BranchOut: Capturing Realistic Multimodality in Autonomous Driving Decisions

Hee Jae Kim Zekai Yin Lei Lai Jason Lee Eshed Ohn-Bar
Boston University
{hjkim37, zekaiyin, leilai, jaslee20, eohnbar}@bu.edu

Abstract: Modeling the nuanced, multimodal nature of human driving remains a core challenge for autonomous systems, as existing methods often fail to capture the diversity of plausible behaviors in complex real-world scenarios. In this work, we introduce a novel end-to-end planner and benchmark for modeling *realistic multimodality* in autonomous driving decisions. We propose a Gaussian Mixture Model (GMM)-based diffusion model designed to explicitly capture human-like, multimodal driving decisions in diverse contexts. Our model achieves state-of-the-art performance on current benchmarks, but reveals weaknesses in standard evaluation practices which rely on single ground-truth trajectories or coarse closed-loop metrics while also penalizing diverse yet plausible alternatives. To address this limitation, we further develop a human-in-the-loop simulation benchmark that enables finer-grained evaluations and measures multimodal realism in challenging driving settings. Our code, models, and benchmark data will be released to promote more accurate and human-aligned autonomous driving models.

1 Introduction

In a typical driving scenario, there are many plausible and safe paths a human driver might take [1–4]. For example, navigating around a stopped truck on the shoulder of the road can involve a range of headway distances, speeds, and lateral offsets. At the same time, even a subtle difference in a predicted trajectory can distinguish safe behavior from a safety-critical outcome [5]. The combination of variability, uncertainty, and required precision makes modeling and evaluation of realistic driving decisions inherently challenging [6–13].

Due to the complexity in capturing the distribution of driving decisions, recent approaches for motion planning leverage powerful generative architectures, such as transformers and diffusion models [14–29]. However, while prior models demonstrate coarse multimodality, e.g., distinguishing between major maneuvers such as turning versus going straight at intersections [9, 17, 30, 31], it remains unclear whether finer variations observed in real-world human driving behaviors, such as around dynamic agents, are represented by the model. Our analysis reveals such limitations, which are addressed using a proposed higher-capacity model architecture and objective. Specifically, we demonstrate that a well-designed model effectively outperforms all prior vision-based planners on nuScenes [32] by a notable margin.

Beyond persistent issues in model coverage [17, 21, 33–37], *evaluating multimodality* remains a significant challenge. Standard open-loop metrics typically compare predictions to a single ground-truth trajectory, thus failing to account for the diversity of plausible alternatives [5, 7, 7, 9, 26, 28, 38–43]. Closed-loop simulation offers a promising alternative [8, 10, 32, 44–46], but often falls short in modeling realistic environmental dynamics, agent interactions, and subtleties of decision-making. Moreover, it can be difficult to precisely specify safe and desirable driving via simplistic closed-loop metrics such as time to collision or drive area compliance [6, 47]. Without direct human demonstrations and feedback, such setups may yield reactive behaviors that appear safe, but lack fidelity to how

9th Conference on Robot Learning (CoRL 2025), Seoul, Korea.

Figure 1: **Capturing Multimodality in Complex Real-World Driving Scenarios.** We study modeling and evaluation of intricate multimodal driving settings, including subtle interactions, e.g., vehicle–vehicle and human–vehicle. As an example, we visualize collected multimodal trajectories in two scenarios: navigating around a vehicle parked on the shoulder (**left**) and interacting with a dynamic agent at an intersection (**right**).

humans actually drive in real-world social scenarios. We aim to improve evaluation by introducing a benchmark that captures diverse, plausible behaviors beyond single-trajectory comparisons.

Our goal is to develop effective models for producing realistic and human-like autonomous driving decisions. Our key contributions are twofold. First, we design an ***expressive diffusion-based model*** for end-to-end vision-based planning that captures a richer distribution of diverse, human-like driving behaviors. Second, we demonstrate that a human-in-the-loop simulation framework combining photorealistic 3D reconstruction with a physically-plausible kinematics model enables collection of a ***realistic multimodal benchmark***. By re-driving scenes, we densify annotations in existing driving benchmarks in a cheap and scalable manner while uncovering limitations of prior evaluation protocols. Our multimodal model and benchmark thus advance safe and seamless decision-making in autonomous driving.

2 Related Work

Generative Models for Autonomous Driving: While there is extensive work on probabilistic behavior prediction [21, 31, 48–54], most state-of-the-art motion planners remain deterministic [8, 9, 25–27, 55–57]. Recently, diffusion-based generative models have shown promise in autonomous driving tasks, with methods such as Diffusion Planner [19] achieving competitive results in nuPlan [32]. However, this method relies on privileged state information, is not end-to-end, and does not explicitly evaluates the diversity or realism of the sampled trajectories. Moreover, prior work has shown diffusion models suffer from mode collapse [17, 33, 34, 36, 37, 58], often generating limited variations around dominant modes. DiffusionDrive [17] alleviates this issue using guided sampling with pre-clustered anchors [59], yet relies on input anchors which are clustered independently (e.g., via K-Means). In contrast, we propose a holistic model which optimizes for multimodality jointly with model structure and optimization, i.e., using a GMM head. We show this structure to outperform DiffusionDrive, despite the lower number of samples.

Real-World Planning Benchmarks and Evaluation: Real-world benchmarks for end-to-end motion planning remain limited in both trajectory diversity and evaluation methodology. Most studies leverage offline datasets (e.g., nuScenes [40], Waymo [60, 61]) to compute open-loop metrics [7, 9, 11, 26, 27, 29, 41, 62, 63]. However, as widely acknowledged in the robotics community, open-loop evaluation fails to capture the sequential and reactive nature of planning policies when deployed [39]. Critically, nearly all existing motion forecasting and planning benchmarks provide only a *single demonstration* per scenario, despite the fact that multiple plausible decisions often exist in real-world contexts. This unimodal framing risks *penalizing valid predictions* and obscuring unsafe behaviors that otherwise match the logged trajectory Furthermore, collecting diverse, high-quality real-world trajectories is prohibitively expensive and potentially dangerous, creating a major barrier to dataset scalability. Nonetheless, recent advances in scene reconstruction and photorealistic simulation, however, offer a promising path forward, as discussed next.

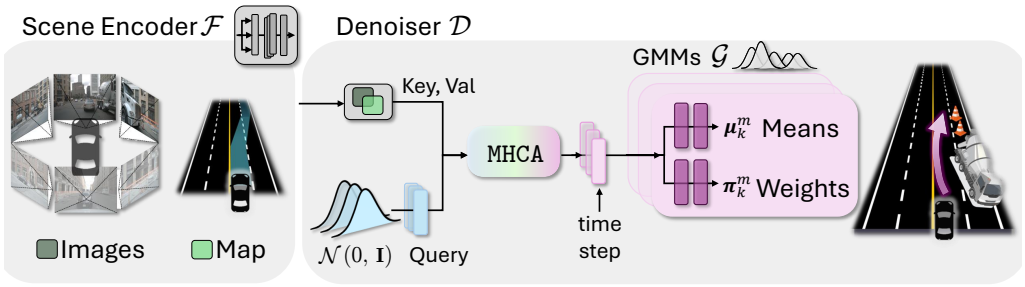


Figure 2: **Our End-to-End, GMM-Based Diffusion Planner.** BranchOut consists of a scene encoder $\mathcal{F}$ and a scene-aware transformer-based denoiser $\mathcal{D}$. The encoder $\mathcal{F}$ processes multi-view camera images and an HD map to extract scene features that condition the denoiser using multi-head cross-attention (MHCA), with scene features as keys and values to condition the ego query. A GMM head $\mathcal{G}$, selected using high-level driving command c, takes the transformed features and predicts K trajectory means ($\boldsymbol{\mu}^m$) and corresponding weights ($\boldsymbol{\pi}^m$), enabling the model to select the most likely future trajectory $\hat{\mathbf{Y}}$. We find that the simple yet effective GMM head outperforms more complex diffusion-based models. When combined, we find complementary improvements in multimodal plan modeling and achieve state-of-the-art results across benchmarks, including an introduced multimodal decisions benchmark.

3D Reconstruction for Simulation: Recent advances in 3D scene reconstruction have enabled photorealistic rendering of complex driving environments from monocular or multi-view video. Techniques such as NeRF [64–68] and Gaussian Splatting [69, 70] produce high-fidelity representations of real-world scenes, opening new possibilities for simulation without requiring handcrafted assets. By enabling high-resolution re-driving in richly contextualized environments, these tools offer a promising foundation for scalable trajectory generation—without repeated physical deployment. However, many existing approaches exhibit artifacts, scene desegmentation in cluttered driving contexts, and limited generalization to out-of-view areas. NeuroNCAP [46] incorporates LiDAR-based reconstruction into a closed-loop simulator, but focuses only on simplified, non-reactive settings without dense traffic, dynamic pedestrians, or complex maneuvers such as overtaking and merging. OmniRe [70] supports editing and decomposition but often suffers from object noise in real-world datasets like nuScenes. HUGSIM [44] advances photorealistic rendering, yet the realism of the underlying agent behaviors and physical responses remains unclear. In contrast, our framework is the first to empirically demonstrate how human-in-the-loop re-driving in a reconstruction-based simulator can offer a realistic augmentation of data coverage. By integrating interactive human input, we also take a step toward scalable human-aligned collection and validation of diverse, multimodal trajectories and enables fine-grained analysis of planning behavior in complex, reactive scenarios.

3 Method

Our goal is to model realistic, human-like driving behavior that captures the diversity of plausible decisions within each scenario. To this end, we design **BranchOut**, a GMM-based diffusion planner architecture that sets a new state-of-the-art in motion planning (Sec. 3.1). Next, to understand the multimodal capabilities of our framework, we collect a novel benchmark, which incorporates 3D reconstruction, collision detection, and interactive ego-agent control to ensure immersive re-driving (Sec 3.2). To rigorously validate the realism of the simulation and benchmark, we propose to compare with a fully reactive digital twin with comprehensive physics simulation and rendering. In Sec. 4, we will further compare to real-world logs to show the high realism of trajectories in simulation. An overview of our method can be seen in Fig. 2.

3.1 BranchOut: An End-to-End, GMM-Based Diffusion Planner

Our formulation follows the standard end-to-end planning setup [26, 27]. Specifically, we assume a *scene context* $\mathbf{C}$, consisting of six multi-view camera images [40] and an HD map [27]. Unlike

some prior work [9], we do not incorporate ego-status information or the vehicle's past trajectory in planning. We also assume access to a high-level driving command from a discrete set of $M = 3$ options, i.e., $c \in \{\texttt{Left}, \texttt{Straight}, \texttt{Right}\}$.

To train a probabilistic, multimodal planner from real-world data, we use driving logs that provide the ego-vehicle's future trajectory. Each sample is structured as $\mathbf{Y}_{\text{ego}} \in \mathbb{R}^{T_f \times 2}$, representing three seconds of future bird's-eye view (BEV) waypoints (i.e., $T_f = 6$). We train a *scene encoder* $\mathcal{F}$ to process $\mathbf{C}$, followed by a *transformer-based denoiser* $\mathcal{D}$ that predicts K future trajectory distributions. The denoiser is equipped with a GMM head $\mathcal{G}$, which predicts multiple candidate trajectories and associated mixture weights [71, 72]. We next define these functions and our training process. Our analysis demonstrates that a GMM head contributes more significantly to multimodal prediction quality than the diffusion process alone, though both components together yield the strongest performance.

Diffusion Model: Prior work has leveraged diffusion to generate a single ego-trajectory along with surrounding agents [19]. In contrast, we focus on predicting *multiple diverse future trajectories for the ego vehicle*, enabling richer reasoning over planning modes. Given a ground-truth trajectory $\mathbf{Y}_{\text{ego}}$, we generate noisy versions for training by sampling a diffusion timestep $t \sim \mathcal{U}(0, 1)$ and Gaussian noise $\mathbf{z} \sim \mathcal{N}(0, \mathbf{I})$. Noisy inputs are computed as:

$$\mathbf{X}_{\text{ego}}^{(t)} = \sqrt{\alpha(t)} \cdot \mathbf{Y}_{\text{ego}} + \sqrt{1 - \alpha(t)} \cdot \mathbf{z}, \quad \mathbf{X}_{\text{ego}}^{(t)} \in \mathbb{R}^{M \times T_f \times 2}, \tag{1}$$

where $\mathbf{X}_{\text{ego}}^{(t)}$ contains perturbed trajectories corresponding to the M high-level commands. Following standard diffusion formulations [15, 73], we define $\alpha(t) = 1 - \sigma^2(t)$, where $\sigma(t)$ determines the noise level at timestep t. At inference time, we initialize from a pure Gaussian noise sample $\mathbf{X}_{\text{ego}}^{(1)} \sim \mathcal{N}(0, \mathbf{I})$ and generate the final trajectory by solving the reverse diffusion ODE using a single-step DPM-Solver++ [74, 75] method.

Scene-Aware Diffusion Transformer Model: The transformer-based denoiser $\mathcal{D}$ is trained to reconstruct clean trajectories $\hat{\mathbf{Y}}_{\text{ego}}$ from noisy inputs $\mathbf{X}_{\text{ego}}^{(t)}$ by leveraging scene-conditioned representations (fused via cross-attention [76]). First, we linearly project the noisy set of ego trajectories $\mathbf{X}_{\text{ego}}^{(t)}$ into an embedding $\mathbf{P} \in \mathbb{R}^{M \times N_p}$. Next, we compute a scene-aware representation:

$$\mathbf{P} = [\text{MHCA}(\mathbf{P}, \mathbf{P}_{\text{agent}}, \mathbf{P}_{\text{agent}}), \text{MHCA}(\mathbf{P}, \mathbf{P}_{\text{map}}, \mathbf{P}_{\text{map}})], \tag{2}$$

where MHCA(q, k, v) computes cross-attentions between queries q, keys k, and values v [76]. The $\mathbf{P}_{\text{agent}} \in \mathbb{R}^{N_a \times N_d}$ and map embeddings $\mathbf{P}_{\text{map}} \in \mathbb{R}^{N_m \times N_d}$ are scene-aware embeddings computed from a transformer-based encoder $\mathcal{F}$ following VAD [27]. In our experiments in Sec. 4, we leverage the VAD-Tiny encoder architecture [27], but significantly outperform the VAD baseline due to the GMM-based diffusion process. To incorporate the diffusion timestep information, we further scale and shift $\mathbf{P}$ based on a learned function of the timestep embedding $\gamma(t)$ (see our supplementary for additional implementation details).

Branched GMM Head: The computed scene-aware feature $\mathbf{P}$ is inputted to a branched GMM head $\mathcal{G}$, which predicts K future trajectory distributions for each command mode:

$$\mathcal{G}(\mathbf{P}) = \{(\boldsymbol{\mu}_k^m, \boldsymbol{\pi}_k^m)\}_{k=1}^K, \tag{3}$$

where $\boldsymbol{\mu}_k^m \in \mathbb{R}^{T_f \times 2}$ denotes the mean trajectory of the k-th Gaussian component for the m-th command, and $\boldsymbol{\pi}_k^m$ denotes its associated probability (weight). Each command-specific branch in our network comprises two separate MLPs, one for the trajectory means and one for the mixing coefficients, i.e., $\boldsymbol{\mu}^m = \text{MLP}_{\boldsymbol{\mu}}^m(\mathbf{P}) \in \mathbb{R}^{K \times T_f \times 2}$ and $\boldsymbol{\pi}^m = \text{MLP}_{\boldsymbol{\pi}}^m(\mathbf{P}) \in \mathbb{R}^K$.

Training Loss: We optimize our model using a loss function comprising three terms:

$$\mathcal{L} = \mathcal{L}_{\text{plan}} + \lambda_{\text{NLL}} \mathcal{L}_{\text{NLL}} + \lambda_{\text{c}} \mathcal{L}_{\text{constraints}}, \tag{4}$$

where $\mathcal{L}_{\text{plan}}$ is the data reconstruction loss commonly used in diffusion [19], $\mathcal{L}_{\text{NLL}}$ is a negative log-likelihood (NLL) [77] loss over predicted multimodal trajectory distribution parameters, and $\mathcal{L}_{\text{constraints}}$ is a safety constraint loss based on Jiang et al. [27]. The hyperparameters λ_{NLL} and λ_{c} are set to 0.1. Complete loss term definitions implementation details are in the supplementary.

3.2 Human-in-the-Loop Simulation and Multimodal Benchmark

As our study emphasizes multimodal decision-making in driving, we aim to assess how well model predictions align with the diversity of human behavior. However, collecting multimodal trajectories in real-world scenes is not feasible. Instead, we leverage a human-in-the-loop, closed-loop photorealistic environment based on the state-of-the-art monocular-based HUGSIM [44] paired with a kinematics model, as discussed below. To ensure meaningful findings, we validate the realism of the resulting trajectories by comparing to real-world logs, i.e., how well one of the simulation trajectories matches the real-world log. Given potential issues with reconstruction quality and agent reactivity, we further compare with human-based driving in a full digital twin world, i.e., by creating a subset of the real-world scenes in CARLA [78].

Kinematics Model: To enable scalable collection of diverse, realistic trajectories, we augment HUGSIM [44] with a human-in-the-loop interface that is easy to set up and operate. Ego-vehicle motion is simulated using a kinematic bicycle model [12, 57], and interactive re-driving is supported through a fully immersive hardware setup.

Collision Feedback: To enhance realism and ensure trajectory feasibility, we employ automatic collision detection using depth predictions from the reconstruction module. When a collision is detected, the user is reset to the starting point and the previous path is cleared. Such collisions provide feedback that helps participants adapt quickly in the early stages of the study.

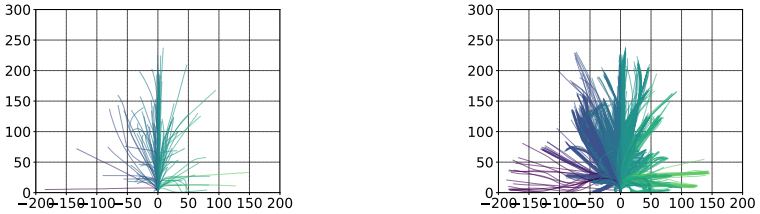
Trajectory Collection Study: Using our reconstructed and virtual environments, we conduct a user study to collect diverse driving trajectories. Each participant is spawned at the same starting position and orientation in the simulation. Participants re-drive the scene five times to cover a range of safe and plausible maneuvers. In total, 40 participants completed drives across randomly assigned scenes. Further details on the setup and protocol are in the supplementary material.

4 Experiments

In this section, we first discuss suitable metrics for our multimodal settings (Sec. 4.1). Second, we assess the realism of the collected diverse driving benchmark in simulation (Sec. 4.2). Third, we evaluate the performance of our proposed end-to-end planner on nuScenes and HUGSIM, augmented with our realistic annotations (Sec. 4.3). Finally, we discuss limitations of existing unimodal evaluation metrics in capturing the inherent complexity of real-world driving behaviors and explore more robust evaluation metrics. We include ablations in our supplementary material.

4.1 Setup and Metrics

We build on standard evaluation practices for end-to-end motion planners [17, 26]. Following prior work [26, 79], we report L2 displacement error against the single ground-truth trajectory provided by nuScenes. For multimodal evaluation, we leverage 16 ground-truth trajectories per scene—15 from our benchmark and one from nuScenes—and compute the minimum Fréchet distance [80] among the predicted motion plan and the set of ground-truth trajectories. To evaluate the quality of the trajectory distribution and fully analyze models' performance, we further report distribution-based metrics: the Negative Log-Likelihood (NLL) and Jensen-Shannon Divergence (JSD) [45, 81]. As many motion planners today are deterministic, we can compute such metrics by sampling from the models with Monte Carlo dropout (rate = 0.1) [82]. However, we are aware of the limitations of such techniques, i.e., compared to methods that train with sampling. For such stochastic models, including DiffusionDrive [17] and our BranchOut, we evaluate directly over the set of generated trajectories. For closed-loop evaluation, we follow HUGSIM [44] and compute No Collision (NC), Drivable Area Compliance (DAC), Time to Collision (TTC), Comfort (COM), Route Completion (R_c), and the HUGSIM Driving Score (HD-Score).



(a) Original Real-World Trajectories (b) Our Collected Multimodal Ground-Truth

Figure 3: **Our Multimodal Benchmark Statistics with Higher Coverage and Diversity.** Existing unimodal real-world trajectories lack diversity and coverage of modes (**left**). The collected trajectories, validated as both diverse and realistic (Sec.3.2, Table1), enable multimodal evaluation (**right**).

Table 1: **Realism of Collected Trajectories in Simulation.** Our simulated trajectories are multimodal and diverse, yet consistently include at least one mode that closely matches the real-world reference from nuScenes, achieving low L2 error at 3s (0.79m). Low Fréchet scores further demonstrate their realism across both photorealistic and digital twin environments.

Benchmark	3s L2 (m) ↓	Fréchet (m) ↓	NLL ↓
Driving in *Photorealistic* Simulation	**0.79**	1.46	3.48
Driving in *Virtual* Simulation	0.93	**1.11**	**3.19**

4.2 Driving in Simulation Produces Realistic Trajectories

It is crucial to evaluate not only trajectory diversity (e.g., Fig.3) but also realism. To assess this, we perform quantitative comparisons between trajectories collected from our photorealistic simulation, the original nuScenes ground truth, and our high-fidelity virtual environment. In Table1, we compare 15 trajectories per scene from each simulation setup to the single ground-truth trajectory provided by nuScenes. A realistic simulation with sufficient coverage should produce at least one trajectory closely matching the real-world log. For each scene, we compute minimum L2 distance, Fréchet distance, and the NLL between the simulated and ground-truth trajectories. We omit JSD in this evaluation, as the ground truth is unimodal and does not support distributional comparison.

Photorealistic Simulation vs. Real-World Driving: We quantify the quality of our collected trajectories by directly comparing (i.e., as predictions) against the original real-world ground truth in Table 1. While augmented trajectories from our simulation demonstrate significantly greater diversity, enhancing the coverage of feasible plans, we also observe low error rates, indicating the high realism of the collected trajectories. The low minL2 (i.e., 3s L2 error) measures how closely *one of the trajectories* aligns with the original driving trajectory for each sample. We note that our evaluation set aligns with the standard real-world nuScenes motion planning benchmark, such that these notably low error rates can be compared with planning-based metrics such as for UniAD, e.g., 0.79m vs. 1.65m 3s L2 in [26].

Photorealistic vs Virtual: When evaluating the virtual environment trajectories, we generally observe lower errors in Table 1, though each simulation setting presents its own trade-offs. Notably, trajectories collected from both the photorealistic and virtual environments achieve a zero collision rate and zero curb collision rate [9], as episodes are reset upon collision. While the virtual CARLA environment yields strong results, it requires significant manual effort to design scenes and agent behaviors. Moreover, crafting realistic reactive dynamics remains challenging in both environments. Interestingly, Table 1 shows that trajectories from our rendering-based simulation achieve a lower 3s L2 error (0.79m) compared to those from the CARLA-based virtual environment (0.93m), indicating that at least one of the collected trajectories more closely matches the real-world ground truth. This is a remarkably low error compared to prior models on this benchmark, all exceeding 1.41m 3s

Table 2: **Planning Performance Comparison Leveraging Multimodal Ground-Truth.** We compare models using single-annotation from nuScenes and our multimodal annotations. Results show notable re-ordering in multimodal metrics (e.g., VAD vs. UniAD), where unimodal L2 penalize plausible predictions, underscoring the need for multimodal evaluation. BranchOut significantly enhances multimodal reasoning by capturing plausible driving behaviors while preventing mode collapse.

Method	# Params (M)	L2 (m) ↓				Fréchet ↓	NLL ↓	Speed JSD ↓
		1s	2s	3s	Avg.			
IDM	-	3.98	8.21	12.65	8.28	10.04	-	-
Ego-MLP [9]	0.2	0.27	0.31	0.40	0.33	0.73	8.99	0.50
OccWorld [83]	58.0	0.44	1.12	2.08	1.21	2.65	12.53	0.52
UniAD [26]	55.7	0.46	0.94	1.65	1.02	2.60	10.86	0.45
VAD-Tiny [27]	39.6	0.51	1.04	1.76	1.11	2.65	7.22	0.43
VAD-Base [27]	58.1	0.46	0.98	1.69	1.04	2.50	7.72	0.41
DiffusionDrive [17]	60.0	0.31	0.82	1.58	0.90	2.41	3.95	0.39
BranchOut w/o Command	40.8	0.35	0.90	1.70	0.98	2.52	5.01	0.41
BranchOut w/o GMM	41.9	0.36	0.82	1.51	0.90	2.43	4.11	0.40
BranchOut w/o Diffusion	41.2	0.37	0.80	1.45	0.87	2.35	3.80	0.37
BranchOut w/ Classifier Guidance	41.9	**0.30**	**0.74**	1.51	0.85	2.46	4.02	0.39
BranchOut	41.9	0.31	0.76	**1.41**	**0.83**	**2.29**	**3.72**	**0.36**
BranchOut w/ EgoStatus	42.2	0.21	0.63	1.40	0.75	2.35	3.79	0.38
BranchOut w/ EgoHistory	42.4	0.26	0.65	1.30	0.74	2.25	3.74	0.35

L2 error (Table 2). These findings support the conclusion that reconstruction-based simulation can serve as a viable and realistic source of human driving behavior.

4.3 Driving Policy Evaluation

We re-evaluate standard driving policy baselines using our enriched benchmark, which captures a broader set of feasible future trajectories for each scene. This allows us to analyze how existing models perform under a more complete characterization of the multimodal decision space. By moving beyond single ground-truth evaluation, we expose failure modes that are otherwise masked—highlighting the limitations of conventional metrics in capturing the diversity and plausibility of real-world driving behavior. All evaluations are conducted on the full nuScenes validation split and on HUGSIM across difficulty levels, for open-loop and closed-loop respectively.

Baselines: We comprehensively evaluate the planning performance of the proposed BranchOut against state-of-the-art vision-based planners, including UniAD [26], VAD [27], OccWorld [83], and DiffusionDrive [17], as well as Ego-MLP [9], a perception-free planner. VAD-Tiny denotes a lightweight variant of VAD-Base, with reduced BEV query counts and fewer encoder/decoder layers. DiffusionDrive employs a truncated diffusion policy that denoises an anchored Gaussian distribution into a multimodal action distribution. For Ego-MLP, we follow [9] and remove the history trajectory input from AD-MLP [7] to prevent label leakage, resulting in an ego-state-only model that is not directly comparable to perception-based planners. We use publicly available code and pre-trained weights for the remaining baselines. Additionally, we include the Intelligent Driver Model (IDM), a rule-based car-following model, as a classical reactive baseline for comparison.

Revisiting Open-Loop Planner Evaluation: Table 2 revisits state-of-the-art planner performance within a multimodal context using our enriched driving trajectory dataset. Under standard L2 error, UniAD outperforms other unimodal planners, achieving 4.1% lower error than VAD-Base. However, the trend reverses when evaluated with multiple ground truths under Fréchet distance, where VAD-Base surpasses UniAD by a significant margin. This demonstrates that although the prediction of VAD-Base may diverge from the single ground truth (nuScenes), they remain behaviorally plausible and better aligned with natural human driving behavior. Similarly, VAD-Tiny and OccWorld, which show higher L2 error than UniAD by 8.8% and 18.6% in L2, achieve comparable performance under Fréchet evaluation. These findings highlight a key limitation of unimodal evalu-

Table 3: **Closed-Loop Evaluation in HUGSIM.** Results are averaged across all difficulty levels. BranchOut demonstrates robust route completion, resulting in the best overall HD-Score.

Method	NC ↑	DAC ↑	TTC ↑	COM ↑	R_c ↑	HD-Score ↑
Ego-MLP [9]	0.48	0.77	0.39	0.80	0.21	0.08
UniAD [26]	0.70	0.95	0.58	0.81	0.34	0.25
VAD-Tiny [27]	0.44	0.80	0.34	**1.00**	0.32	0.11
VAD-Base [27]	0.56	0.87	0.43	**1.00**	0.28	0.14
DiffusionDrive [17]	0.56	0.67	0.48	0.80	0.24	0.10
BranchOut	**0.76**	**0.99**	**0.69**	**1.00**	**0.58**	**0.47**

ation, which penalizes diverse yet valid predictions, and underscore the importance of incorporating multimodality into planning evaluation to better assess trajectory quality.

Multimodal Planner Evaluation: We evaluate the quality and diversity of predicted trajectory distributions across planning models. As shown in Table 2, stochastic planners—such as DiffusionDrive and our proposed BranchOut—consistently outperform unimodal baselines across all metrics. Notably, BranchOut achieves this with a compact sampling strategy of a single trajectory per high-level command (three in total), whereas DiffusionDrive requires over 20 samples to reach comparable performance. Under single ground-truth evaluation, BranchOut reduces L2 error by 7.8%, and under multimodal evaluation, it improves Fréchet distance by 4.98%. Beyond geometric accuracy, BranchOut also excels on distribution-based metrics, indicating not only greater trajectory diversity but also better alignment with human-like behavior. Specifically, it achieves a 5.8% lower NLL and a 7.7% reduction in Speed JSD compared to DiffusionDrive. These results underscore the effectiveness of our diffusion-based framework, particularly when paired with a compact GMM head, in generating diverse, efficient, and realistic trajectory predictions.

Closed-Loop Evaluation: As shown in Table 3, BranchOut demonstrates robustness across all metrics, outperforming both unimodal (e.g., UniAD) and multimodal (e.g., DiffusionDrive) baselines in terms of safe driving and goal completion. Notably, BranchOut achieves a superior R_c, with a 70.5% improvement over UniAD. Thus, our GMM-based diffusion modeling effectively reasons over feasible plans in closed-loop, predicting multimodal trajectory distributions that enable adaptive decision-making. Our findings show that BranchOut can dynamically respond to agent interactions, sudden obstacles, and unpredictable environmental changes, achieving more robust and safe driving.

5 Conclusion

We present BranchOut, a GMM-based branched diffusion planner that explicitly models the rich multimodality of human driving. By integrating a GMM head into a diffusion-based framework, BranchOut efficiently generates diverse and plausible future trajectories, achieving state-of-the-art performance across both error-based and distribution-based metrics. Beyond modeling, we identify key limitations in current evaluation protocols, which often rely on a single ground-truth trajectory and may penalize diverse yet plausible alternatives. Our analysis shows that incorporating multiple ground truths from an introduced, human-aware benchmark can reverse performance trends, emphasizing the importance of assessing behavioral diversity in planner evaluation. To support this, we introduce a human-in-the-loop, photorealistic simulation framework with kinematics and collision feedback, enabling scalable collection of diverse, reactive trajectories. Our enriched benchmark more accurately reflects real-world driving variability and reveals that standard error-based metrics struggle to fully capture alignment with human behavior, highlighting the importance of multimodal evaluation protocols. BranchOut demonstrates that multimodal modeling and evaluation are not only tractable, but essential for building safe, realistic, and human-aligned autonomous driving systems.

Acknowledgments: We thank the Red Hat Collaboratory (award #2024-01-RH07, #2025-01-RH04) and the National Science Foundation (IIS-2152077) for supporting this research.

6 Limitations

While our work establishes a foundation for modeling and evaluating planning in multimodal contexts, several limitations remain.

Simulation Fidelity: Our human-in-the-loop simulation builds upon HUGSIM, utilizing agents based on the IDM, a simplistic speed-based car-following policy. This often results in oversimplified or unrealistic behavior, particularly in complex scenes, and can lead to scenarios where human driving becomes infeasible. Enhancing the realism of background agents is critical for improving simulation fidelity.

Benchmark Coverage: Although we take a significant step toward benchmarking multimodality by augmenting real-world data through scalable human-driven collection, fully capturing the space of plausible behaviors remains challenging. Scenarios involving nuanced human-vehicle interactions or dynamic complexities in dense urban settings are particularly underrepresented and offer promising directions for further expansion.

Planner Robustness: While BranchOut achieves robust performance across all metrics, we observe occasional failure modes, such as aggressive driving in high-speed merge scenarios and subtle lateral oscillations during narrow lane following. We analyze these cases in the supplementary material and identify them as opportunities for future improvement.

Addressing these limitations is essential for advancing the development of autonomous driving systems that are both realistic and human-aligned.

References

[1] F. Sagberg, Selpi, G. F. Bianchi Piccinini, and J. Engström. A review of research on driving styles and road safety. *Human Factors*, 2015.

[2] O. Taubman-Ben-Ari and D. Yehiel. Driving styles and their associations with personality and motivation. *Accident Analysis & Prevention*, 2012.

[3] H. A. Deery. Hazard and risk perception among young novice drivers. *Journal of Safety Research*, 1999.

[4] J. Duncan, P. Williams, and I. Brown. Components of driving skill: experience does not mean expertise. *Ergonomics*, 1991.

[5] S. Casas, C. Gulino, S. Suo, and R. Urtasun. The importance of prior knowledge in precise multimodal prediction. In *IROS*, 2020.

[6] Introducing Drivership: A New Framework for Good Driving. https://waymo.com/blog/2025/02/introducing-drivership-a-new-framework-for-good-driving.

[7] J.-T. Zhai, Z. Feng, J. Du, Y. Mao, J.-J. Liu, Z. Tan, Y. Zhang, X. Ye, and J. Wang. Rethinking the open-loop evaluation of end-to-end autonomous driving in nuscenes. *arXiv preprint arXiv:2305.10430*, 2023.

[8] D. Dauner, M. Hallgarten, A. Geiger, and K. Chitta. Parting with misconceptions about learning-based vehicle motion planning. *arXiv preprint arXiv:2306.07962*, 2023.

[9] 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 *CVPR*, 2024.

[10] D. Dauner, M. Hallgarten, T. Li, X. Weng, Z. Huang, Z. Yang, H. Li, I. Gilitschenski, B. Ivanovic, M. Pavone, et al. Navsim: Data-driven non-reactive autonomous vehicle simulation and benchmarking. *NeurIPS*, 2024.

[11] C. Zhang, R. Guo, W. Zeng, Y. Xiong, B. Dai, R. Hu, M. Ren, and R. Urtasun. Rethinking closed-loop training for autonomous driving. In *ECCV*, 2022.

[12] N. Hanselmann, K. Renz, K. Chitta, A. Bhattacharyya, and A. Geiger. King: Generating safety-critical driving scenarios for robust imitation via kinematics gradients. In *ECCV*, 2022.

[13] S. Khandelwal, W. Qi, J. Singh, A. Hartnett, and D. Ramanan. What-if motion prediction for autonomous driving. *arXiv preprint arXiv:2008.10587*, 2020.

[14] W. Peebles and S. Xie. Scalable diffusion models with transformers. In *ICCV*, 2023.

[15] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 2020.

[16] J. Zhang, Z. Huang, and E. Ohn-Bar. Coaching a teachable student. In *CVPR*, 2023.

[17] B. Liao, S. Chen, H. Yin, B. Jiang, C. Wang, S. Yan, X. Zhang, X. Li, Y. Zhang, Q. Zhang, Z. Yang, Y. Chen, J. Wang, S. Manivasagam, W.-C. Ma, A. J. Yang, and R. Urtasun. Diffusiondrive: Truncated diffusion model for end-to-end autonomous driving. In *CVPR*, 2025.

[18] C. Chi, Z. Xu, S. Feng, E. Cousineau, Y. Du, B. Burchfiel, R. Tedrake, and S. Song. Diffusion policy: Visuomotor policy learning via action diffusion. *IJRR*, 2023.

[19] Y. Zheng, R. Liang, K. Zheng, J. Zheng, L. Mao, J. Li, W. Gu, R. Ai, S. E. Li, X. Zhan, et al. Diffusion-based planning for autonomous driving with flexible guidance. *arXiv preprint arXiv:2501.15564*, 2025.

[20] B. Yang, H. Su, N. Gkanatsios, T.-W. Ke, A. Jain, J. Schneider, and K. Fragkiadaki. Diffusion-es: Gradient-free planning with diffusion for autonomous driving and zero-shot instruction following. *arXiv preprint arXiv:2402.06559*, 2024.

[21] H. J. Kim and E. Ohn-Bar. Motion diversification networks. In *CVPR*, 2024.

[22] C. Jiang, A. Cornman, C. Park, B. Sapp, Y. Zhou, D. Anguelov, et al. Motiondiffuser: Controllable multi-agent motion prediction using diffusion. In *CVPR*, 2023.

[23] M. Janner, Y. Du, J. B. Tenenbaum, and S. Levine. Planning with diffusion for flexible behavior synthesis. *arXiv preprint arXiv:2205.09991*, 2022.

[24] Z. Huang, H. Liu, and C. Lv. Gameformer: Game-theoretic modeling and learning of transformer-based interactive prediction and planning for autonomous driving. In *ICCV*, 2023.

[25] K. Chitta, A. Prakash, B. Jaeger, Z. Yu, K. Renz, and A. Geiger. Transfuser: Imitation with transformer-based sensor fusion for autonomous driving. *PAMI*, 2022.

[26] 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 *CVPR*, 2023.

[27] 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. *ICCV*, 2023.

[28] J. Zhang, Z. Huang, A. Ray, and E. Ohn-Bar. Feedback-guided autonomous driving. In *CVPR*, 2024.

[29] S. Chen, B. Jiang, H. Gao, B. Liao, Q. Xu, Q. Zhang, C. Huang, W. Liu, and X. Wang. Vadv2: End-to-end vectorized autonomous driving via probabilistic planning. *arXiv preprint arXiv:2402.13243*, 2024.

[30] C. Tang and R. R. Salakhutdinov. Multiple futures prediction. *NeurIPS*, 2019.

[31] N. Rhinehart, K. M. Kitani, and P. Vernaza. R2p2: A reparameterized pushforward policy for diverse, precise generative path forecasting. In *ECCV*, 2018.

[32] H. Caesar, J. Kabzan, K. S. Tan, W. K. Fong, E. Wolff, A. Lang, L. Fletcher, O. Beijbom, and S. Omari. nuPlan: A closed-loop ml-based planning benchmark for autonomous vehicles. *arXiv preprint arXiv:2106.11810*, 2021.

[33] S. K. Aithal, P. Maini, Z. Lipton, and J. Z. Kolter. Understanding hallucinations in diffusion models through mode interpolation. *NeurIPS*, 2024.

[34] Y. Qin, H. Zheng, J. Yao, M. Zhou, and Y. Zhang. Class-balancing diffusion models. In *CVPR*, 2023.

[35] H. J. Kim, K. Sengupta, M. Kuribayashi, H. Kacorri, and E. Ohn-Bar. Text to blind motion. *NeurIPS*, 2024.

[36] A. Sedlmeier, M. Kölle, R. Müller, L. Baudrexel, and C. Linnhoff-Popien. Quantifying multi-modality in world models. *arXiv preprint arXiv:2112.07263*, 2021.

[37] R. Barceló, C. Alcázar, and F. Tobar. Avoiding mode collapse in diffusion models fine-tuned with reinforcement learning. *arXiv preprint arXiv:2410.08315*, 2024.

[38] R. Zhu, P. Huang, E. Ohn-Bar, and V. Saligrama. Learning to drive anywhere. *CoRL*, 2023.

[39] F. Codevilla, A. M. Lopez, V. Koltun, and A. Dosovitskiy. On offline evaluation of vision-based driving models. In *ECCV*, 2018.

[40] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *CVPR*, 2020.

[41] P. Hu, A. Huang, J. Dolan, D. Held, and D. Ramanan. Safe local motion planning with self-supervised freespace forecasting. In *CVPR*, 2021.

[42] L. Lai, E. Ohn-Bar, S. Arora, and J. S. K. Yi. Uncertainty-guided never-ending learning to drive. In *CVPR*, 2024.

[43] S. Suo, K. Wong, J. Xu, J. Tu, A. Cui, S. Casas, and R. Urtasun. Mixsim: A hierarchical framework for mixed reality traffic simulation. In *CVPR*, 2023.

[44] H. Zhou, L. Lin, J. Wang, Y. Lu, D. Bai, B. Liu, Y. Wang, A. Geiger, and Y. Liao. Hugsim: A real-time, photo-realistic and closed-loop simulator for autonomous driving. 2024.

[45] N. Montali, J. Lambert, P. Mougin, A. Kuefler, N. Rhinehart, M. Li, C. Gulino, T. Emrich, Z. Yang, S. Whiteson, et al. The waymo open sim agents challenge. *NeurIPS*, 2023.

[46] 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 *ECCV*, 2024.

[47] P. Abbeel and A. Y. Ng. Apprenticeship learning via inverse reinforcement learning. In *ICML*, 2004.

[48] N. Nayakanti, R. Al-Rfou, A. Zhou, K. Goel, K. S. Refaat, and B. Sapp. Wayformer: Motion forecasting via simple & efficient attention networks. In *ICRA*, 2023.

[49] L. Lai, Z. Yin, and E. Ohn-Bar. ZeroVO: Visual odometry with minimal assumptions. In *CVPR*, 2025.

[50] Y. Yuan, X. Weng, Y. Ou, and K. M. Kitani. Agentformer: Agent-aware transformers for socio-temporal multi-agent forecasting. In *ICCV*, 2021.

[51] P. Xu, J.-B. Hayet, and I. Karamouzas. SocialVAE: Human trajectory prediction using time-wise latents. In *ECCV*, 2022.

[52] L. Lai, Z. Shangguan, J. Zhang, and E. Ohn-Bar. XVO: Generalized visual odometry via cross-modal self-training. In *ICCV*, 2023.

[53] Y. Yuan and K. Kitani. Diverse trajectory forecasting with determinantal point processes. *ICLR*, 2020.

[54] A. Rudenko, L. Palmieri, M. Herman, K. M. Kitani, D. M. Gavrila, and K. O. Arras. Human motion trajectory prediction: A survey. *IJRR*, 2020.

[55] M. Bansal, A. Krizhevsky, and A. Ogale. Chauffeurnet: Learning to drive by imitating the best and synthesizing the worst. *arXiv preprint arXiv:1812.03079*, 2018.

[56] B. Jaeger, K. Chitta, and A. Geiger. Hidden biases of end-to-end driving models. In *ICCV*, 2023.

[57] D. Chen, V. Koltun, and P. Krähenbühl. Learning to drive from a world on rails. In *ICCV*, 2021.

[58] G. Barquero, S. Escalera, and C. Palmero. Belfusion: Latent diffusion for behavior-driven human motion prediction. In *ICCV*, 2023.

[59] S. Xu, Y.-X. Wang, and L.-Y. Gui. Diverse human motion prediction guided by multi-level spatial-temporal anchors. In *ECCV*, 2022.

[60] P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *CVPR*, 2020.

[61] S. Ettinger, S. Cheng, B. Caine, C. Liu, H. Zhao, S. Pradhan, Y. Chai, B. Sapp, C. R. Qi, Y. Zhou, et al. Large scale interactive motion forecasting for autonomous driving: The waymo open motion dataset. In *ICCV*, 2021.

[62] J. Mao, Y. Qian, J. Ye, H. Zhao, and Y. Wang. Gpt-driver: Learning to drive with gpt. *arXiv preprint arXiv:2310.01415*, 2023.

[63] B. Ivanovic and M. Pavone. Rethinking trajectory forecasting evaluation. *arXiv preprint arXiv:2107.10297*, 2021.

[64] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 2021.

[65] M. Tancik, V. Casser, X. Yan, S. Pradhan, B. Mildenhall, P. P. Srinivasan, J. T. Barron, and H. Kretzschmar. Block-nerf: Scalable large scene neural view synthesis. In *CVPR*, 2022.

[66] K. Deng, A. Liu, J.-Y. Zhu, and D. Ramanan. Depth-supervised NeRF: Fewer views and faster training for free. In *CVPR*, 2022.

[67] M. M. Johari, Y. Lepoittevin, and F. Fleuret. Geonerf: Generalizing nerf with geometry priors. In *CVPR*, 2022.

[68] J. Yang, B. Ivanovic, O. Litany, X. Weng, S. W. Kim, B. Li, T. Che, D. Xu, S. Fidler, M. Pavone, et al. Emernerf: Emergent spatial-temporal scene decomposition via self-supervision. *ICLR*, 2024.

[69] X. Zhou, Z. Lin, X. Shan, Y. Wang, D. Sun, and M.-H. Yang. DrivingGaussian: Composite gaussian splatting for surrounding dynamic autonomous driving scenes. In *CVPR*, 2024.

[70] Z. Chen, J. Yang, J. Huang, R. d. Lutio, J. M. Esturo, B. Ivanovic, O. Litany, Z. Gojcic, S. Fidler, M. Pavone, L. Song, and Y. Wang. Omnire: Omni urban scene reconstruction. *arXiv preprint arXiv:2408.16760*, 2024.

[71] D. A. Reynolds et al. Gaussian mixture models. *Encyclopedia of biometrics*, 2009.

[72] C. M. Bishop and N. M. Nasrabadi. *Pattern Recognition and Machine Learning*. 2006.

[73] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, 2015.

[74] C. Lu, Y. Zhou, F. Bao, J. Chen, C. Li, and J. Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *NeurIPS*, 2022.

[75] C. Lu, Y. Zhou, F. Bao, J. Chen, C. Li, and J. Zhu. Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. *Machine Intelligence Research*, 2025.

[76] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *NeurIPS*, 2017.

[77] Z. Zhang, P. Karkus, M. Igl, W. Ding, Y. Chen, B. Ivanovic, and M. Pavone. Closed-loop supervised fine-tuning of tokenized traffic models. In *CVPR*, 2025.

[78] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun. Carla: An open urban driving simulator. In *CoRL*, 2017.

[79] 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 *ECCV*, 2022.

[80] T. Eiter and H. Mannila. Computing discrete fréchet distance. 1994.

[81] M. Igl, D. Kim, A. Kuefler, P. Mougin, P. Shah, K. Shiarlis, D. Anguelov, M. Palatucci, B. White, and S. Whiteson. Symphony: Learning realistic and diverse agents for autonomous driving simulation. In *ICRA*, 2022.

[82] Y. Gal and Z. Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *ICML*, 2016.

[83] W. Zheng, W. Chen, Y. Huang, B. Zhang, Y. Duan, and J. Lu. OccWorld: Learning a 3d occupancy world model for autonomous driving. In *ECCV*, 2023.