semantics are not always translated into safe action trajectories, occasionally resulting in collisions.

Understanding Expert, thereby shortening its input sequence. Next, adaptive normalization uses sparse spatial representations from the decoupled Perception Expert to modulate the action suffix before it enters the Action Expert, preserving spatial conditioning without increasing the sequence length. Finally, learnable action queries support direct trajectory decoding in a single forward pass, eliminating iterative flow-matching sampling. At runtime, we enable TensorFloat-32 (TF32) arithmetic and fused scaled dot-product attention (SDPA), reduce launch overhead through `torch.compile` and CUDA Graphs, and use TensorRT for graph-level inference optimization. The complete configuration achieves 10.40 Hz inference on a single NVIDIA H100, reducing end-to-end latency from 761.0 to 96.2ms and yielding a $7.91\times$ speedup over the baseline (Tab. XI). Thus, UniDriveVLA-Flash supports 10 Hz planning updates while increasing the ST-P3 average L2 error by 0.05 m and the collision rate by 0.02 percentage points.

F. Qualitative Results

We present qualitative closed-loop results on NeuroNCAP and Fail2Drive to illustrate how UniDriveVLA responds to safety-critical interactions and long-tail driving scenarios.

Fig. 4 presents paired rollouts of UniDriveVLA and ImpromptuVLA across stationary, frontal, and side collision configurations. In the displayed UniDriveVLA outputs, the scene-level reasoning isolates the decision-relevant condition, with representative cases involving a changing traffic signal, a leading vehicle, a cut-in vehicle, and an entering scooter. The corresponding trajectories proceed, maintain clearance, or yield in accordance with the stated reasoning. By contrast, ImpromptuVLA collides in all displayed comparison rollouts. In these examples, the predicted trajectories are consistent with the generated scene interpretations.

Fig. 5 further examines closed-loop planning in three Fail2Drive long-tail scenarios: an animal entering the roadway, an improperly parked vehicle, and a pedestrian near a construction zone. In the displayed cases, UniDriveVLA adjusts its trajectory to the scenario-specific hazard while accounting for surrounding traffic. Together, these examples illustrate context-aware planning responses across distinct long-tail interactions.

V. LIMITATIONS AND FUTURE WORK

As shown in Fig. 6, although UniDriveVLA can identify critical traffic participants and generate appropriate Chain-of-Causation (CoC) reasoning, it does not always translate the corresponding driving intentions into safe trajectories. Common failures include delayed yielding, overly close following, and insufficient lateral avoidance, which may ultimately lead to collisions. These cases indicate a remaining alignment gap between CoC reasoning and continuous action generation. As model-generated supervision, the CoC annotations may contain occasional grounding errors or incomplete causal descriptions. Future work will improve annotation quality through grounding-based consistency filtering and targeted human verification, and further align reasoning with action through closed-loop reinforcement learning.

VI. CONCLUSION

We presented UniDriveVLA, a unified Vision-Language-Action model that integrates semantic understanding and spatial perception for reliable action planning in autonomous driving. UniDriveVLA assigns task-specific transformations to separate expert pathways and coordinates the corresponding token streams through Masked Joint Attention, while ReSPECT regulates low-rank updates during driving-domain adaptation to preserve pretrained semantic reasoning. Sparse spatial states extracted from multi-scale visual features provide the Action Expert with explicit spatial context alongside semantic reasoning for continuous action generation. UniDriveVLA-Flash further streamlines the architecture and runtime execution for real-time end-to-end inference. Extensive evaluations show that UniDriveVLA retains semantic understanding, demonstrates strong spatial perception, and achieves leading planning results under matched protocols in both open-loop and closed-loop settings. Results on NeuroNCAP and Fail2Drive further demonstrate robust closed-loop responses to safety-critical interactions and generalization under long-tail scenario shifts, respectively. Future work will extend UniDriveVLA toward longer-horizon reasoning and action planning in more diverse and interactive driving environments.

REFERENCES

- [1] Z. Li, W. Wang, H. Li, E. Xie, C. Sima, T. Lu, Q. Yu, and J. Dai, “Bevformer: learning bird’s-eye-view representation from lidar-camera via spatiotemporal transformers,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 47, no. 3, pp. 2020–2036, 2024. 1, 3
- [2] T. Liang, H. Xie, K. Yu, Z. Xia, Z. Lin, Y. Wang, T. Tang, B. Wang, and Z. Tang, “Befusion: A simple and robust lidar-camera fusion framework,” *Advances in neural information processing systems*, vol. 35, pp. 10 421–10 434, 2022. 1, 3
- [3] B. Liao, S. Chen, X. Wang, T. Cheng, Q. Zhang, W. Liu, and C. Huang, “Maptr: Structured modeling and learning for online vectorized hd map construction,” *arXiv preprint arXiv:2208.14437*, 2022. 1
- [4] Y. Chai, B. Sapp, M. Bansal, and D. Anguelov, “Multipath: Multiple probabilistic anchor trajectory hypotheses for behavior prediction,” *arXiv preprint arXiv:1910.05449*, 2019. 1
- [5] J. Gu, C. Hu, T. Zhang, X. Chen, Y. Wang, Y. Wang, and H. Zhao, “Vip3d: End-to-end visual trajectory prediction via 3d agent queries,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 5496–5506. 1
- [6] Y. Hu, J. Yang, L. Chen, K. Li, C. Sima, X. Zhu, S. Chai, S. Du, T. Lin, W. Wang *et al.*, “Planning-oriented autonomous driving,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 17 853–17 862. 1, 2, 7, 8, 9

- [7] B. Jiang, S. Chen, Q. Xu, B. Liao, J. Chen, H. Zhou, Q. Zhang, W. Liu, C. Huang, and X. Wang, "Vad: Vectorized scene representation for efficient autonomous driving," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 8340–8350. 1, 2, 7, 8
- [8] Y. Wang, W. Luo, J. Bai, Y. Cao, T. Che, K. Chen, Y. Chen, J. Diamond, Y. Ding, W. Ding *et al.*, "Alpamayo-r1: Bridging reasoning and action prediction for generalizable autonomous driving in the long tail," *arXiv preprint arXiv:2511.00088*, 2025. 1, 2, 3, 5, 6, 9
- [9] Y. Li, K. Xiong, X. Guo, F. Li, S. Yan, G. Xu, L. Zhou, L. Chen, H. Sun, B. Wang *et al.*, "Recogdrive: A reinforced cognitive framework for end-to-end autonomous driving," *arXiv preprint arXiv:2506.08052*, 2025. 1, 2, 6, 7, 9
- [10] Z. Zhou, T. Cai, S. Z. Zhao, Y. Zhang, Z. Huang, B. Zhou, and J. Ma, "Autovla: A vision-language-action model for end-to-end autonomous driving with adaptive reasoning and reinforcement fine-tuning," *arXiv preprint arXiv:2506.13757*, 2025. 1, 2, 7, 8
- [11] J.-J. Hwang, R. Xu, H. Lin, W.-C. Hung, J. Ji, K. Choi, D. Huang, T. He, P. Covington, B. Sapp *et al.*, "Emma: End-to-end multimodal model for autonomous driving," *arXiv preprint arXiv:2410.23262*, 2024. 1, 2
- [12] H. Fu, D. Zhang, Z. Zhao, J. Cui, D. Liang, C. Zhang, D. Zhang, H. Xie, B. Wang, and X. Bai, "Orion: A holistic end-to-end autonomous driving framework by vision-language instructed action generation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025, pp. 24 823–24 834. 1, 2, 7, 8, 9
- [13] S. Bai, Y. Cai, R. Chen, K. Chen, X. Chen, Z. Cheng, L. Deng, W. Ding, C. Gao, C. Ge, W. Ge, Z. Guo, Q. Huang, J. Huang, F. Huang, B. Hui, S. Jiang, Z. Li, M. Li, M. Li, K. Li, Z. Lin, J. Lin, X. Liu, J. Liu, C. Liu, Y. Liu, D. Liu, S. Liu, D. Lu, R. Luo, C. Lv, R. Men, L. Meng, X. Ren, X. Ren, S. Song, Y. Sun, J. Tang, J. Tu, J. Wan, P. Wang, P. Wang, Q. Wang, Y. Wang, T. Xie, Y. Xu, H. Xu, J. Xu, Z. Yang, M. Yang, J. Yang, A. Yang, B. Yu, F. Zhang, H. Zhang, X. Zhang, B. Zheng, H. Zhong, J. Zhou, F. Zhou, J. Zhou, Y. Zhu, and K. Zhu, "Qwen3-vl technical report," *arXiv preprint arXiv:2511.21631*, 2025. 1, 6
- [14] J. Zhu, W. Wang, Z. Chen, Z. Liu, S. Ye, L. Gu, H. Tian, Y. Duan, W. Su, J. Shao *et al.*, "Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models," *arXiv preprint arXiv:2504.10479*, 2025. 1, 9
- [15] R. Islam and O. M. Moushi, "Gpt-4o: The cutting-edge advancement in multimodal llm," in *Intelligent Computing-Proceedings of the Computing Conference*. Springer, 2025, pp. 47–60. 1
- [16] C. Dang, J. Wang, G. Li, Z. Hou, Z. You, H. Ye, J. Ma, L. Chen, and Y. Wang, "Sparseoccvla: Bridging occupancy and vision-language models via sparse queries for unified 4d scene understanding and planning," *arXiv preprint arXiv:2601.06474*, 2026. 1, 2
- [17] R. Liu, L. Kong, D. Li, and H. Zhao, "Occvla: Vision-language-action model with implicit 3d occupancy supervision," *arXiv preprint arXiv:2509.05578*, 2025. 1, 2
- [18] X. Zhou, X. Han, F. Yang, Y. Ma, V. Tresp, and A. Knoll, "Opendrivevla: Towards end-to-end autonomous driving with large vision language action model," *arXiv preprint arXiv:2503.23463*, 2025. 1, 2, 5, 8
- [19] S. Wang, Z. Yu, X. Jiang, S. Lan, M. Shi, N. Chang, J. Kautz, Y. Li, and J. M. Alvarez, "OmniDrive: A holistic vision-language dataset for autonomous driving with counterfactual reasoning," in *Proceedings of the computer vision and pattern recognition conference*, 2025, pp. 22 442–22 452. 1, 2, 5, 8
- [20] J. Wang, G. Li, Z. Huang, C. Dang, H. Ye, Y. Han, and L. Chen, "Vg-drive: Empowering vision-language models with cross-view geometric grounding for autonomous driving," *arXiv preprint arXiv:2602.20794*, 2026. 1
- [21] J. Han, M. Tian, J. Zhu, F. He, H. Zhang, S. Guo, D. Zhu, H. Tang, P. Xu, Y. Guo *et al.*, "Percept-wam: Perception-enhanced world-awareness-action model for robust end-to-end autonomous driving," *arXiv preprint arXiv:2511.19221*, 2025. 1, 2
- [22] S. Zeng, X. Chang, M. Xie, X. Liu, Y. Bai, Z. Pan, M. Xu, X. Wei, and N. Guo, "Futuresightdrive: Thinking visually with spatio-temporal cot for autonomous driving," *arXiv preprint arXiv:2505.17685*, 2025. 2, 8
- [23] Z. Yang, Y. Chai, X. Jia, Q. Li, Y. Shao, X. Zhu, H. Su, and J. Yan, "Drivemoe: Mixture-of-experts for vision-language-action model in end-to-end autonomous driving," *arXiv preprint arXiv:2505.16278*, 2025. 2, 3, 7
- [24] X. Tian, J. Gu, B. Li, Y. Liu, Y. Wang, Z. Zhao, K. Zhan, P. Jia, X. Lang, and H. Zhao, "Drivevlm: The convergence of autonomous driving and large vision-language models," *arXiv preprint arXiv:2402.12289*, 2024. 2
- [25] B. Jiang, S. Chen, B. Liao, X. Zhang, W. Yin, Q. Zhang, C. Huang, W. Liu, and X. Wang, "Senna: Bridging large vision-language models and end-to-end autonomous driving," *arXiv preprint arXiv:2410.22313*, 2024. 2
- [26] W. Sun, X. Lin, Y. Shi, C. Zhang, H. Wu, and S. Zheng, "Sparsedrive: End-to-end autonomous driving via sparse scene representation," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2025, pp. 8795–8801. 2, 3, 5, 8
- [27] S. Zhang, W. Huang, Z. Gao, H. Chen, and C. Lv, "Wisead: Knowledge augmented end-to-end autonomous driving with vision-language model," *arXiv preprint arXiv:2412.09951*, 2024. 2, 9
- [28] S. Xing, C. Qian, Y. Wang, H. Hua, K. Tian, Y. Zhou, and Z. Tu, "Openemba: Open-source multimodal model for end-to-end autonomous driving," in *Proceedings of the Winter Conference on Applications of Computer Vision*, 2025, pp. 1001–1009. 2
- [29] Y. Li, S. Shang, W. Liu, B. Zhan, H. Wang, Y. Wang, Y. Chen, X. Wang, Y. An, C. Tang *et al.*, "Drivevla-w0: World models amplify data scaling law in autonomous driving," *arXiv preprint arXiv:2510.12796*, 2025. 2, 3
- [30] K. Renz, L. Chen, E. Arani, and O. Sinavski, "Simlingo: Vision-only closed-loop autonomous driving with language-action alignment," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 11 993–12 003. 2, 7, 9
- [31] A. Jiang, Y. Gao, Z. Sun, Y. Wang, J. Wang, J. Chai, Q. Cao, Y. Heng, H. Jiang, Y. Dong *et al.*, "Diffvla: Vision-language guided diffusion planning for autonomous driving," *arXiv preprint arXiv:2505.19381*, 2025. 2
- [32] J. Li, J. Wu, D. Hu, X. Huang, B. Sun, Z. Hao, X. Lang, X. Zhu, and L. Zhang, "Sgdrive: Scene-to-goal hierarchical world cognition for autonomous driving," *arXiv preprint arXiv:2601.05640*, 2026. 2
- [33] L. Zeng, J. Yao, B. Liao, H. Tao, W. Liu, and X. Wang, "Diffusionvl: Translating any autoregressive models into diffusion vision language models," *arXiv preprint arXiv:2512.15713*, 2025. 2
- [34] F. Zhu, R. Wang, S. Nie, X. Zhang, C. Wu, J. Hu, J. Zhou, J. Chen, Y. Lin, J.-R. Wen *et al.*, "Llada 1.5: Variance-reduced preference optimization for large language diffusion models," *arXiv preprint arXiv:2505.19223*, 2025. 2
- [35] C. Dang, S. Ang, Y. Li, H. Tian, J. Wang, G. Li, H. Ye, J. Ma, L. Chen, and Y. Wang, "Drivefine: Refining-augmented masked diffusion vla for precise and robust driving," *arXiv preprint arXiv:2602.14577*, 2026. 2
- [36] Y. Ma, Y. Cao, W. Ding, S. Zhang, Y. Wang, B. Ivanovic, M. Jiang, M. Pavone, and C. Xiao, "dvlm-ad: Enhance diffusion vision-language-model for driving via controllable reasoning," *arXiv preprint arXiv:2512.04459*, 2025. 2
- [37] H. Fu, D. Zhang, Z. Zhao, J. Cui, H. Xie, B. Wang, G. Chen, D. Liang, and X. Bai, "Minddrive: A vision-language-action model for autonomous driving via online reinforcement learning," *arXiv preprint arXiv:2512.13636*, 2025. 2
- [38] S. Zhang, W. Huang, Z. Chen, C. J. Collister, Q. Huang, and C. Lv, "Openread: Reinforced open-ended reasoning for end-to-end autonomous driving with llm-as-critic," *arXiv preprint arXiv:2512.01830*, 2025. 2
- [39] T. Yan, T. Tang, X. Gui, Y. Li, J. Zhesng, W. Huang, L. Kong, W. Han, X. Zhou, X. Zhang *et al.*, "Ad-r1: Closed-loop reinforcement learning for end-to-end autonomous driving with impartial world models," *arXiv preprint arXiv:2511.20325*, 2025. 2
- [40] M. Xu, J. Cui, F. Cai, H. Shang, Z. Zhu, S. Luan, Y. Xu, N. Zhang, Y. Li, J. Cai *et al.*, "Wam-diff: A masked diffusion vla framework with moe and online reinforcement learning for autonomous driving," *arXiv preprint arXiv:2512.11872*, 2025. 2
- [41] S. Shang, B. Zhan, Y. Yan, Y. Wang, Y. Li, Y. An, X. Wang, J. Liu, L. Hou, L. Fan *et al.*, "Dynvla: Learning world dynamics for action reasoning in autonomous driving," *arXiv preprint arXiv:2603.11041*, 2026. 2
- [42] L. Zhou, H. Luo, Z. Zhu, C. Chi, M. Tu, K. Xiong, L. Gong, Z. Wu, Z. Zhang, F. Li *et al.*, "Xiaomi auto world model: A joint world model integrating reconstruction and generation for autonomous driving," *arXiv preprint arXiv:2605.18137*, 2026. 2
- [43] Y. Bai, D. Wu, Y. Liu, F. Jia, W. Mao, Z. Zhang, Y. Zhao, J. Shen, X. Wei, T. Wang *et al.*, "Is a 3d-tokenized llm the key to reliable autonomous driving?" *arXiv preprint arXiv:2405.18361*, 2024. 2
- [44] Z. Liu, R. Huang, R. Yang, S. Yan, Z. Wang, L. Hou, D. Lin, X. Bai, and H. Zhao, "Drivepi: Spatial-aware 4d mllm for unified autonomous driving understanding, perception, prediction and planning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026, pp. 3688–3698. 2

- [45] W. Liang, L. Yu, L. Luo, S. Iyer, N. Dong, C. Zhou, G. Ghosh, M. Lewis, W.-t. Yih, L. Zettlemoyer *et al.*, “Mixture-of-transformers: A sparse and scalable architecture for multi-modal foundation models,” *arXiv preprint arXiv:2411.04996*, 2024. 2, 3
- [46] C. Deng, D. Zhu, K. Li, C. Gou, F. Li, Z. Wang, S. Zhong, W. Yu, X. Nie, Z. Song *et al.*, “Emerging properties in unified multimodal pretraining,” *arXiv preprint arXiv:2505.14683*, 2025. 2
- [47] B. Tang, B. Zheng, S. Paul, and S. Xie, “Exploring the deep fusion of large language models and diffusion transformers for text-to-image synthesis,” in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 28 586–28 595. 2
- [48] L. Qin, J. Gong, Y. Sun, T. Li, M. Yang, X. Yang, C. Qu, Z. Tan, and H. Li, “Uni-cot: Towards unified chain-of-thought reasoning across text and vision,” *arXiv preprint arXiv:2508.05606*, 2025. 2
- [49] K. Black, N. Brown, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai, L. Groom, K. Hausman, B. Ichter *et al.*, “pi_0: A vision-language-action flow model for general robot control,” *arXiv preprint arXiv:2410.24164*, 2024. 3
- [50] Physical Intelligence, K. Black, N. Brown, J. Darpanian, K. Dhabalia, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai *et al.*, “ $\pi_{0.5}$: A vision-language-action model with open-world generalization,” *arXiv preprint arXiv:2504.16054*, 2025. 3, 9
- [51] C. Cheang, S. Chen, Z. Cui, Y. Hu, L. Huang, T. Kong, H. Li, Y. Li, Y. Liu, X. Ma *et al.*, “Gr-3 technical report,” *arXiv preprint arXiv:2507.15493*, 2025. 3
- [52] G. Team, A. Ye, B. Wang, C. Ni, G. Huang, G. Zhao, H. Li, J. Li, J. Zhu, L. Feng *et al.*, “Gigabrain-0: A world model-powered vision-language-action model,” *arXiv preprint arXiv:2510.19430*, 2025. 3
- [53] T. Yuan, Y. Liu, C. Lu, Z. Chen, T. Jiang, and H. Zhao, “Depthvla: Enhancing vision-language-action models with depth-aware spatial reasoning,” *arXiv preprint arXiv:2510.13375*, 2025. 3
- [54] H. Bi, H. Tan, S. Xie, Z. Wang, S. Huang, H. Liu, R. Zhao, Y. Feng, C. Xiang, Y. Rong *et al.*, “Motus: A unified latent action world model,” *arXiv preprint arXiv:2512.13030*, 2025. 3
- [55] J. Cai, Z. Cai, J. Cao, Y. Chen, Z. He, L. Jiang, H. Li, H. Li, Y. Li, Y. Liu *et al.*, “Internvla-a1: Unifying understanding, generation and action for robotic manipulation,” *arXiv preprint arXiv:2601.02456*, 2026. 3
- [56] F. Bartoccioni, E. Ramzi, V. Besnier, S. Venkataramanan, T.-H. Vu, Y. Xu, L. Chambon, S. Gidaris, S. Odabas, D. Hurych *et al.*, “Vavim and vavam: Autonomous driving through video generative modeling,” *arXiv preprint arXiv:2502.15672*, 2025. 3, 9
- [57] W. Huang, S. Zhang, Q. Huang, Z. Wang, Z. Mao, C. Chua, Z. Chen, L. Chen, and C. Lv, “Automot: A unified vision-language-action model with asynchronous mixture-of-transformers for end-to-end autonomous driving,” *arXiv preprint arXiv:2603.14851*, 2026. 3
- [58] S. Zuo, Y. Li, W. Zheng, Z. Zhu, J. Zhou, and J. Lu, “Vega: Learning to drive with natural language instructions,” *arXiv preprint arXiv:2603.25741*, 2026. 3
- [59] J. Philion and S. Fidler, “Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d,” in *European conference on computer vision*. Springer, 2020, pp. 194–210. 3
- [60] Y. Li, Z. Ge, G. Yu, J. Yang, Z. Wang, Y. Shi, J. Sun, and Z. Li, “Bevdepth: Acquisition of reliable depth for multi-view 3d object detection,” in *Proceedings of the AAAI conference on artificial intelligence*, vol. 37, no. 2, 2023, pp. 1477–1485. 3
- [61] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, “End-to-end object detection with transformers,” in *European conference on computer vision*. Springer, 2020, pp. 213–229. 3
- [62] Y. Wang, V. C. Guizilini, T. Zhang, Y. Wang, H. Zhao, and J. Solomon, “Detr3d: 3d object detection from multi-view images via 3d-to-2d queries,” in *Conference on robot learning*. PMLR, 2022, pp. 180–191. 3
- [63] Y. Liu, T. Wang, X. Zhang, and J. Sun, “Petr: Position embedding transformation for multi-view 3d object detection,” in *European conference on computer vision*. Springer, 2022, pp. 531–548. 3
- [64] X. Lin, T. Lin, P. Pei, L. Huang, and Z. Su, “Sparse4d: Multi-view 3d object detection with sparse spatial-temporal fusion,” *arXiv preprint arXiv:2211.10581*, 2022. 3
- [65] H. Liu, Y. Teng, T. Lu, H. Wang, and L. Wang, “Sparsebev: High-performance sparse 3d object detection from multi-camera videos,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 18 580–18 590. 3
- [66] R. Zhu, J. Zhao, D. Zhang, G. Wang, X. Chen, S. Zhang, J. Gong, Q. Zhou, W. Zhang, N. Wang *et al.*, “Sparsead: Sparse query-centric paradigm for efficient end-to-end autonomous driving,” *IEEE Transactions on Artificial Intelligence*, 2025. 3
- [67] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen *et al.*, “Lora: Low-rank adaptation of large language models,” *Iclr*, vol. 1, no. 2, p. 3, 2022. 4, 6
- [68] Y. Xiong and X. Xie, “Oplora: Orthogonal projection lora prevents catastrophic forgetting during parameter-efficient fine-tuning,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 40, no. 40, 2026, pp. 34 088–34 096. 4
- [69] H. Wang, Y. Li, S. Wang, G. Chen, and Y. Chen, “Milora: Harnessing minor singular components for parameter-efficient llm finetuning,” in *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 2025, pp. 4823–4836. 4
- [70] Y. Tang, Z. Xu, Z. Meng, and E. Cheng, “Hip-ad: Hierarchical and multi-granularity planning with deformable attention for autonomous driving in a single decoder,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025, pp. 25 605–25 615. 5, 8, 9
- [71] S. Gu, W. Yin, B. Jin, X. Guo, J. Wang, H. Li, Q. Zhang, and X. Long, “Dome: Taming diffusion model into high-fidelity controllable occupancy world model,” *arXiv preprint arXiv:2410.10429*, 2024. 5
- [72] Y. Lipman, R. T. Q. Chen, H. Ben-Hamu, M. Nickel, and M. Le, “Flow matching for generative modeling,” *arXiv preprint arXiv:2210.02747*, 2023. 5
- [73] Y. Zheng, T. Tan, B. Huang, E. Liu, R. Liang, J. Zhang, J. Cui, G. Chen, K. Ma, H. Ye *et al.*, “Unleashing the potential of diffusion models for end-to-end autonomous driving,” *arXiv preprint arXiv:2602.22801*, 2026. 5, 6
- [74] S. Yang, Y. Chen, Z. Tian, C. Wang, J. Li, B. Yu, and J. Jia, “Visionzip: Longer is better but not necessary in vision language models,” *arXiv preprint arXiv:2412.04467*, 2026. 5, 11
- [75] X. Jia, Z. Yang, Q. Li, Z. Zhang, and J. Yan, “Bench2drive: Towards multi-ability benchmarking of closed-loop end-to-end autonomous driving,” *Advances in Neural Information Processing Systems*, vol. 37, pp. 819–844, 2024. 7
- [76] Z. Song, C. Jia, L. Liu, H. Pan, Y. Zhang, J. Wang, X. Zhang, S. Xu, L. Yang, and Y. Luo, “Don’t shake the wheel: Momentum-aware planning in end-to-end autonomous driving,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025, pp. 22 432–22 441. 7
- [77] X. Jia, P. Wu, L. Chen, J. Xie, C. He, J. Yan, and H. Li, “Think twice before driving: Towards scalable decoders for end-to-end autonomous driving,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 21 983–21 994. 7
- [78] X. Jia, Y. Gao, L. Chen, J. Yan, P. L. Liu, and H. Li, “Driveadapter: Breaking the coupling barrier of perception and planning in end-to-end autonomous driving,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 7953–7963. 7
- [79] X. Jia, J. You, Z. Zhang, and J. Yan, “Drivetransformer: Unified transformer for scalable end-to-end autonomous driving,” *arXiv preprint arXiv:2503.07656*, 2025. 7
- [80] H. Liu, T. Li, H. Yang, L. Chen, C. Wang, K. Guo, H. Tian, H. Li, H. Li, and C. Lv, “Reinforced refinement with self-aware expansion for end-to-end autonomous driving,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026. 7
- [81] S. Liu, W. Chen, W. Li, Z. Wang, L. Yang, J. Huang, Z. Huang, Z. Cheng, and H. Yang, “Bridgedrive: Diffusion bridge policy for closed-loop trajectory planning in autonomous driving,” in *International Conference on Learning Representations*, vol. 2026, 2026, pp. 130 989–131 009. 7
- [82] X. Wang, Q. Liu, W. Ding, Z. Yang, W. Li, C. Liu, B. Li, K. Zhan, X. Lang, and W. Chen, “Unifying language-action understanding and generation for autonomous driving,” *arXiv preprint arXiv:2603.01441*, 2026. 7
- [83] L. Nguyen, M. Fauth, B. Jaeger, D. Dauner, M. Igl, A. Geiger, and K. Chitta, “Lead: Minimizing learner-expert asymmetry in end-to-end driving,” in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026. 7
- [84] L. Wiedmann, O. Zohar, A. Mahla, X. Wang, R. Li, T. Frere, L. von Werra, A. R. Gosthipaty, and A. Marafioti, “Finevision: Open data is all you need,” *arXiv preprint arXiv:2510.17269*, 2025. 6
- [85] I. Loshchilov and F. Hutter, “Decoupled weight decay regularization,” *arXiv preprint arXiv:1711.05101*, 2017. 6
- [86] S. Hu, L. Chen, P. Wu, H. Li, J. Yan, and D. Tao, “St-p3: End-to-end vision-based autonomous driving via spatial-temporal feature learning,” in *European Conference on Computer Vision*. Springer, 2022, pp. 533–549. 7, 8

- [87] J. Mao, Y. Qian, J. Ye, H. Zhao, and Y. Wang, "Gpt-driver: Learning to drive with gpt," *arXiv preprint arXiv:2310.01415*, 2023. 8
- [88] Z. Huang, T. Tang, S. Chen, S. Lin, Z. Jie, L. Ma, G. Wang, and X. Liang, "Making large language models better planners with reasoning-decision alignment," in *European Conference on Computer Vision*. Springer, 2024, pp. 73–90. 8
- [89] Z. Li, Z. Yu, S. Lan, J. Li, J. Kautz, T. Lu, and J. M. Alvarez, "Is ego status all you need for open-loop end-to-end autonomous driving?" in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 14 864–14 873. 8
- [90] H. Liu, Z. Huang, W. Huang, H. Yang, X. Mo, and C. Lv, "Hybrid-prediction integrated planning for autonomous driving," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 47, no. 4, pp. 2597–2614, 2025. 8
- [91] Y. Zhou, L. Huang, Q. Bu, J. Zeng, T. Li, H. Qiu, H. Zhu, M. Guo, Y. Qiao, and H. Li, "Embodied understanding of driving scenarios," in *European Conference on Computer Vision*. Springer, 2024, pp. 129–148. 8
- [92] W. Zheng, W. Chen, Y. Huang, B. Zhang, Y. Duan, and J. Lu, "Occworld: Learning a 3d occupancy world model for autonomous driving," in *European conference on computer vision*. Springer, 2024, pp. 55–72. 8
- [93] H. Su, W. Wu, and J. Yan, "Difsd: Ego-centric fully sparse paradigm with uncertainty denoising and iterative refinement for efficient end-to-end self-driving," *arXiv preprint arXiv:2409.09777*, 2024. 8
- [94] A.-M. Marcu, L. Chen, J. Hünemann, A. Karnsund, B. Hanotte, P. Chidananda, S. Nair, V. Badrinarayanan, A. Kendall, J. Shotton *et al.*, "Lingoqa: Visual question answering for autonomous driving," in *European Conference on Computer Vision*. Springer, 2024, pp. 252–269. 6
- [95] L. Chen, J. Li, X. Dong, P. Zhang, Y. Zang, Z. Chen, H. Duan, J. Wang, Y. Qiao, D. Lin *et al.*, "Are we on the right way for evaluating large vision-language models?" *Advances in Neural Information Processing Systems*, vol. 37, pp. 27 056–27 087, 2024. 6
- [96] X. Yue, Y. Ni, K. Zhang, T. Zheng, R. Liu, G. Zhang, S. Stevens, D. Jiang, W. Ren, Y. Sun *et al.*, "Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 9556–9567. 6
- [97] C. Fu, P. Chen, Y. Shen, Y. Qin, M. Zhang, X. Lin, J. Yang, X. Zheng, K. Li, X. Sun *et al.*, "Mme: A comprehensive evaluation benchmark for multimodal large language models," *arXiv preprint arXiv:2306.13394*, 2023. 6
- [98] P. Rahmanzadehgervi, L. Bolton, M. R. Taesiri, and A. T. Nguyen, "Vision language models are blind," in *Proceedings of the Asian Conference on Computer Vision*, 2024, pp. 18–34. 6
- [99] A. Masry, X. L. Do, J. Q. Tan, S. Joty, and E. Hoque, "Chartqa: A benchmark for question answering about charts with visual and logical reasoning," in *Findings of the association for computational linguistics: ACL 2022*, 2022, pp. 2263–2279. 6
- [100] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nusenes: A multimodal dataset for autonomous driving," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 621–11 631. 7
- [101] W. Ljungbergh, A. Tonderski, J. Johnander, H. Caesar, K. Åström, M. Felsberg, and C. Petersson, "Neuroncap: Photorealistic closed-loop safety testing for autonomous driving," in *European Conference on Computer Vision*. Springer, 2024, pp. 161–177. 7, 9
- [102] S. Gerstenecker, A. Geiger, and K. Renz, "Fail2drive: Benchmarking closed-loop driving generalization," *arXiv preprint arXiv:2604.08535*, 2026. 7, 9
- [103] R. Mao, H. Wang, Y. Yang, Q. Ma, J. Zhou, and Z. Zhang, "The blind spot of adaptation: Quantifying and mitigating forgetting in fine-tuned driving models," *arXiv preprint arXiv:2604.04857*, 2026. 7, 9
- [104] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang *et al.*, "others. 2025. qwen2. 5-vl technical report," *arXiv preprint arXiv:2502.13923*, vol. 4, no. 5, 1. 9
- [105] H. Chi, H.-a. Gao, Z. Liu, J. Liu, C. Liu, J. Li, K. Yang, Y. Yu, Z. Wang, W. Li *et al.*, "Impromptu vla: Open weights and open data for driving vision-language-action models," *arXiv preprint arXiv:2505.23757*, 2025. 9
- [106] Z. Huang, C. Feng, F. Yan, B. Xiao, Z. Jie, Y. Zhong, X. Liang, and L. Ma, "Robotron-drive: All-in-one large multimodal model for autonomous driving," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025, pp. 8011–8021. 9
- [107] A. Singh, A. Fry, A. Perelman, A. Tart, A. Ganesh, A. El-Kishky, A. McLaughlin, A. Low, A. Ostrow, A. Ananthram *et al.*, "Openai gpt-5 system card," *arXiv preprint arXiv:2601.03267*, 2025. 9
- [108] S. Bai, Y. Cai, R. Chen, K. Chen, X. Chen, Z. Cheng, L. Deng, W. Ding, C. Gao, C. Ge *et al.*, "Qwen3-vl technical report," *arXiv preprint arXiv:2511.21631*, 2025. 9
- [109] Z. Zhou, Y. Zhu, M. Zhu, J. Wen, N. Liu, Z. Xu, W. Meng, Y. Peng, C. Shen, F. Feng *et al.*, "Chatvla: Unified multimodal understanding and robot control with vision-language-action model," in *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 2025, pp. 5377–5395. 9
- [110] D. Zhang, Z. Yuan, Z. Chen, C.-T. Liao, Y. Chen, F. Shen, Q. Zhou, and T.-S. Chua, "Reasoning-vla: A fast and general vision-language-action reasoning model for autonomous driving," *arXiv preprint arXiv:2511.19912*, 2025. 9
- [111] K. Ding, Z. Lin, H. Wang, J. Gui, Q. Liu, Z. Wang, L. Tang, and L. He, "PrismAD: Decoupled planning via semantic mixture-of-planners for end-to-end autonomous driving," *arXiv preprint arXiv:2607.10336*, 2026. 9
- [112] Q. Peng, X. Chen, C. Yang, S. Shi, and H. Li, "ColaVLA: Leveraging cognitive latent reasoning for hierarchical parallel trajectory planning in autonomous driving," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026. 9
- [113] J. Li, B. Zhang, X. Jin, J. Deng, X. Zhu, and L. Zhang, "Imagidrive: A unified imagination-and-planning framework for autonomous driving," *arXiv preprint arXiv:2508.11428*, 2025. 9
- [114] P. Wu, X. Jia, L. Chen, J. Yan, H. Li, and Y. Qiao, "Trajectory-guided control prediction for end-to-end autonomous driving: A simple yet strong baseline," *Advances in Neural Information Processing Systems*, vol. 35, pp. 6119–6132, 2022. 9
- [115] Y. Wu, H. Zhang, F. He, R. Wu, Y. Shan, C. Qiu, L. Gao, W. Ke, and T. Zhang, "Aligndrive: Aligned lateral-longitudinal planning for end-to-end autonomous driving," *arXiv preprint arXiv:2601.01762*, 2026. 9
- [116] B. Jaeger, K. Chitta, and A. Geiger, "Hidden biases of end-to-end driving models," in *2023 IEEE/CVF international conference on computer vision (ICCV)*. IEEE, 2023, pp. 8206–8215. 9



Yongkang Li received the bachelor's degree in communication and information engineering from Huazhong University of Science and Technology, Wuhan, China, in 2025. He is currently pursuing the M.S. degree at Huazhong University of Science and Technology, where he works with the HUST Vision Lab. His research interests include multimodal perception, open-world recognition, and autonomous driving.



Yu Zhu is currently pursuing the B.S. degree with the School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan, China, where she is a member of the HUST Vision Lab. Her research interests include vision-language models, vision-language-action models, and autonomous driving.



Lijun Zhou received her Ph.D. degree from the University of Chinese Academy of Sciences, Beijing, China, in 2021. She is currently an Algorithm Researcher at XiaomiEV. Her research interests include World Model, Video Generation, and Autonomous Driving.



Sixu Yan received the B.E. degree in mechanical engineering from Ocean University of China, Qingdao, China, in 2021, and the M.S. degree in mechanical engineering from Shanghai Jiao Tong University, Shanghai, China, in 2024. He is currently pursuing the Ph.D. degree with the School of Electronic Information and Communications, Huazhong University of Science and Technology, Wuhan, China. His research interests include robotics and computer vision.



Bing Wang received the Ph.D. degree from the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore, in 2016. He is currently an Expert Algorithm Engineer with Xiaomi EV. His research interests include computer vision, machine learning, world models, autonomous driving, and robotics.



Bencheng Liao received the B.E. degree from the School of Electronic Information and Communications, Huazhong University of Science and Technology, Wuhan, China, in 2020. He is currently a Ph.D. candidate with the Institute of Artificial Intelligence and the School of Electronic Information and Communications, Huazhong University of Science and Technology. His research interests include object detection, 3D vision, and autonomous driving.



Guang Chen received the Ph.D. degree from the University of Missouri in 2014. He is currently an Expert Algorithm Engineer with Xiaomi EV. His research interests include computer vision, machine learning, and autonomous driving.



Tianyi Yan received the B.Eng. degree in computer science from Beijing University of Posts and Telecommunications, Beijing, China, in 2020, and the M.S. degree in pattern recognition from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2023. He is currently pursuing the Ph.D. degree with the State Key Laboratory of Internet of Things for Smart City, University of Macau, Macau, China. His research interests include generative world models, end-to-end autonomous driving, and large language model post-training.



Wenyu Liu (Senior Member, IEEE) received the B.S. degree in Computer Science from Tsinghua University in 1986, and the M.S. and Ph.D. degrees, both in Electronics and Information Engineering, from Huazhong University of Science and Technology (HUST) in 1991 and 2001. He is now a professor and associate dean of the School of Electronic Information and Communications, HUST. His research areas include artificial intelligence and efficient deep learning.



Kaixin Xiong is with Xiaomi EV. He received the master's degree in information and communication engineering from Huazhong University of Science and Technology, Wuhan, China, in 2023. He is currently an Algorithm Engineer at XiaomiEV. His research interests include World Model, 3D vision and Autonomous Driving.



Zhenxin Zhu received the M.S. degree in Control Science and Engineering from Beihang University, Beijing, China, in 2025. He is currently an Algorithm Engineer at XiaomiEV. His research interests include Physical AI and World Models.



Hangjun Ye received the Ph.D. degree from the Department of Computer Science and Technology, Tsinghua University, Beijing, China, in 2003. He is currently the Head of the Autonomous Driving and Robotics Division at Xiaomi EV. His research interests include computer vision, machine learning, autonomous driving, and robotics.



Hongwei Xie received his M.S. degree from Nanjing University, China, in 2016. He currently works at Xiaomi EV, focusing on computer vision and autonomous driving. He has authored or co-authored over 20 papers in prestigious venues including CVPR, ICCV, and NeurIPS.



Xinggang Wang (Senior Member, IEEE) received the B.S. and Ph.D. degrees from Huazhong University of Science and Technology (HUST), Wuhan, China, in 2009 and 2014, respectively. He is now a professor at the School of Electronic Information and Communications, HUST. He serves as Co-Editor-in-Chief of Image and Vision Computing, associate editor of IEEE TPAMI, and area chair of CVPR and ICCV. His research interests include computer vision and deep learning.



Haiyang Sun received the master's degree in information and communication engineering from Tsinghua University, Beijing, China, in 2016. He is currently with Xiaomi EV. His research interests include world models, 3D vision, and autonomous driving.