# PersonaDrive: Controllable Trajectory Prediction with Multi-Dimensional Driving Personas

Chan Lee[1], Kimin Yun[2], Yuseok Bae[2], Seong Tae Kim[1†], and Jung Uk Kim[1†]

[1] Kyung Hee University, Yong-in, South Korea
{cksdlakstp12, st.kim, ju.kim}@khu.ac.kr
[2] ETRI, Daejeon, South Korea
{kimin.yun, baeys}@etri.re.kr

**Abstract.** Although recent trajectory prediction and end-to-end autonomous driving methods improve robustness in urban environments, they still lack meaningful controllability. Existing benchmarks either provide no persona-conditioned annotations or support only a single urgency spectrum (*i.e.,* emergency, normal, relaxed), which cannot distinguish personas that share the same urgency level but require different driving dynamics. To address this, we propose (*i*) the Persona-Conditioned Trajectory (PCT) dataset, which decomposes driving personas along two axes—Temporal Urgency and Ride Comfort—and combines three levels of each to form a grid of nine personas, each paired with natural-language descriptions and trajectories, and (*ii*) PersonaDrive, a framework that can learn driving personas from language and can generate persona-specific trajectories. PersonaDrive incorporates Persona-Conditioned Anchor Transform (PCAT), which hierarchically reshapes anchors along both axes, and Persona-Conditioned Multi-Modal Fusion (PCMF) for BEV-level persona fusion. Training is supervised by a Hierarchical Guide Loss enforcing axis-aligned physical orderings and an Axis-Decomposed Diversity Loss preventing diagonal mode collapse. Experimental results show that PersonaDrive consistently improves over the compared baselines across multi-dimensional scenarios. The code and PCT dataset are available at https://github.com/VisualAIKHU/PersonaDrive.

**Keywords:** Autonomous Driving · Dataset and Benchmark · Multi-modal Learning

## 1 Introduction

As techniques that learn driving policies directly from raw sensor inputs–such as object detection [6, 11, 17, 19, 24, 28, 39, 44, 47], object tracking [54–56], and online mapping [29, 31, 33]–continue to advance, the importance of autonomous driving has become increasingly prominent. Among various autonomous driving

† Corresponding authors

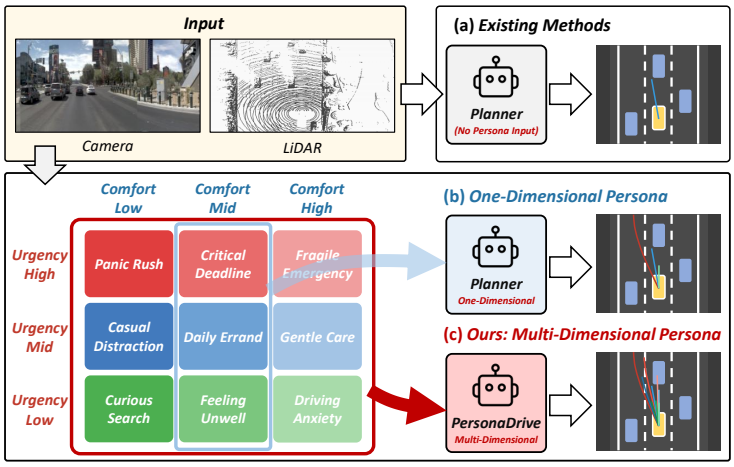


**Fig. 1:** Conceptual comparison of (a) existing trajectory predictors, which output a single persona-agnostic plan, (b) one-dimensional persona methods conditioned on a single urgency axis, and (c) our PersonaDrive, which conditions on natural-language persona descriptions spanning Temporal Urgency and Ride Comfort, yielding nine distinct driving personas.

tasks, trajectory prediction has gained significant attention as a core component of future planning [18, 38, 43]. This task leverages spatiotemporal information from the current scene, such as camera and LiDAR observations, to forecast the future positions of the ego vehicle over upcoming time steps.

This importance of trajectory prediction is further highlighted by recent autonomous driving studies that have advanced the field through unified BEV representations [16], multi-sensor fusion [7], anchor-based distillation [27, 49], and diffusion- or distribution-based prediction [5, 30].

Despite these recent advancements, as shown in Figure 1(a), many current trajectory prediction models still fall short in reflecting diverse human driving personas. As they are trained on datasets collected under fixed, normal driving conditions, these models often learn a single conservative trajectory pattern instead of adapting to diverse user personas. In reality, driving behavior varies widely depending on the driving situation, shaped by multiple concurrent factors including trip urgency, passenger comfort requirements, cargo fragility, road familiarity, and emotional state that interact to produce a unique driving style in each situation [12, 13, 20–22, 32, 42]. Without controllable persona inputs, existing models cannot capture this behavioral diversity.

As illustrated in Figure 1(b), prior work on persona-conditioned planning has attempted to address this gap by introducing a single spectrum of coarse categories such as emergency, normal, and relaxed [4, 10, 13, 25, 36, 45, 52, 53]. However, such a one-dimensional formulation maps all personas within the same urgency level to identical outputs: for instance, a paramedic rushing a critical patient to a hospital and a firefighter racing to a blaze are both urgent, yet the former demands smooth, jerk-free motion to protect the patient while the latter

tolerates aggressive maneuvering. Mapping both to a single "emergency" label erases this distinction and limits the advantage of using natural language over a simple one-hot encoding.

To capture these orthogonal aspects of driving behavior, as shown in Figure 1(c), we decompose persona into two independent and representative axes-Temporal Urgency and Ride Comfort-and discretize each into three levels to form nine distinct persona categories that span the full behavioral space.

In this paper, we introduce the Persona-Conditioned Trajectory (PCT) Dataset, which encodes driving persona along two behavioral axes—Temporal Urgency and Ride Comfort—yielding nine distinct persona categories (Figure 1(c)). Each category is paired with natural-language persona descriptions and corresponding ground-truth trajectories for the same scenes. We further propose PersonaDrive, a framework that learns persona-driven behavior from language and generates persona-specific future trajectories. By crossing three urgency and three comfort levels, the nine categories span the full spectrum of driving personas, from someone rushing to an urgent appointment regardless of discomfort, to someone delivering fragile instruments who needs the smoothest possible ride.

Notably, scenarios such as the urgent-but-gentle need of a persona transporting delicate equipment and the urgent-but-rough tolerance of a persona racing to a critical meeting share the same urgency level yet require different trajectories, making text-based conditioning essential beyond simple categorical labels [37]. We address two key challenges: (*i*) how to control predictions by integrating the text-based persona description with the current scene across nine diverse categories, and (*ii*) how to generate diverse, persona-conditioned trajectories while maintaining clear separation along both the urgency and comfort axes.

To address the challenge (*i*), we introduce Persona-Conditioned Multi-Modal Fusion (PCMF), which aligns persona cues with visual and spatial representations through query-conditioned compression and gated fusion, enabling persona-aware trajectory prediction. We also propose Persona-Conditioned Anchor Transform (PCAT), which modulates the base anchors that serve as initial trajectory prototypes according to the target persona. Together, these components enable trajectories to respond to persona changes across both axes in a stable, predictable way, improving controllability.

For challenge (*ii*), we devise a diversity loss whose core idea is to preserve the persona separation observed in ground-truth behaviors along both axes, penalizing predicted trajectories that converge despite different personas. As a result, our framework enables direct control of diverse, persona-conditioned trajectory predictions via natural-language descriptions, differentiating multi-dimensional persona scenarios that one-dimensional labels cannot.

The major contributions can be summarized as follows:

- We introduce a two-axis persona formulation that decomposes driving behavior along Temporal Urgency and Ride Comfort, yielding nine distinct persona categories, and construct the PCT dataset that provides natural-language persona descriptions paired with corresponding ground-truth trajectories for all nine categories.

- We propose PersonaDrive, a framework that learns driving persona from text and injects it into planning through PCAT and PCMF.
- Experiments on the NAVSIM closed-loop benchmark show that PersonaDrive improves over the compared baselines across all nine persona categories for persona-aware trajectory prediction.

## 2 Persona-Conditioned Trajectory Dataset

### 2.1 Multi-Dimensional Behavioral Decomposition

People vary driving behavior according to trip purpose and current state [20,22,42], yet existing trajectory prediction datasets such as OpenScene do not provide controllable inputs [2, 3, 9, 40, 48, 51], causing models to converge to average behavior rather than learning persona-conditioned distributions. To move beyond a single urgency axis (*e.g.*, emergency / normal / relaxed), we define two orthogonal behavioral axes that jointly form a multi-dimensional persona space:

**(*i*) Temporal Urgency (Urgency)**: the degree of time pressure perceived by the driver, which primarily governs speed and forward progress. High urgency yields faster travel and decisive maneuvers; low urgency allows leisurely pacing.

**(*ii*) Ride Comfort (Comfort)**: the priority placed on ride smoothness and passenger well-being, which primarily governs acceleration profiles and lateral dynamics. High comfort demands gentle acceleration, smooth cornering, and minimal jerk; low comfort tolerates aggressive dynamics.

These two axes are orthogonal: urgency dictates how fast the vehicle should progress, while comfort dictates how smoothly it should do so. By discretizing each axis into three levels (High / Mid / Low), we obtain a 3×3 grid of nine personas, as illustrated in Figure 1. This grid structure enables rigorous validation of axis independence: by fixing one axis and varying the other, we can verify that each dimension contributes distinct and measurable behavioral changes to the predicted trajectory (Section 2.1).

### 2.2 Dataset Construction

The PCT dataset represents each persona using natural-language descriptions that take the form of a passenger request to an autonomous taxi, naturally encoding both urgency and comfort through factors such as age, trip purpose, stress level, and emotional state. For example, a panicked passenger desperately rushing to save an overdosed friend, unconcerned about ride quality, maps to (Urgency High, Comfort Low), while a parent urgently transporting a severely burned child who cannot tolerate any bumps maps to (Urgency High, Comfort High). By varying these factors, the text spans the full $3 \times 3$ persona grid without requiring explicit axis labels.

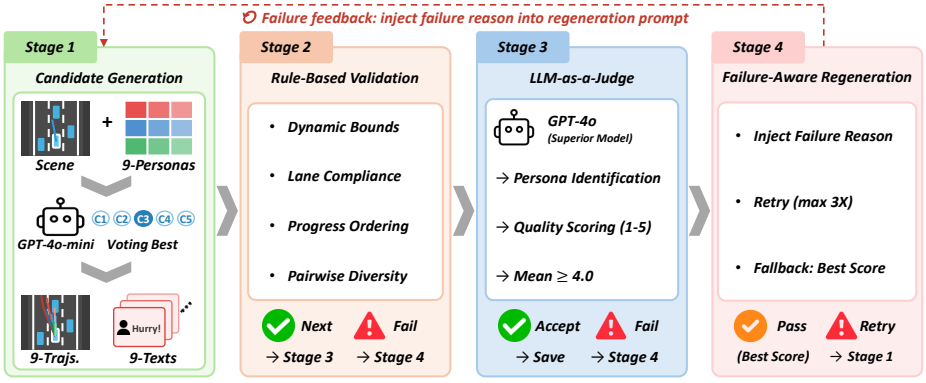


**Fig. 2:** Four-stage generation and validation pipeline for the PCT dataset. Stage 1 generates five candidate sets of nine trajectory parameters via GPT-4o-mini and selects the best through voting. Stages 2 and 3 run in parallel: rule-based validation checks dynamic bounds, lane compliance, progress ordering, and pairwise diversity, while GPT-4o evaluates trajectory quality and verifies persona consistency. Samples failing either stage are regenerated in Stage 4, where failure context is injected into the prompt.

### 2.3 Generation and Validation Pipeline

We construct the dataset through a four-stage pipeline illustrated in Figure 2. For each scene, structured scene information (lane geometry, drivable area, surrounding objects, traffic signals, ego state) is extracted into a JSON representation, and a system prompt specifies the nine persona definitions and trajectory generation parameters. The actual waypoint coordinates are computed from these parameters by applying lane-keeping constraints and drivable-area clamping.

**Stage 1: Candidate Generation and Voting.** For each scene, we generate five candidate sets of nine trajectory parameter sets using GPT-4o-mini and select the best through a voting mechanism that considers plausibility and diversity. The corresponding passenger request texts are then generated via a separate API call, grounded in the actual trajectory characteristics.

**Stage 2: Rule-Based Validation.** The selected candidate undergoes deterministic checks: dynamic bounds on speed, acceleration, and jerk; adherence to the $3 \times 3$ progress pattern; TTC $\geq 1.0\,\mathrm{s}$; and minimum pairwise diversity (ADE $\geq 0.3\,\mathrm{m}$). For texts, we verify sufficient diversity via Jaccard similarity.

**Stage 3: LLM-as-a-Judge.** In parallel with Stage 2, GPT-4o [1]—a more capable model than the GPT-4o-mini generator—evaluates each candidate following the LLM-as-a-Judge paradigm [57]. A trajectory judge scores realism, safety, and persona consistency, while a text judge classifies each description into its correct persona category. A sample passes only when the combined score exceeds a threshold.

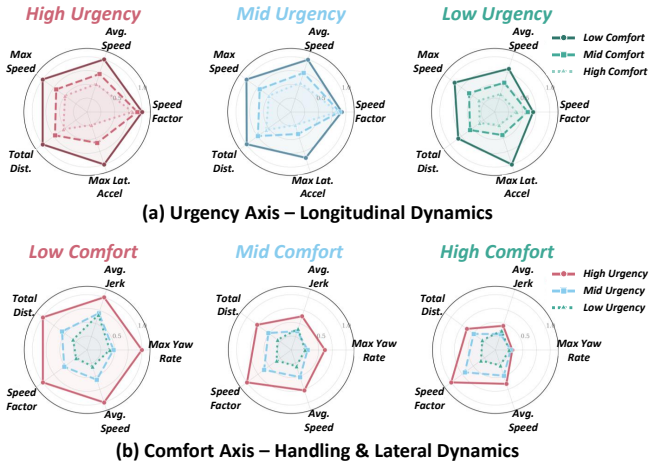


**Fig. 3:** Multi-dimensional behavioral profiles of the nine personas. (a) With urgency fixed, longitudinal metrics (speed, distance) separate by urgency level while comfort lines overlap. (b) With comfort fixed, handling metrics (jerk, yaw rate) separate by comfort level while urgency lines overlap. This confirms that the two axes control orthogonal aspects of driving behavior.

**Stage 4: Failure-Aware Regeneration.** Scenes that fail validation are regenerated with failure reasons injected into the prompt. Rule-based failures allow up to three rounds and quality-based failures up to two. If a scene still does not pass, the result with the highest combined score is selected. We applied this pipeline to the entire OpenScene dataset [40], generating nine persona-conditioned trajectory-text pairs for every scene in both the train and test splits. To verify that the resulting persona labels are not artifacts of the GPT-family judge, we re-judge a held-out sample with two independent model families (Claude, Gemini), which show substantial-to-near-perfect agreement ($\kappa = 0.978/0.804$); details are in the supplementary material.

### 2.4 Dataset Analysis

Figure 3(a) and Figure 3(b) visualize the trajectory statistics of the generated dataset using radar charts, organized along each axis to verify that the two dimensions produce the expected behavioral signatures.

**Urgency axis (Figure 3(a))** Each chart fixes one urgency level and overlays the three comfort levels. The longitudinal metrics (Avg. Speed, Max Speed, Speed Factor, Total Dist.) decrease consistently from High Urgency to Low Urgency, confirming that urgency primarily governs forward progress. In contrast, Max Lat. Accel shows clear separation across comfort levels: Low Comfort yields the largest lateral acceleration, while High Comfort compresses it, reflecting smoother

**Table 1:** Persona distinguishability of the PCT dataset. Each cell reports the accuracy (%) for identifying the correct persona from shuffled trajectory–text pairs.

| | **Comf. Low** | **Comf. Mid** | **Comf. High** |
|---|---|---|---|
| **Urg. High** | 89.9 | 89.9 | 92.2 |
| **Urg. Mid** | 98.4 | 82.2 | 82.9 |
| **Urg. Low** | 94.6 | 73.6 | 67.4 |

cornering under higher comfort priority.

**Comfort axis (Figure 3(b)).** Each chart fixes one comfort level and overlays the three urgency levels. The handling metrics (Avg. Jerk, Max Yaw Rate, Max Lat. Accel) decrease consistently from Low Comfort to High Comfort, confirming that comfort governs ride smoothness. In contrast, Avg Speed, Speed Factor, and Total Dist. show clear separation across urgency levels: High Urgency consistently occupies the outermost region while Low Urgency stays near the center.

These patterns jointly confirm axis independence: urgency controls longitudinal dynamics (speed, distance) while comfort controls lateral dynamics (jerk, yaw rate, lateral acceleration). Within each chart, the lines for the non-varying axis largely overlap, indicating that each axis varies independently without interfering with the other. A Wilcoxon signed-rank test [50] on the per-scene mean speed differences yields $p < 0.001$ for all urgency-adjacent pairs (within each comfort level), and a corresponding test on mean absolute jerk yields $p < 0.001$ for all comfort-adjacent pairs (within each urgency level), confirming statistical significance along both axes.

### 2.5 Human Evaluation

To assess the human plausibility of the generated annotations, we conduct a user study with 13 participants. Each participant was presented with shuffled trajectory–text pairs from the same scene and asked to identify which of the nine personas each pair belongs to. As shown in Table 1, the overall accuracy is 85.7%, highest for scenarios at the extremes of one or both axes and relatively lower for cells near the center of the grid, which are closer to default driving behavior. Most importantly, accuracy remains well above chance even for cells sharing one axis value, demonstrating that humans can perceive the comfort dimension even when urgency is held constant and confirming that both axes carry distinguishable information. A separate quality and realism study is also conducted, where participants consistently rate both trajectory realism and text clarity favorably; detailed results are provided in the supplementary material.

## 3 Proposed Method: PersonaDrive

Figure 4 shows the overall architecture. The inputs comprise cameras $I_{cam}$, LiDAR $I_{lidar}$, and a text description $T$ encoding the user driving persona. A

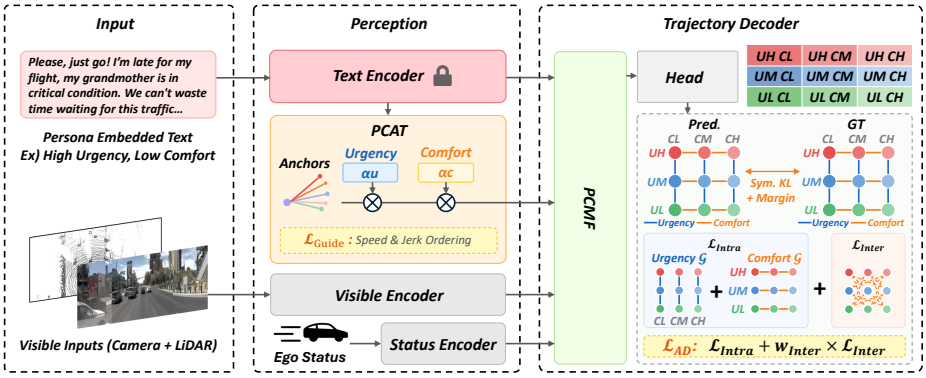


**Fig. 4:** Overall architecture of the proposed PersonaDrive framework. PCAT modulates the anchor bank along the urgency and comfort axes using scalars derived from the text query. PCMF fuses BEV, ego, persona, and transformed anchor queries to produce persona-aware trajectories. Training is supervised by a Hierarchical Guide Loss and an Axis-Decomposed Diversity Loss. $\otimes$ denotes element-wise multiplication.

visual encoder produces BEV queries $Q_{bev}$, and a frozen text encoder (*e.g.,* MiniLM [46]) produces $Q_{txt}$. PCAT receives $Q_{txt}$ and hierarchically modulates the trajectory anchor set $\mathcal{P}$ to generate the persona-adapted anchor set $\widetilde{\mathcal{P}}$. Together with the ego query $Q_{ego}$, $Q_{bev}$, $Q_{txt}$, and $\widetilde{\mathcal{P}}$ are fed to the PCMF module, and the head network estimates the persona-conditioned trajectory.

### 3.1 Persona-Conditioned Anchor Transform

To estimate a trajectory, a trajectory anchor set $\mathcal{P} = \{p_1, \ldots, p_K\}$ serves as initial priors for prediction. Since the target persona can vary widely, we propose PCAT, which applies a hierarchical two-stage anchor modulation.

We decompose the text query $Q_{txt}$ into two independent branches. The urgency branch produces a scalar $\alpha_u$ that controls the global trajectory scale (*i.e.,* overall travel distance), and the comfort branch produces a per-timestep modulation vector $\alpha_c \in \mathbb{R}^T$ that adjusts the local trajectory shape at each waypoint. Each branch consists of a two-layer MLP with GELU activation, followed by a task-specific projection:

$$\alpha_u = f_u(Q_{txt}) \in \mathbb{R}, \quad \alpha_c = \sigma\big(f_c(Q_{txt})\big) \times \Delta_c + \gamma_c \in \mathbb{R}^T, \tag{1}$$

where $f_u$ and $f_c$ denote the urgency and comfort projection networks, and $\sigma$ is the sigmoid function. The comfort modulation $\alpha_c$ is bounded to $[\gamma_c,\ \gamma_c + \Delta_c]$ via the sigmoid-affine transformation. We set $\gamma_c = 0.8$ and $\Delta_c = 0.4$, constraining $\alpha_c$ to $[0.8, 1.2]$ for stable yet meaningful shape variation, while $\alpha_u$ is unconstrained for flexible global scaling.

The persona-adapted anchor set $\widetilde{\mathcal{P}} = \{\tilde{p}_1, \ldots, \tilde{p}_K\}$ is then obtained by applying the two modulations sequentially. We first scale the entire anchor by $\alpha_u$ and

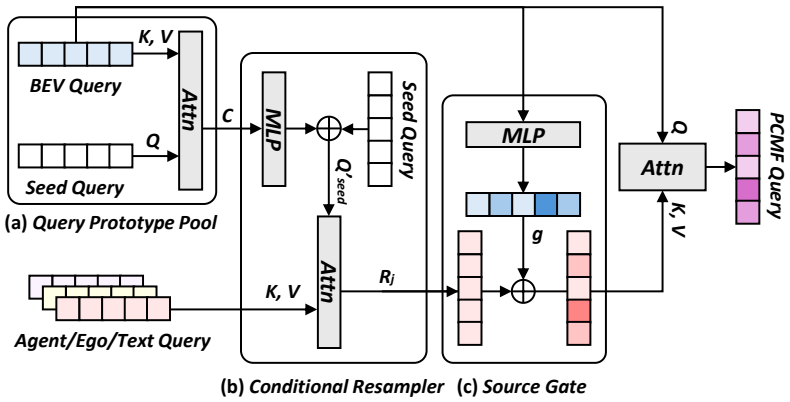


**Fig. 5:** Illustration of the PCMF module. (a) Query Prototype Pool compresses $Q_{bev}$ into a conditional vector $C$. (b) Conditional Resampler modulates learnable slots with **C** and resamples each source. (c) Source Gate applies persona-dependent gating, and the PCMF fuser combines the resulting slots with BEV features.

then modulate each timestep independently by $\alpha_c$:

$$\tilde{p}_i^{(u)} = p_i \times \alpha_u, \quad \tilde{p}_i(t) = \tilde{p}_i^{(u)}(t) \times \alpha_c(t). \tag{2}$$

where $t$ indexes the timestep. For the neutral persona (medium urgency, medium comfort), we skip the transformation and use the base anchor $p_i$ as-is, which guarantees that the model behavior is identical to the baseline when no persona modification is applied.

### 3.2 Persona-Conditioned Multi-Modal Fusion

After modulating the anchors, PCMF updates the input queries to reflect the user driving persona. We treat $Q_{bev}$ as the main query and combine it with $\widetilde{\mathcal{P}}$ to generate the agent query $Q_{\text{agent}}$. As shown in Figure 5, the module proceeds in three steps. First, the Query Prototype Pool compresses $Q_{bev}$ into a conditional vector via a set of $M$=8 learnable seed queries:

$$C = \text{Attn}(Q_{\text{seed}}, Q_{bev}, Q_{bev}). \tag{3}$$

Next, the **Conditional Resampler** modulates learnable slots with a projected context vector $\tilde{C}$ and resamples each source into persona-aligned representations:

$$Q'_{slot} = Q_{seed} + \tilde{C}, \quad R_j = \text{Attn}(Q'_{slot}, Q_j, Q_j),\ j \in \{\text{agent}, \text{ego}, \text{text}\}, \tag{4}$$

where $Q'_{slot}$ is obtained by adding $\tilde{C}$ to learnable slot seeds $Q_{seed}$, enabling the conditioned slots to attend selectively to tokens consistent with the target persona. A **Source Gate** then modulates source-wise contributions with a scene-conditioned gating vector $g = \text{softmax}(\text{MLP}(Q_{bev}))$, yielding $\tilde{R}_j = g_j R_j$, where

$g = [g_{\text{agent}}, g_{\text{ego}}, g_{\text{text}}]$ dynamically adjusts each source's contribution based on the scene complexity and target persona. Finally, the PCMF Fuser integrates the gated slots:

$$Q_{PCMF} = \text{Attn}\big(Q_{bev},\ [\tilde{R}_{\text{agent}}; \tilde{R}_{\text{ego}}; \tilde{R}_{\text{text}}],\ [\tilde{R}_{\text{agent}}; \tilde{R}_{\text{ego}}; \tilde{R}_{\text{text}}]\big), \tag{5}$$

producing the updated BEV query $Q_{\text{PCMF}}$ that enables persona-aware trajectory prediction.

### 3.3 Persona-Guided Training Objectives

Let $\tau_{m,t} \in \mathbb{R}^2$ denote the predicted waypoint of persona $m$ at timestep $t$, with $T$ the total number of timesteps. The following two losses leverage the $3 \times 3$ grid structure to provide axis-aligned supervision.

**Hierarchical Guide Loss.** Although PCAT can produce anchors appropriate for each persona, naive end-to-end training may fail to enforce the expected physical orderings across the two behavioral axes. We therefore introduce a Hierarchical Guide Loss that enforces axis-aligned orderings using ground-truth margins. For each persona $m$, we define the trajectory length $l_m = \sum_{t=1}^{T-1} \|\tau_{m,t+1} - \tau_{m,t}\|_2$ and the mean jerk magnitude $j_m$ via third-order finite differences ($\Delta t = 0.5\,\text{s}$). Within the same comfort column, higher-urgency trajectories should travel farther; within the same urgency row, lower-comfort trajectories should exhibit higher jerk. Both orderings are enforced with the same margin form:

$$\mathcal{L}_{\text{urg}} = \frac{1}{|\mathcal{C}_u|} \sum_{(m_h, m_l) \in \mathcal{C}_u} \text{ReLU}\Big( \big[ l^{\text{gt}}_{m_h} - l^{\text{gt}}_{m_l} \big]_+ - \big( l^{\text{pred}}_{m_h} - l^{\text{pred}}_{m_l} \big) \Big), \tag{6}$$

$$\mathcal{L}_{\text{cmf}} = \frac{1}{|\mathcal{C}_c|} \sum_{(m_l, m_h) \in \mathcal{C}_c} \text{ReLU}\Big( \big[ j^{\text{gt}}_{m_l} - j^{\text{gt}}_{m_h} \big]_+ - \big( j^{\text{pred}}_{m_l} - j^{\text{pred}}_{m_h} \big) \Big), \tag{7}$$

where $\mathcal{C}_u$ and $\mathcal{C}_c$ denote the six adjacent urgency pairs (3 columns $\times$ 2) and six adjacent comfort pairs (3 rows $\times$ 2), respectively. The total guide loss $\mathcal{L}_{\text{guide}} = \mathcal{L}_{\text{urg}} + \mathcal{L}_{\text{cmf}}$ yields 12 margin constraints that prevent the model from conflating urgency-driven and comfort-driven trajectory differences.

**Axis-Decomposed Diversity Loss.** Applying a single diversity objective over all $M$=9 persona pairs risks diagonal mode collapse, where personas differing along both axes still produce similar trajectories. We therefore decompose the loss into intra-axis and inter-axis components.

For the intra-axis loss, we define six groups of three personas sharing one axis (three urgency groups and three comfort groups). Within each group $\mathcal{G}_k$, we compute pairwise average $L_1$ distance matrices from predicted and GT trajectories, convert each row to a temperature-scaled softmax distribution (with diagonal masking), and align the two via symmetric KL divergence:

$$\mathcal{L}^{(k)}_{\text{KL}} = \frac{1}{2|\mathcal{G}_k|} \sum_{i \in \mathcal{G}_k} \Big[ \text{KL}(q_i \| p_i) + \text{KL}(p_i \| q_i) \Big], \tag{8}$$

where $p_i$ and $q_i$ are the softmax distributions derived from predicted and GT distances. A margin term additionally penalizes cases where predicted pairwise distance falls below the GT distance. The intra-axis loss $\mathcal{L}_{\text{Intra}}$ averages both terms over all six groups; the inter-axis loss $\mathcal{L}_{\text{Inter}}$ applies the same KL-margin formulation over all $M$=9 personas. The final loss combines both:

$$\mathcal{L}_{\text{AD}} = \mathcal{L}_{\text{Intra}} + w_{\text{Inter}} \cdot \mathcal{L}_{\text{Inter}}, \tag{9}$$

where $w_{\text{Inter}} = 0.2$. The intra-axis term forces diversity among personas differing along one axis, while the inter-axis term regularizes diagonal pairs.

### 3.4 Total Loss

The total loss function is represented as:

$$\mathcal{L}_{\text{Total}} = \lambda_1 \mathcal{L}_{\text{traj}} + \lambda_2 \mathcal{L}_{\text{guide}} + \lambda_3 \mathcal{L}_{\text{AD}}, \tag{10}$$

where $\mathcal{L}_{\text{traj}}$ is the baseline loss of DiffusionDrive [30] for trajectory reconstruction and classification, $\mathcal{L}_{\text{guide}}$ is the Hierarchical Guide Loss that enforces axis-aligned physical orderings, and $\mathcal{L}_{\text{AD}}$ is the Axis-Decomposed Diversity Loss that maintains persona separation. $\lambda_1$, $\lambda_2$, and $\lambda_3$ are balancing hyperparameters.

## 4 Experiments

### 4.1 Evaluation Metrics

We evaluate on NAVSIM [8], a closed-loop planning benchmark built on OpenScene [40], strictly following its protocol with all simulator settings unchanged. Since the NAVSIM nonreactive simulator is calibrated for normal driving conditions, PDMS may penalize intentionally aggressive or cautious trajectories. We therefore adopt Average Displacement Error (ADE) and Final Displacement Error (FDE) as primary metrics, which directly measure agreement with the persona-specific ground truth. We report aggregate PDMS only for comparability with prior work, not as a measure of persona fidelity, as a score calibrated for normal driving does not directly reflect the quality of intentionally non-normal trajectories.

### 4.2 Implementation Details

As our baseline, we adopt DiffusionDrive [30] with a ResNet-34 backbone [14]. The input comprises three cropped, downscaled forward-view camera images stitched horizontally into a 1024×256 frame, together with a BEV raster derived from LiDAR. We train from scratch on the navtrain split for 100 epochs using AdamW [35] (learning rate 0.0006), a global batch size of 128 distributed over 6 NVIDIA RTX A6000 GPUs, and no test-time augmentation. For evaluation on the navtest split, the model outputs an 8-waypoint trajectory over a 4-s horizon.

**Table 2:** Comparison of closed-loop metrics on the planning-oriented NAVSIM navtest split. We report Avg. ADE, Avg. FDE, and Avg. PDMS averaged across all nine persona categories. **Bold**/underlined fonts indicate the best/second-best results.

| Methods | Avg. ADE↓ | Avg. FDE↓ | Avg. PDMS↑ |
|---|---|---|---|
| LTF (TPAMI'22) [7] | 2.49 | 3.87 | 52.0 |
| Transfuser (TPAMI'22) [7] | 2.53 | 3.96 | 51.9 |
| UniAD (CVPR'23) [16] | <u>2.47</u> | <u>3.83</u> | 52.5 |
| Hydra-MDP (arXiv'24) [27] | 6.70 | 11.85 | 41.4 |
| PARA-Drive (CVPR'24) [49] | 4.84 | 9.00 | 51.9 |
| Traj-LLM (T-IV'24) [23] | 2.66 | 4.20 | 54.1 |
| VisionTRAP (ECCV'24) [37] | 5.17 | 9.95 | 50.7 |
| WoTE (ICCV'25) [26] | 3.35 | 5.71 | 52.4 |
| DiffusionDrive (CVPR'25) [30] | 3.93 | 5.79 | <u>55.3</u> |
| VADv2 (ICLR'26) [5] | 6.18 | 10.76 | 47.9 |
| **PersonaDrive** | **2.39** | **3.71** | **57.7** |

Loss hyperparameters are fixed to $\tau = 0.25$, $\lambda_1 = 10$, $\lambda_2 = \lambda_3 = 1$. The comfort modulation $\alpha_c$ is bounded to $[0.8, 1.2]$. Although $\alpha_u$ is left unconstrained, across all 109,269 evaluation samples its observed range is $[-1.42, 1.26]$ ($|\alpha_u| > 1$ for only 0.13%) and final waypoints are hard-clipped during denormalization, so no physically invalid trajectory occurs. Across multiple training seeds, the coefficient of variation of Avg. ADE/FDE stays below 0.12%, and our model wins ADE/FDE in all nine cells. All experiments are implemented in PyTorch [41].

### 4.3 Comparison with Previous Methods

Table 2 compares ADE and FDE across all nine persona categories. For fair comparison, all baselines share the same frozen text encoder [46]; the text features are concatenated with each method's BEV representation before the prediction head, ensuring identical persona information. Our method achieves the lowest ADE and FDE, reducing ADE by 3.2% and FDE by 3.1% over the second-best method, confirming closer fidelity to the persona-specific ground truth. It also attains the highest aggregate PDMS. Since PDMS aggregates persona-invariant safety terms with persona-conditional quality terms, we decompose it per-persona in the supplementary material, where the safety terms (NC/DAC/DDC) stay within a consistent floor while the quality terms vary as the persona intends.

### 4.4 Comparison with One-/Multi-Dimensional

To quantify the limitation of one-dimensional persona formulations, we construct two proxies from our inference outputs. 1D-Urgency uses the mid-comfort (CM) prediction for all cells in the same urgency row; 1D-Comfort analogously uses the mid-urgency (UM) prediction for each comfort column. These proxies provide a principled upper bound on one-dimensional performance. As shown in Table 3, both proxies incur consistently higher ADE and FDE than the multi-dimensional model. The degradation is largest at against-the-grain cells where urgency and

**Table 3:** One-dimensional vs. multi-dimensional persona conditioning on the NAVSIM navtest split. **Multi-D** is our full model. 1D-Urg. uses the mid-comfort (CM) prediction for all cells in the same urgency row; 1D-Comf. uses the mid-urgency (UM) prediction for all cells in the same comfort column. **Bold** indicates the best result per cell.

| Method | Metric | UH-CL | UH-CM | UH-CH | UM-CL | UM-CM | UM-CH | UL-CL | UL-CM | UL-CH | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1D-Urg. | | 3.30 | **2.96** | 3.07 | 2.76 | **2.35** | 2.35 | 2.73 | **1.93** | 2.56 | 2.67 |
| 1D-Comf. | ADE↓ | 4.79 | 4.85 | 4.66 | 2.43 | **2.35** | **2.34** | 5.70 | 6.26 | 7.11 | 4.50 |
| **Multi-D** | | **2.98** | **2.96** | **2.69** | **2.43** | **2.35** | **2.34** | **2.02** | **1.93** | **1.82** | **2.39** |
| 1D-Urg. | | 5.81 | **5.15** | 5.42 | 4.14 | **3.52** | 3.53 | 4.06 | **2.63** | 4.21 | 4.27 |
| 1D-Comf. | FDE↓ | 11.76 | 11.40 | 10.32 | 3.69 | **3.52** | **3.50** | 9.84 | 10.88 | 12.87 | 8.64 |
| **Multi-D** | | **5.25** | **5.15** | **4.47** | **3.69** | **3.52** | **3.50** | **2.82** | **2.63** | **2.33** | **3.71** |

**Table 4:** Effect of our proposed components on the NAVSIM navtest split. All metrics are averaged across all nine persona categories. All variants include the text encoder. Detailed per-persona results are in the supplementary material.

| PCAT | PCMF | $\mathcal{L}_{AD}$ | Avg. ADE↓ | Avg. FDE↓ | Avg. PDMS↑ |
|---|---|---|---|---|---|
| - | - | - | 3.93 | 5.79 | 55.3 |
| ✓ | - | - | 3.52 | 5.43 | 56.0 |
| ✓ | ✓ | - | 3.23 | 4.85 | 56.9 |
| - | ✓ | ✓ | 2.55 | 3.87 | 57.2 |
| ✓ | - | ✓ | 2.61 | 3.97 | 57.5 |
| ✓ | ✓ | ✓ | **2.39** | **3.71** | **57.7** |

comfort impose opposing demands (*e.g.*, UH–CH: fast yet gentle, UL–CL: slow yet aggressive), confirming that collapsing either axis erases precisely the distinctions that matter most. Neither axis alone is sufficient; the multi-dimensional formulation achieves the lowest error across all nine categories.

### 4.5 Ablation Study

We conduct an ablation study to quantify the effect of the proposed components (Persona-Conditioned Anchor Transform (PCAT), Persona-Conditioned Multi-Modal Fusion (PCMF), axis-decomposed diversity loss $\mathcal{L}_{AD}$). As shown in Table 4, all components that incorporate at least one of the proposed modules outperform the text-encoder-only baseline. When all our components are considered, we achieve the highest performance.

### 4.6 Discussions

**Text Encoding Models Comparison.** As shown in Table 5, MiniLM achieves the best ADE and FDE despite being the smallest model (≈33M *vs.* ≈125–140M). Since all variants share the same framework and differ only in the frozen text encoder, the gap reflects embedding quality rather than model capacity. Trained for sentence-level similarity, MiniLM naturally places same-row or same-column persona descriptions close together—the structure PCAT needs for axis

**Table 5:** Comparison of different text encoders on the NAVSIM navtest split. We report Avg. ADE, Avg. FDE, and Avg. PDMS averaged across all nine persona categories. Detailed per-persona results are in the supplementary material.

| Text Encoders | Avg. ADE↓ | Avg. FDE↓ | Avg. PDMS↑ |
|---|---|---|---|
| DeBERTa [15] | 5.99 | 11.56 | 57.1 |
| RoBERTa [34] | 4.72 | 8.76 | **62.0** |
| **MiniLM** [46] | **2.39** | **3.71** | <u>57.7</u> |

**Table 6:** Comparison of one-hot persona encoding and text-based persona description on the NAVSIM navtest split. We report Avg. ADE, Avg. FDE, and Avg. PDMS for DiffusionDrive with one-hot labels and our framework with one-hot and text-based persona inputs, averaged across all nine persona categories. More results are in the supplementary material.

| Methods | Avg. ADE↓ | Avg. FDE↓ | Avg. PDMS↑ |
|---|---|---|---|
| DiffusionDrive (One-Hot) [30] | 2.67 | 4.20 | 55.6 |
| PersonaDrive (One-Hot) | 2.54 | 4.01 | 56.6 |
| **PersonaDrive (Text)** | **2.39** | **3.71** | **57.7** |

decomposition—whereas RoBERTa and DeBERTa, trained with token-level objectives, lack this property. In fact, both models produce higher ADE and FDE than the text-encoder-only baseline (Table 4), indicating that token-level embeddings inject noise into the axis decomposition rather than useful persona structure. The higher PDMS of RoBERTa does not indicate better persona-aware planning, since PDMS does not measure persona fidelity.

**One-Hot vs. Text.** Table 6 reports performance when persona is given as a one-hot vector. Our framework outperforms the baseline even under one-hot inputs, and text-based conditioning further improves over one-hot, with the largest gains in multi-dimensional scenarios where personas share the same urgency but differ in comfort. This supports the argument that text embeddings naturally place same-row or same-column descriptions closer in the embedding space, preserving the similarity structure that multi-dimensional persona control requires, while one-hot vectors are mutually orthogonal regardless of axis sharing.

**FPS and Model Size.** Our method processes all nine persona texts in a single pass during training, increasing training time from 0.636s/batch to 0.853s/batch (34.1% slower). Inference overhead is minimal: the baseline runs at 0.022s/image (45 FPS) and ours at 0.023s/image (43 FPS), 4.7% increase. In practical use, persona is specified once before driving, so the text encoding can be computed once and reused. Model size grows from 61.4M to 63.2M parameters (2.75%).

### 4.7 Visualization Results

Figure 6 presents qualitative results on NAVSIM scenes. Higher urgency yields longer trajectories with greater forward progress, while higher comfort produces

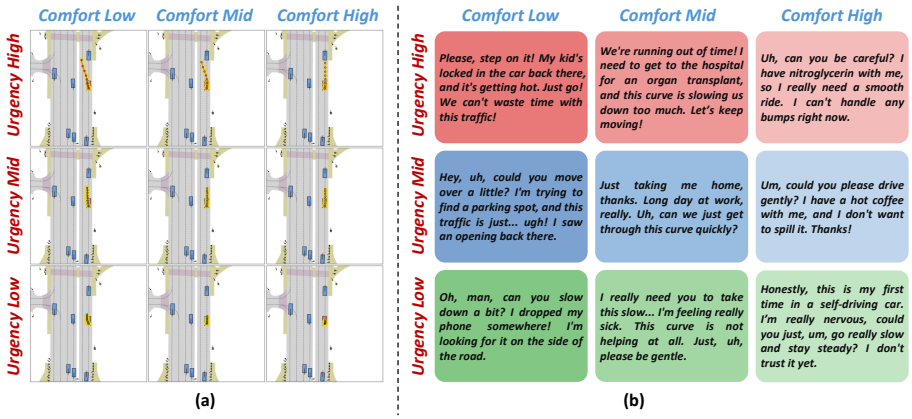


**Fig. 6:** Qualitative results on the NAVSIM dataset. (a) Predicted trajectories under nine personas spanning the Urgency and Comfort axes. (b) Corresponding natural-language persona descriptions used as text input for each cell of the 3×3 grid. Each cell shows the ground-truth persona trajectory (waypoints highlighted in yellow) for one Urgency×Comfort persona.

smoother motion. Personas sharing the same urgency but differing in comfort exhibit similar forward progress yet distinct handling profiles, remaining within the behavioral envelope spanned by the ground-truth data. More examples are in the supplementary material.

## 5 Conclusion

We propose PersonaDrive, a controllable trajectory prediction framework that conditions on natural-language persona descriptions along two orthogonal axes—Temporal Urgency and Ride Comfort—yielding a $3 \times 3$ grid of nine distinct personas. The PCT dataset pairs each persona with trajectory-text annotations via a four-stage LLM-as-a-judge pipeline, and two dedicated modules—PCAT and PCMF—inject persona cues into anchor priors and BEV features, supervised by a Hierarchical Guide Loss and an Axis-Decomposed Diversity Loss. Experiments results on NAVSIM show consistent improvements across all nine personas, with the largest gains where both axes must be jointly resolved.

## Acknowledgements

This work was partly supported by IITP-ITRC grant funded by the Korea government (MSIT)(IITP-2026-RS-2023-00258649, 30%) and partly supported by IITP grant funded by the Korea government (MSIT)(No. RS-2022-II220124, Development of Artificial Intelligence Technology for Self-Improving Competency-Aware Learning Capabilities (50%), IITP-2023-RS-2023-00266615: Convergence Security Core Talent Training Business Support Program (20%)).

## References


1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuscenes: A multimodal dataset for autonomous driving. In: CVPR (2020)
3. Caesar, H., Kabzan, J., Tan, K.S., Fong, W.K., Wolff, E., Lang, A., Fletcher, L., Beijbom, O., Omari, S.: nuplan: A closed-loop ml-based planning benchmark for autonomous vehicles. arXiv preprint arXiv:2106.11810 (2021)
4. Chang, W.J., Zhan, W., Tomizuka, M., Chandraker, M., Pittaluga, F.: Langtraj: Diffusion model and dataset for language-conditioned trajectory simulation. In: ICCV (2025)
5. Chen, S., Jiang, B., Gao, H., Liao, B., Xu, Q., Zhang, Q., Huang, C., Liu, W., Wang, X.: Vadv2: End-to-end vectorized autonomous driving via probabilistic planning. In: ICLR (2026)
6. Chen, S., Wang, X., Cheng, T., Zhang, Q., Huang, C., Liu, W.: Polar parametrization for vision-based surround-view 3d detection. arXiv preprint arXiv:2206.10965 (2022)
7. Chitta, K., Prakash, A., Jaeger, B., Yu, Z., Renz, K., Geiger, A.: Transfuser: Imitation with transformer-based sensor fusion for autonomous driving. TPAMI (2022)
8. Dauner, D., Hallgarten, M., Li, T., Weng, X., Huang, Z., Yang, Z., Li, H., Gilitschenski, I., Ivanovic, B., Pavone, M., et al.: Navsim: Data-driven non-reactive autonomous vehicle simulation and benchmarking. In: NeurIPS (2024)
9. Ettinger, S., Cheng, S., Caine, B., Liu, C., Zhao, H., Pradhan, S., Chai, Y., Sapp, B., Qi, C.R., Zhou, Y., et al.: Large scale interactive motion forecasting for autonomous driving: The waymo open motion dataset. In: ICCV (2021)
10. Felemban, A., Hroub, N., Ding, J., Abdelrahman, E., Shen, X., Mohamed, A., Elhoseiny, M.: imotion-llm: Instruction-conditioned trajectory generation. In: WACV (2026)
11. Gu, J., Sun, C., Zhao, H.: Densetnt: End-to-end trajectory prediction from dense goal sets. In: ICCV (2021)
12. Hao, R., Jing, B., Yu, H., Nie, Z.: Styledrive: Towards driving-style aware benchmarking of end-to-end autonomous driving. arXiv preprint arXiv:2506.23982 (2025)
13. Hasenjäger, M., Wersing, H.: Personalization in advanced driver assistance systems and autonomous vehicles: A review. In: ITSC (2017)
14. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
15. He, P., Liu, X., Gao, J., Chen, W.: Deberta: Decoding-enhanced bert with disentangled attention. arXiv preprint arXiv:2006.03654 (2020)
16. Hu, Y., Yang, J., Chen, L., Li, K., Sima, C., Zhu, X., Chai, S., Du, S., Lin, T., Wang, W., et al.: Planning-oriented autonomous driving. In: CVPR (2023)
17. Huang, J., Huang, G., Zhu, Z., Ye, Y., Du, D.: Bevdet: High-performance multi-camera 3d object detection in bird-eye-view. arXiv preprint arXiv:2112.11790 (2021)
18. Jiang, C., Cornman, A., Park, C., Sapp, B., Zhou, Y., Anguelov, D., et al.: Motiondiffuser: Controllable multi-agent motion prediction using diffusion. In: CVPR (2023)

19. Jung, J., Kim, S., Kim, J.U.: Monosaod: Monocular 3d object detection with sparsely annotated label. In: CVPR (2026)
20. Jurecki, R.S., Stańczyk, T.L.: A methodology for evaluating driving styles in various road conditions. Energies (2021)
21. Kou, G., Jia, F., Mao, W., Liu, Y., Zhao, Y., Zhang, Z., Yoshie, O., Wang, T., Li, Y., Zhang, X.: Padriver: Towards personalized autonomous driving. In: IJCNN (2025)
22. Lajunen, T., Karola, J., Summala, H.: Speed and acceleration as measures of driving style in young male drivers. Perceptual and motor skills (1997)
23. Lan, Z., Liu, L., Fan, B., Lv, Y., Ren, Y., Cui, Z.: Traj-llm: A new exploration for empowering trajectory prediction with pre-trained large language models. T-IV (2024)
24. Lee, C., Shin, S., Park, G.M., Kim, J.U.: Multispectral pedestrian detection with sparsely annotated label. In: AAAI (2025)
25. Li, D., Li, C., Wang, Y., Ren, J., Wen, X., Li, P., Xu, L., Zhan, K., Jia, P., Lang, X., et al.: Learning personalized driving styles via reinforcement learning from human feedback. arXiv preprint arXiv:2503.10434 (2025)
26. Li, Y., Wang, Y., Liu, Y., He, J., Fan, L., Zhang, Z.: End-to-end driving with online trajectory evaluation via bev world model. In: ICCV (2025)
27. Li, Z., Li, K., Wang, S., Lan, S., Yu, Z., Ji, Y., Li, Z., Zhu, Z., Kautz, J., Wu, Z., et al.: Hydra-mdp: End-to-end multimodal planning with multi-target hydra-distillation. arXiv preprint arXiv:2406.06978 (2024)
28. Li, Z., Wang, W., Li, H., Xie, E., Sima, C., Lu, T., Yu, Q., Dai, J.: Bevformer: learning bird's-eye-view representation from lidar-camera via spatiotemporal transformers. TPAMI (2024)
29. Liao, B., Chen, S., Wang, X., Cheng, T., Zhang, Q., Liu, W., Huang, C.: Maptr: Structured modeling and learning for online vectorized hd map construction. arXiv preprint arXiv:2208.14437 (2022)
30. Liao, B., Chen, S., Yin, H., Jiang, B., Wang, C., Yan, S., Zhang, X., Li, X., Zhang, Y., Zhang, Q., et al.: Diffusiondrive: Truncated diffusion model for end-to-end autonomous driving. In: CVPR (2025)
31. Liao, B., Chen, S., Zhang, Y., Jiang, B., Zhang, Q., Liu, W., Huang, C., Wang, X.: Maptrv2: An end-to-end framework for online vectorized hd map construction. IJCV (2025)
32. Liao, H., Li, Z., Shen, H., Zeng, W., Liao, D., Li, G., Xu, C.: Bat: Behavior-aware human-like trajectory prediction for autonomous driving. In: AAAI (2024)
33. Liu, Y., Yuan, T., Wang, Y., Wang, Y., Zhao, H.: Vectormapnet: End-to-end vectorized hd map learning. In: ICML (2023)
34. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 (2019)
35. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
36. Mao, J., Qian, Y., Ye, J., Zhao, H., Wang, Y.: Gpt-driver: Learning to drive with gpt. arXiv preprint arXiv:2310.01415 (2023)
37. Moon, S., Woo, H., Park, H., Jung, H., Mahjourian, R., Chi, H.g., Lim, H., Kim, S., Kim, J.: Visiontrap: Vision-augmented trajectory prediction guided by textual descriptions. In: ECCV (2024)
38. Ngiam, J., Caine, B., Vasudevan, V., Zhang, Z., Chiang, H.T.L., Ling, J., Roelofs, R., Bewley, A., Liu, C., Venugopal, A., et al.: Scene transformer: A unified architecture for predicting multiple agent trajectories. arXiv preprint arXiv:2106.08417 (2021)

39. Oh, Y., Kim, H.I., Kim, S.T., Kim, J.U.: Monowad: Weather-adaptive diffusion model for robust monocular 3d object detection. In: ECCV (2024)
40. OpenScene Contributors: Openscene: The largest up-to-date 3d occupancy prediction benchmark in autonomous driving. `https://github.com/OpenDriveLab/OpenScene` (2023), accessed: 2026-06-29
41. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. In: NeurIPS (2019)
42. Peng, C., Merat, N., Romano, R., Hajiseyedjavadi, F., Paschalidis, E., Wei, C., Radhakrishnan, V., Solernou, A., Forster, D., Boer, E.: Drivers' evaluation of different automated driving styles: Is it both comfortable and natural? Human factors (2024)
43. Phan-Minh, T., Grigore, E.C., Boulton, F.A., Beijbom, O., Wolff, E.M.: Covernet: Multimodal behavior prediction using trajectory sets. In: CVPR (2020)
44. Shi, S., Jiang, L., Dai, D., Schiele, B.: Mtr++: Multi-agent motion prediction with symmetric scene modeling and guided intention querying. TPAMI (2024)
45. Wan, W., Dou, Z., Komura, T., Wang, W., Jayaraman, D., Liu, L.: Tlcontrol: Trajectory and language control for human motion synthesis. In: ECCV (2024)
46. Wang, W., Wei, F., Dong, L., Bao, H., Yang, N., Zhou, M.: Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In: NeurIPS (2020)
47. Wang, Y., Guizilini, V.C., Zhang, T., Wang, Y., Zhao, H., Solomon, J.: Detr3d: 3d object detection from multi-view images via 3d-to-2d queries. In: CoRL (2022)
48. Wei, C., Qin, Z., Li, S., Zhang, Z., Zhao, X., Abdelraouf, A., Gupta, R., Han, K., Barth, M.J., Wu, G.: Pdb: Not all drivers are the same–a personalized dataset for understanding driving behavior. arXiv preprint arXiv:2503.06477 (2025)
49. Weng, X., Ivanovic, B., Wang, Y., Wang, Y., Pavone, M.: Para-drive: Parallelized architecture for real-time autonomous driving. In: CVPR (2024)
50. Wilcoxon, F.: Individual comparisons by ranking methods. Biometrics Bulletin (1945)
51. Wilson, B., Qi, W., Agarwal, T., Lambert, J., Singh, J., Khandelwal, S., Pan, B., Kumar, R., Hartnett, A., Pontes, J.K., et al.: Argoverse 2: Next generation datasets for self-driving perception and forecasting. arXiv preprint arXiv:2301.00493 (2023)
52. Xia, J., Xu, C., Xu, Q., Wang, Y., Chen, S.: Language-driven interactive traffic trajectory generation. NeurIPS (2024)
53. Xu, Z., Zhang, Y., Xie, E., Zhao, Z., Guo, Y., Wong, K.Y.K., Li, Z., Zhao, H.: Drivegpt4: Interpretable end-to-end autonomous driving via large language model. RA-L (2024)
54. Zeng, F., Dong, B., Zhang, Y., Wang, T., Zhang, X., Wei, Y.: Motr: End-to-end multiple-object tracking with transformer. In: ECCV (2022)
55. Zhang, Y., Sun, P., Jiang, Y., Yu, D., Weng, F., Yuan, Z., Luo, P., Liu, W., Wang, X.: Bytetrack: Multi-object tracking by associating every detection box. In: ECCV (2022)
56. Zhang, Y., Wang, C., Wang, X., Zeng, W., Liu, W.: Fairmot: On the fairness of detection and re-identification in multiple object tracking. IJCV (2021)
57. Zheng, L., Chiang, W.L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., et al.: Judging llm-as-a-judge with mt-bench and chatbot arena. NeurIPS (2023)