Technical Report for UniPlan

Lan Feng
EPFL
lan.feng@epfl.ch

Alexandre Alahi
EPFL
alexandre.alahi@epfl.ch

1. Introduction

Autonomous driving systems are traditionally designed with a modular architecture, involving separate components for perception, prediction, and planning. However, the emergence of powerful foundation models has sparked increasing interest in end-to-end approaches that directly map raw sensory inputs to driving actions. Such methods have demonstrated promising capabilities in robustness and reasoning.

To evaluate the effectiveness of these approaches in handling complex, real-world challenges, the 2025 Waymo Open Dataset Vision-based End-to-End (WOD-E2E) Driving Challenge presents a new benchmark focused on long-tail driving scenarios—rare but impactful situations such as detours around marathons, collisions with fallen scooter riders, or interactions with emergency vehicles. The dataset includes 4021 20-second driving segments, with 2037 for training and 479 for validation. The test set is used for final evaluation and includes only partial observations for forecasting. Participants are required to predict 5-second waypoint trajectories in bird-eye-view coordinates using input from 8 surrounding cameras, past vehicle poses, and a route plan. Submissions are scored primarily using the Rater Feedback Score (RFS), with Average Displacement Error (ADE) as a tie-breaker.

In this report, we present **UniPlan**, a unified end-to-end planning framework that leverages large-scale public driving datasets beyond WOD-E2E dataset to improve generalization in rare, long-tail scenarios. Our method achieves 3rd place in the 2025 challenge leaderboard without relying on expensive MLLM-based auto-labeling techniques.

2. Method

2.1. Model Architecture

Our model builds on **DiffusionDrive** [3], which introduces a *truncated diffusion policy* for efficient and diverse trajectory generation. Its architecture is illstrated in Fig. 1. Unlike vanilla diffusion methods that denoise from random Gaussian noise, DiffusionDrive starts from an *anchored Gaussian distribution*—generated around prior multi-mode trajectory anchors—and applies only two denoising steps to produce final trajectory predictions. The key advantages include:

- Significantly fewer denoising steps (2 instead of 20), enabling real-time inference.
- Anchor-guided sampling ensures mode diversity and avoids mode collapse.
- A cascade diffusion decoder refines the prediction in each denoising step via cross-attention over BEV and agent/map queries.

During training, each noisy trajectory is paired with its anchor and denoised using a transformer-based diffusion decoder conditioned on scene context. The model outputs both trajectory coordinates and confidence scores.

Diffusion Decoder We follow the diffusion-based trajectory generation design introduced in DiffusionDrive [3]. The method begins by sampling a set of noisy trajectories $\{\hat{\tau}_k\}_{k=1}^{N_{\text{infer}}}$ from an anchored Gaussian distribution. For each trajectory, deformable spatial cross-attention is applied to retrieve features from Bird's Eye View (BEV) or Perspective View (PV), based on the trajectory coordinates.

Subsequently, the model performs cross-attention between the trajectory features and queries representing agents or maps, as extracted from the perception module, followed by a feed-forward network (FFN). To incorporate the diffusion timestep, the architecture includes a *Timestep Modulation* layer, followed by a multi-layer perceptron (MLP) that predicts a confidence score and a spatial offset from the initial noisy coordinates.

This process is repeated across cascade diffusion decoder layers, enabling iterative denoising. During inference, the decoder is reused with shared weights across denoising steps. The final trajectory is selected based on the highest predicted confidence score.

2.2. Data Processing

NuPlan dataset. We build a dataset of 90k samples from the 100-hour nuPlan dataset [1] using a sliding window of 9s (4s history, 5s future) sampled every 1s. Samples with final displacement error $> 0.5m$ from a constant velocity

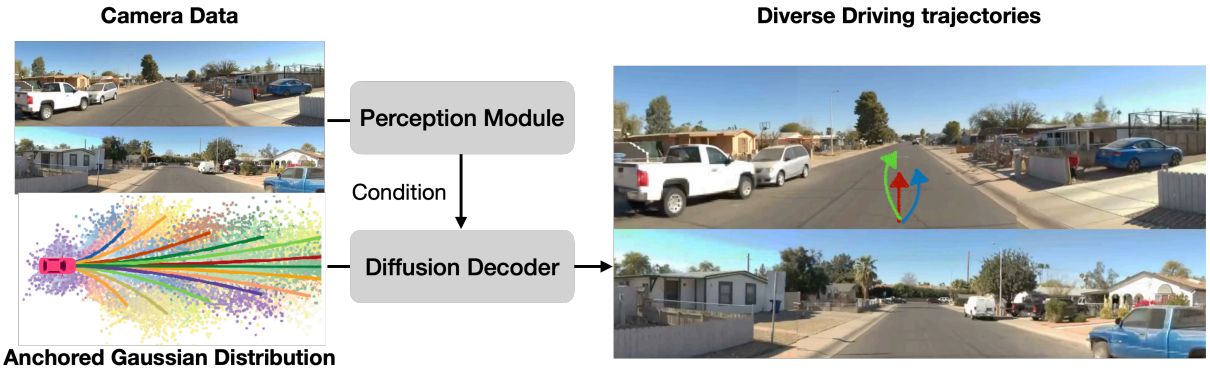

Figure 1. Overall architecture of DiffusionDrive. DiffusionDrive can integrate various existing perception modules and sensor inputs. The designed diffusion decoder takes the sampled noisy trajectories from an anchored Gaussian distribution as input and progressively denoises them with enhanced interactions with the conditional scene context in a cascade manner to generate the final predictions.

model are retained. There is also 10% possibility to keep the samples with displacement error $< 0.5m$ to improve data diversity.

WOD-E2E dataset. We use a similar filtering strategy to create 35k training and 10k validation samples. to generate new anchors based on the WOD-E2E dataset for the training and inference of the model.

DiffusionDrive was originally proposed in the Navsim benchmark [2], to adapt DiffusionDrive to the WOD-E2E challenge:

- We use K-Means with 20 clusters to generate new planner anchors using WOD-E2E data for 5-second prediction horizons (vs. 4s in nuPlan).
- We use the original ego status feature vector (velocity, acceleration, command).

Camera feature alignment. NuPlan images (1920x1120) are cropped (L0, F0, R0 views), concatenated to form a 4:1 aspect ratio image, and resized to 1024x256. WOD-E2E camera data is directly concatenated with the front 3 cameras and resizing to the same shape.

2.3. Training Setup

We train on a single compute node with **4× H100 GPUs and 360GB RAM**. Each GPU processes a batch size of 64, giving an effective batch size of 256. We use the AdamW optimizer with a learning rate of 3e-4 and a Warmup Cosine LR scheduler with the following settings:

- Warm-up for 3 epochs, followed by cosine decay.
- Training runs for 100 epochs.
- Final model is selected from the last checkpoint.

Detailed parameters can also be found in Tab. 2.

Compute Time. Joint training (WOD-E2E + NuPlan) takes approximately 6 hours. WOD-E2E-only training requires 1.5 hours.

2.4. Inference Strategy

We train 4 DiffusionDrive models with different seeds:

- One model on the full dataset (WOD-E2E + nuPlan / WOD-E2E-only)
- Three models trained on 80% random subsets of the full dataset

During inference, each of the 4 trained DiffusionDrive models generates 20 candidate trajectories from the anchored Gaussian distribution using 2 denoising steps, as in the original paper. These are scored via the model's confidence head. We collect 80 total candidates and select the highest-scoring trajectory as the final output.

3. Experiments

3.1. Setup

We evaluate the following three configurations:

1. Setting A: WOD-E2E-only training without ensemble
2. Setting B: WOD-E2E-only training with ensemble
3. Setting C: Joint training on WOD-E2E + nuPlan with ensemble

Performance is measured using the Rater Feedback Score (RFS) and ADE on the WOD-E2E Challenge leaderboard.

3.2. Results and Analysis

Resutls are listed in Tab. 1. Our experiments yield the following insights:

- **Model ensemble** significantly boosts RFS, demonstrating the benefit of aggregating predictions from multiple models.
- **Joint training with nuPlan data** slightly improves average RFS overall, with notable improvements in specific long-tail categories.
- Categories that showed significant gains include: *cyclist interactions*, *cut-in maneuvers*, *single-lane adjustments*,

Table 1. Performance Comparison Across Settings

Metric Name	Setting C	Setting B	Setting A
Average Score	**7.685**	7.683	7.632
Construction Score	8.157	8.324	**8.566**
Intersection Score	**7.787**	7.846	7.854
Pedestrian Score	7.674	**7.728**	7.506
Cyclist Score	**7.564**	7.332	7.430
Multi Lane Maneuver Score	7.550	**7.566**	7.542
Single Lane Maneuver Score	**8.247**	8.122	7.956
Cut In Score	**8.003**	7.744	7.624
Foreign Object Debris Score	7.833	7.927	**7.922**
Special Vehicle Score	**7.823**	7.646	7.640
Spotlight Score	6.723	**6.775**	6.602
Others Score	7.171	**7.509**	7.308
ADE @ 3s	1.298	1.293	**1.271**
ADE @ 5s	2.926	2.936	**2.857**

and *special vehicles*, confirming the benefit of incorporating diverse public datasets.

4. Conclusion

We present UniPlan, a unified framework for scalable end-to-end planning by leveraging large-scale public driving datasets. Our approach achieves competitive results on the WOD-E2E long-tail challenge without relying on expensive foundation models. The results highlight the promise of scalable, data-centric approaches for improving autonomous vehicle robustness.

5. More Training Details

We provide more training details in Tab. 2 for the reproduction of the leaderboard performance.

Table 2. Training Configuration for UniPlan

Hyperparameter	Value
Batch Size	256
Learning Rate	3e-4
Optimizer	AdamW
Learning Rate Scheduler	WarmupCosLR
Minimum Learning Rate	1e-6
Total Epochs	100
Warmup Epochs	3

References

[1] Holger Caesar, Juraj Kabzan, Kok Seang Tan, Whye Kit Fong, Eric Wolff, Alex Lang, Luke Fletcher, Oscar Beijbom, and Sammy Omari. nuplan: A closed-loop ml-based planning benchmark for autonomous vehicles. *arXiv preprint arXiv:2106.11810*, 2021. 1

[2] Daniel Dauner, Marcel Hallgarten, Tianyu Li, Xinshuo Weng, Zhiyu Huang, Zetong Yang, Hongyang Li, Igor Gilitschenski, Boris Ivanovic, Marco Pavone, et al. Navsim: Data-driven non-reactive autonomous vehicle simulation and benchmarking. *Advances in Neural Information Processing Systems*, 37: 28706–28719, 2024. 2

[3] Bencheng Liao, Shaoyu Chen, Haoran Yin, Bo Jiang, Cheng Wang, Sixu Yan, Xinbang Zhang, Xiangyu Li, Ying Zhang, Qian Zhang, et al. Diffusiondrive: Truncated diffusion model for end-to-end autonomous driving. *arXiv preprint arXiv:2411.15139*, 2024. 1