MeanFuser: Fast One-Step Multi-Modal Trajectory Generation and Adaptive Reconstruction via MeanFlow for End-to-End Autonomous Driving

Supplementary Material

7. CARLA Longest6 Benchmark

CARLA Longest6 Benchmark: The CARLA Longest6 Benchmark, introduced by TransFuser[7], is designed to reduce computational resource consumption and testing time while ensuring a balanced distribution of routes across six towns, with six test routes selected from each town, resulting in a total of 36 routes averaging 1500 meters in length. The benchmark incorporates six distinct weather conditions and six different times of day, and its evaluation metric is the Driving Score (DS), computed as a weighted average of Route Completion (RC) penalized by the Infraction Score (IS).

Table 5. **Longest6 Benchmark Results.** We show the mean and std for all metrics (RC: Route Completion, IS: Infraction Score. DS: Driving Score).

Method	RC ↑	IS ↑	DS ↑
Latent TransFuser[7]	**95.18**$_{\pm 0.45}$	0.38$_{\pm 0.05}$	37.31$_{\pm 5.35}$
TransFuser[7]	93.38$_{\pm 1.20}$	0.50$_{\pm 0.06}$	47.30$_{\pm 5.72}$
DiffusionDrive[22]	94.16$_{\pm 1.46}$	0.69$_{\pm 0.02}$	64.27$_{\pm 2.43}$
MeanFuser (Ours)	94.65$_{\pm 1.32}$	**0.73**$_{\pm 0.05}$	**70.08**$_{\pm 3.20}$

To comprehensively evaluate model performance, we validate our approach on the CARLA Longest6 closed-loop benchmark. As shown in Tab. 5, we conduct three runs for each route to compute the mean and standard deviation. Experimental results demonstrate that the model achieves a Driving Score (DS) that surpasses the unimodal trajectory method TransFuser by 22.78 and outperforms the multi-modal method DiffusionDrive by 5.81, confirming its effectiveness and robustness in closed-loop testing. In Fig. 6, we visualize the planning outcomes of our model across diverse scenarios under both daytime and nighttime conditions.

8. Further Ablation Study

The ablation of the number of Gaussian components. As shown in Tab. 6, we present the performance of the model with different numbers of Gaussian components. The model achieves optimal performance with eight Gaussian components. Adding more components beyond this point does not improve results and may even cause a slight decrease in performance. This indicates that eight components provide sufficient capacity to model the trajectory distribution. Further increases lead to each component becoming data-starved, preventing the model from adequately learning the velocity field and resulting in unreliable velocity predictions.

Figure 6. **Visualization on the CARLA Longest6 benchmark.** The two images in the upper row and the two in the lower row showcase the planning outcomes, depicting daytime and nighttime conditions, respectively.

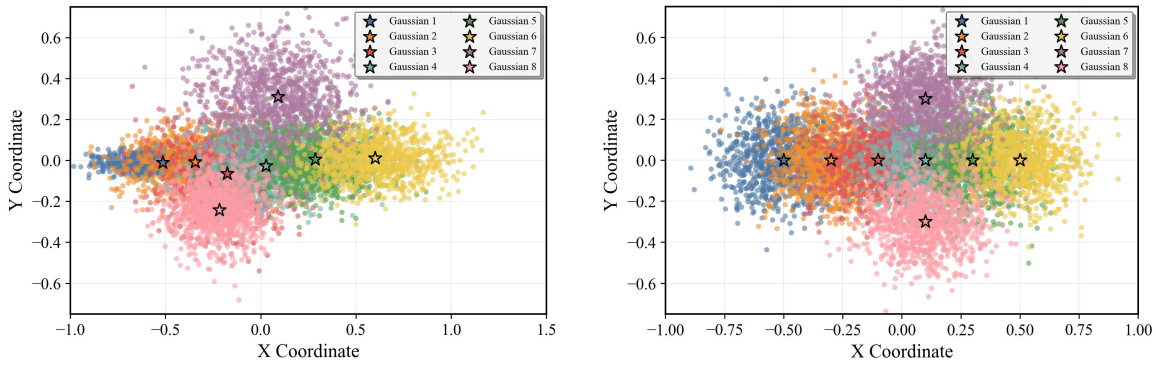
Table 6. **Number of Gaussian components and model performance.** $N_{gaussian}$ denotes the number of Gaussian components in the Gaussian Mixture Noise distribution.

$N_{gaussian}$	NC ↑	DAC ↑	TTC ↑	Comf. ↑	EP ↑	PDMS ↑
2	98.1	96.9	94.0	100	81.8	88.4$_{-0.5}$
8	98.6	97.0	95.0	100	82.8	89.0
16	98.1	97.3	93.8	99.9	83.3	88.8$_{-0.2}$
32	98.0	97.0	94.3	99.9	82.8	88.5$_{-0.5}$

Table 7. **Comparison of the GMN Generation Methods.**

Method	NC ↑	DAC ↑	TTC ↑	Comf. ↑	EP ↑	PDMS ↑
Data Clustering	98.6	97.0	95.0	100	82.8	89.0
Manual Design	98.2	97.0	94.1	100	83.0	88.6$_{-0.4}$

The ablation study of Gaussian Mixture Noise generation. To investigate the dependence of model performance on Gaussian Mixture Noise (GMN) generation strategies, we conduct a comprehensive ablation study. As summarized in Tab. 7, we quantitatively compare two GMN construction methods: one derived from clustering expert tra-



(a) GMN from Data Clustering

(b) GMN from Manual Design

Figure 7. **Visualization of alternative approaches for generating Gaussian Mixture Noise (GMN)**. (a) Mean and standard deviation are derived from clustered expert demonstrations in the training set. (b) Mean and standard deviation are obtained through manually design.

jectories in the navtrain dataset (detailed in Section 4.2), and the other manually designed. In our manual design approach, the mean of each Gaussian component is determined by simple heuristic rules while standard deviations are set to a unified fixed value. Experimental results demonstrate that when using manually designed GMN, model performance decreases by only 0.45% compared to that of the data-driven clustering approach. In contrast, the extreme case where all Gaussian components follow standard normal distributions leads to significant performance degradation. This confirms that our model's effectiveness does not rely on specific fixed datasets. Visual comparisons of the GMN generated by both methods are presented in Fig. 7.

Table 8. **Comparison of multimodality and performance.** (GMN: Gaussian Mixture Noise. K: number of multimodal trajectories; $\mathcal{D}$: multimodality metric.)

Method	GMN	K	PDMS ↑	$\mathcal{D}$ ↑	$\mathcal{M}_{DP}$ ↑
TransFuser	-	8	94.0	0.0	-
MeanFuser	×	8	88.3	0.25	22.07
MeanFuser	√	8	**89.0**	**0.30**	**26.70**$_{+20.84\%}$

The ablation study of Multi-modal planning performence.

We employ a mean Intersection-over-Union (mIoU)-based metric $\mathcal{D}$ to quantify the multimodality of planning outcomes. For a set of K trajectories $\{\tau_k\}_{k=1}^{K}$, the metric is defined as:

$$\mathcal{D} = 1 - \frac{1}{T_f} \sum_{i=1}^{T_f} \frac{\bigcap_{k=1}^{K} \text{Area}(\hat{\tau}_{ki})}{\bigcup_{k=1}^{K} \text{Area}(\hat{\tau}_{ki})}, \tag{16}$$

where T_f denotes the prediction horizon, $\hat{\tau}_{ki}$ represents the bounding box of the k-th trajectory at timestep i, and the operators $\bigcap$ and $\bigcup$ calculate the intersection and union of areas across all K trajectories at each timestep i, respectively. A higher value of $\mathcal{D}$ indicates greater diversity among the predicted trajectories. For an intuitive understanding, Fig. 8 visualizes the bounding boxes at different timesteps along with their intersections.

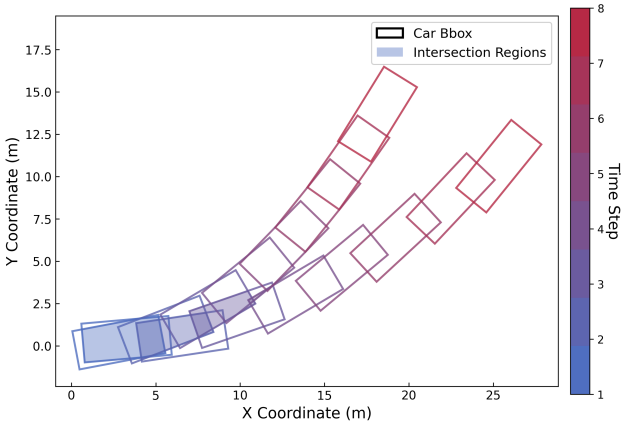


Figure 8. Visualization of the intersection and union dynamics of bounding boxes (Car Bbox) for ego trajectories across different timesteps.

To ensure that the observed diversity does not stem from model error, such as trajectory divergence, we introduce the composite metric:

$$\mathcal{M}_{\mathbf{DP}} = \mathcal{D} \times PDMS, \tag{17}$$

which provides a unified measure that considers both planning performance and trajectory diversity.

The ablation studies in Tab. 8 demonstrate that the GMN not only enhances the primary performance metric (PDMS) but also significantly increases the diversity ($\mathcal{D}$) of the generated trajectory proposals, resulting in a 20.84% improvement in the comprehensive evaluation metric $\mathcal{M}_{DP}$.