

HybridDriveVLA: Vision-Language-Action Model with Visual CoT reasoning and ToT Evaluation for Autonomous Driving

Supplementary Material

A. Evaluation Metrics: Additional Details

This section provides further details on the evaluation metrics used to assess HybridDriveVLA, complementing the descriptions in the main paper.

A.1. Trajectory Planning Metrics

The main paper introduces the primary metrics for trajectory planning: L2 Displacement Error, Collision Rate, UniAD Metrics, and ST-P3 Metrics. It is important to note that a collision is formally defined as any instance where the ego-vehicle’s predicted bounding box overlaps with the ground-truth bounding box of any other object in the scene. Both UniAD and ST-P3 evaluation styles are used to ensure comprehensive and fair comparisons with a wide range of existing methods.

A.2. Visual Prediction Metric

While the core focus of HybridDriveVLA is on planning, the quality of the visual predictions from the Visual Chain-of-Thought (V-CoT) is crucial for its performance. To quantify this, we use the following metric, which was not detailed in the main paper: **Fréchet Inception Distance (FID)**: Evaluates the quality of generated future scenes frames by measuring the perceptual distance between the distributions of real and generated images. A lower FID score indicates that the generated images are more realistic and visually similar to the ground-truth distribution. This metric is essential for validating that the V-CoT produces high-fidelity visual goals for the planning module.

A.3. NAVSIM Benchmark Metrics

The main paper lists the closed-loop metrics for the NAVSIM benchmark. These metrics holistically evaluate driving behavior in a simulated environment, where the model’s own predictions influence future states. This closed-loop setup is a more realistic and challenging test of an agent’s driving policy compared to open-loop metrics, as it requires the model to recover from its own errors. The combination of metrics like collision rate, goal progress, and comfort provides a comprehensive picture of the agent’s real-world viability.

B. Future Scenes Generation Analysis

The Visual Chain-of-Thought (V-CoT) component of HybridDriveVLA generates future scene images, which serve

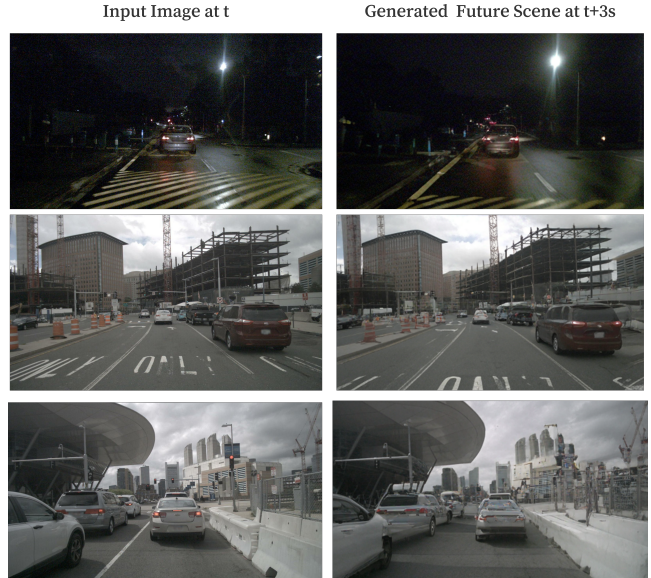


Figure 1. Example of future scene generated by V-CoT at a 3-second horizon. The model accurately predicts the continuity of the road, the position of the lead vehicle, and the surrounding environment.

as a visual goal for the ToT-evaluation reasoning. To validate the quality of these visual predictions, we evaluate the Fréchet Inception Distance (FID) on the nuScenes dataset.

Table 1. Future scenes images generation results on the nuScenes dataset. Lower FID is better.

Method	DriveDreamer [1]	GenAD [2]	Ours
Type	Diffusion	Diffusion	Autoreg.
Resolution	128×192	256×448	256×448
FID ↓	52.6	15.4	9.8

B.1. Analysis of Visual Prediction Quality

Table 1 presents a comparative analysis of future scenes generation quality, measured by FID. HybridDriveVLA achieved a High FID score of 9.8, outperforming both diffusion-based models like GenAD. This high performance was significant because, unlike specialized generative models optimized solely for visual quality, HybridDriveVLA’s V-CoT is an integrated component of a larger reasoning framework.

Notably, while our primary goal for V-CoT is to produce temporally coherent visual goals for planning rather than

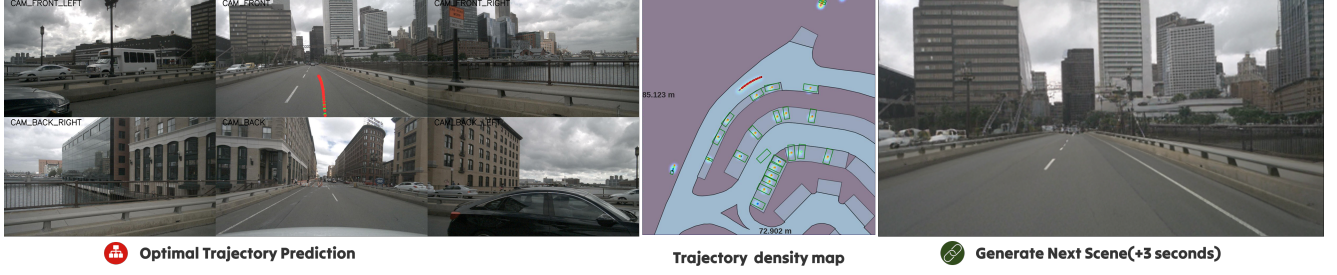


Figure 2. Example qualitative results of HybridDriveVLA optimal trajectory and generated future scene. HybridDriveVLA demonstrates robust performance in complex urban scenarios by generating safe and efficient trajectories. The optimal predicted trajectory is shown in red, while the ground truth is in green.



Figure 3. Qualitative results of HybridDriveVLA in diverse conditions. (Top) Navigating in cloudy and rainy weather. (Middle) Driving in a poorly lit environment at night. (Bottom) Driving in a dense urban area. HybridDriveVLA adapts by prioritizing safety, comfort, and progress. The optimal predicted trajectory is shown in red and the ground truth in green.

photorealistic video, the low FID score indicates that our model generates high-fidelity future scenes. Unlike specialized generative models that are optimized solely for visual quality, HybridDriveVLA’s V-CoT is an integrated component of a larger reasoning framework. The superior FID score suggests that grounding visual generation in a deliberative planning process not only improves action prediction but also enhances the quality of the visual predictions themselves. This is because the model must generate scenes that are not just visually plausible but also causally consistent with safe and logical driving actions.

C. Additional Qualitative and Quantitative Analysis

To comprehensively evaluate HybridDriveVLA, we present additional visual results across a variety of challenging driving scenarios. These examples highlight the model’s robustness and the interpretability of its decision-making process, particularly the interplay between V-CoT and ToT-Evaluation.



(a) Comfort: Prioritizes passenger comfort by executing smooth turns and minimizing abrupt changes in speed



(b) Safety: Prioritizes cautious positions by maintaining a significant distance from other vehicles and lane edges.



(c) Progress: Prioritizes taking the most direct route possible to make efficient progress towards the goal.



(d) Optimal: trade-offs between safety, comfort, and progress to achieve the most optimal overall trajectory

Figure 4. Visualization of aspect-based trajectories deliberation by ToT-Evaluation and the ground truth GT (green). HybrideDriveVLA evaluated multiple waypoint sequences prioritizing Safety (purple), Comfort (yellow), and Progress (blue). The final optimal trajectory (red) is selected after scoring all candidates.

C.1. Qualitative Visualizations in Diverse Conditions

Complex Urban Environments: As shown in Figure 3, HybridDriveVLA accurately predicts safe and efficient trajectories in crowded urban settings. It correctly anticipates the movement of other vehicles and adheres to lane boundaries. V-CoT module generates future state of the environment, including the likely positions of other agents, allowing ToT-Evaluation module to score and select a sequence of waypoints as a trajectory that makes progress while maintaining safety and comfort.

Challenging Weather and Lighting Conditions: We showcased HybridDriveVLA’s performance in diverse conditions such as rain and nighttime driving in Figure 3. The robustness of vision encoder allows HybridDriveVLA to extract features even from noisy or low-light images. V-CoT anticipates the scene in advance, while ToT-Evaluation module appropriately weighs the safety aspect, often resulting in more cautious driving behavior, such as increased following distance, which is consistent with human driving adaptive reasoning.

Aspect-based Trajectory Generation: A key feature of HybridDriveVLA is its ability to generate trajectories that prioritize different driving aspects (Safety, Comfort, Progress). Figure 4 shows the alternative trajectories considered by the ToT-Evaluation module. In the example scenario, the progress trajectory (blue) was more aggressive, while the safety trajectory (purple) stayed around the lane center and maintained a larger distance from obstacles. The Comfort trajectory exhibited smoother steering for smooth turns. The optimal path is selected by balancing these aspects, demonstrating HybridDriveVLA’s deliberative evaluation capabilities.

C.2. Computational Complexity and Runtime Analysis

We report the computational requirements and runtime performance of HybridDriveVLA. All experiments were conducted on an NVIDIA H100 GPUs.

Analysis: The full HybridDriveVLA model, with its integrated V-CoT and ToT-Evaluation modules, operates at 3.8 FPS. While this is not yet real-time, it is competitive with other large autoregressive models. The computational overhead is primarily from the sequential generation of visual and textual tokens. This performance level is more than adequate for offline evaluation and simulation, and it demonstrates the architectural feasibility of our joint hybrid reasoning approach. Future work will explore model optimization techniques such as quantization and distillation to improve inference speed for real-world deployment.

References

- [1] Xiaofeng Wang, Zheng Zhu, Guan Huang, Xinze Chen, Jiagang Zhu, and Jiwen Lu. Drivedreamer: Towards real-world-drive world models for autonomous driving. In *European conference on computer vision*, pages 55–72. Springer, 2024. 1
- [2] Wenzhao Zheng, Ruiqi Song, Xianda Guo, Chenming Zhang, and Long Chen. Genad: Generative end-to-end autonomous driving. pages 87–104, 2024. 1