BranchOut: Capturing Realistic Multimodality in Autonomous Driving Decisions

Hee Jae Kim, Zekai Yin, Lei Lai, Jason Lee, and Eshed Ohn-Bar
Boston University

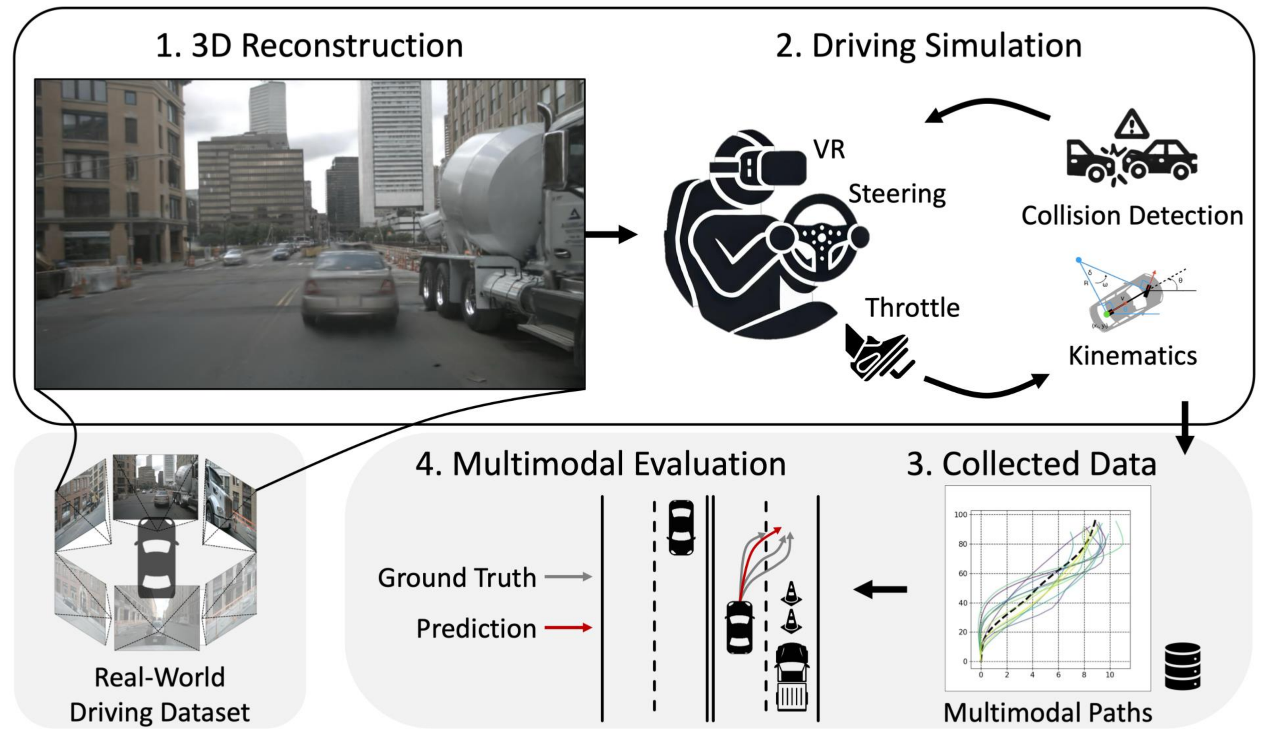
Overview

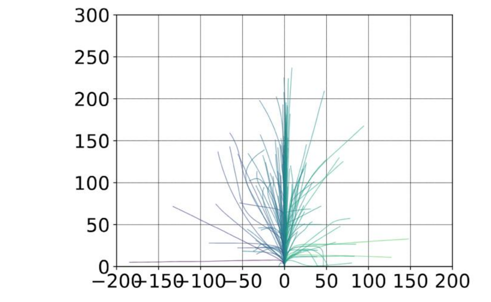
Motivation: Real-world driving involves multiple plausible decisions, making planning inherently multimodal. Yet, current planners often fail to capture diverse mode, often remaining deterministic or collapsing to a dominant mode. Moreover, existing datasets provide only a single annotation per scenario, which may penalize valid alternatives.

Contributions:

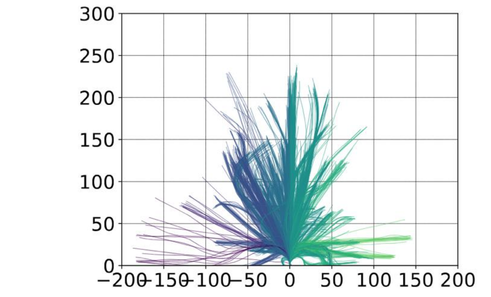
- We introduce **BranchOut**, a GMM-based diffusion planner that explicitly captures multimodal driving behaviors in an end-to-end manner.
- We present human-in-the-loop photorealistic simulation framework to collect diverse trajectories, enabling multimodal evaluation protocol.

Diverse and Realistic Simulation





(a) Original Real-World Trajectories

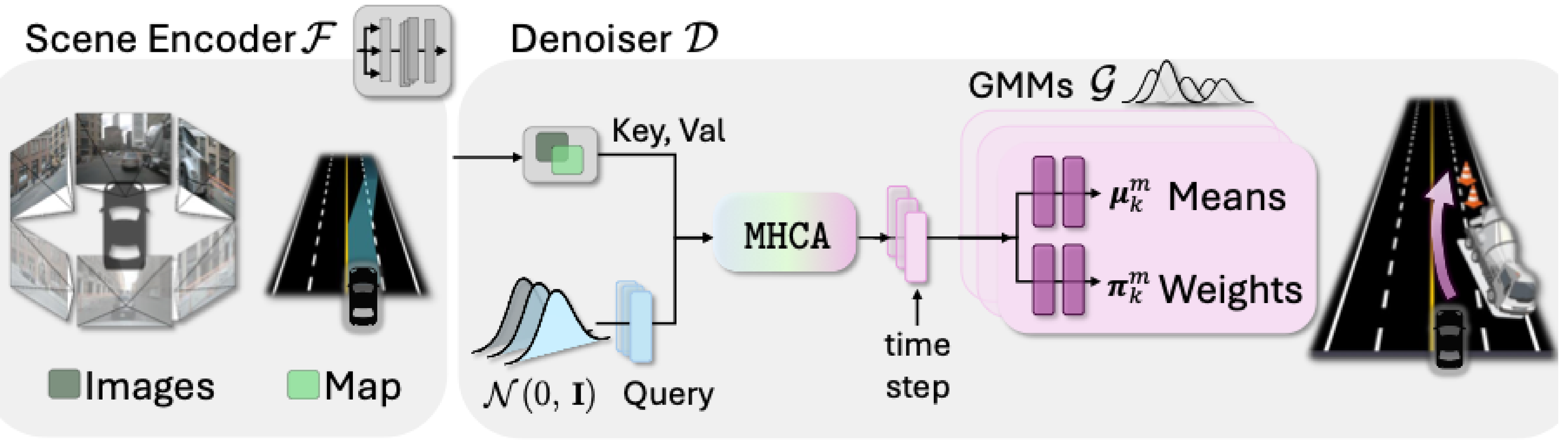


(b) Our Collected Multimodal Ground-Truth

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**

- We leverage a rendering-based photorealistic, reactive human-in-the-loop simulation with collision feedback, enabling scalable collection of diverse and realistic multimodal trajectories.
- Our simulated trajectories not only achieve high diversity but also closely match real-world logs (3s L2 = 0.79 m), compared to a digital twin (0.93 m), without the overhead of manual scene construction.

Method



Model Architecture of BranchOut

Diffusion Process models multimodal driving decisions by perturbing ground-truth trajectories $\mathbf{Y}_{\text{ego}}$ into noisy inputs $\mathbf{X}_{\text{ego}}^{(t)}$ using Gaussian noise $\mathbf{z} \sim \mathcal{N}(0, \mathbf{I})$ and a diffusion timestep $t \sim \mathcal{U}(0,1)$:

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

At inference, we start from Gaussian noise and solve the reverse diffusion ODE with a single-step DPM-Solver++.

Scene-Aware Diffusion Transformer embeds noisy trajectory queries into $\mathbf{P}$, fuses them with agent and map features via cross-attention, and modulates them with timestep embedding $\gamma(t)$:

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

$$\mathbf{P} \leftarrow \text{Scale}(\gamma(t)) \cdot \mathbf{P} + \text{Shift}(\gamma(t))$$

Branched GMM Head decodes scene-aware features $\mathbf{P}$ into $\boldsymbol{\mu}_k^m$ and $\boldsymbol{\pi}_k^m$, with each branch m corresponding to a navigation command and predicting K diverse modes to explicitly capture multiple plausible futures:

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

Loss Functions:

$$\mathcal{L} = \mathcal{L}_{\text{plan}} + \lambda_{\text{NLL}} \mathcal{L}_{\text{NLL}} + \lambda_c \mathcal{L}_{\text{constraints}}$$

- $\mathcal{L}_{\text{plan}}$: diffusion reconstruction loss
- $\mathcal{L}_{\text{NLL}}$: negative log-likelihood over GMM parameters
- $\mathcal{L}_{\text{constraints}}$: collision, boundary, and directional safety constraints

Quantitative Results

Open-Loop Evaluation on nuScenes

Method	# Params (M)	L2 (m) ↓ 1s	2s	3s	Avg.	Fréchet ↓	NLL ↓	Speed JSD ↓
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

Impact of Branched Decoder

Method	L2 (m) ↓ 1s	2s	3s	Avg.	Fréchet ↓	NLL ↓	Speed JSD ↓
BranchOut (Shared Head)	**0.31**	0.78	1.52	0.87	2.41	3.98	0.39
BranchOut (Ours)	**0.31**	**0.76**	**1.41**	**0.83**	**2.29**	**3.72**	**0.36**

Closed-loop Evaluation on HugSim

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**

Qualitative Results