# TEX-Drive: Temporal Perception Meets Experience-Guided Mixture-of-Experts for End-to-End Autonomous Driving

Yitong Li, Xuchong Zhang*, Fanjie Kong, Weihuang Chen, Haonan Hou, and Hongbin Sun

State Key Laboratory of Human-Machine Hybrid Augmented Intelligence, and Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, Xi'an 710049, China
zhangxc0329@xjtu.edu.cn

**Abstract.** Human driving behavior embodies two core cognitive mechanisms: experiential reasoning and functional specialization. However, existing end-to-end autonomous driving (E2E-AD) frameworks typically model these two processes separately, often resulting in inconsistent decisions under long-horizon or dynamic conditions. To address this issue, we propose TEX-Drive, a unified E2E framework that integrates temporal perception with an experience-driven Mixture-of-Experts, achieving coordinated temporal perception–guided experience routing. Specifically, the temporal perception unit employs explicit key-frame selection to fuse critical information from both historical and current inputs, forming a temporally coherent scene representation. Building upon this, the experience-driven Mixture-of-Experts planning retrieves expert modules from a long-term memory space according to contextual similarity, dynamically routing the most relevant experts to adaptively generate future trajectories. This design enables the decision-making process to maintain temporal consistency while being guided by accumulated experiential knowledge. Extensive experiments on the Bench2Drive benchmark show that TEX-Drive consistently improves driving score, route completion, and behavioral robustness, substantially outperforming existing state-of-the-art methods.

**Keywords:** End-to-End Autonomous Driving · Mixture-of-Experts · Temporal Perception

## 1 Introduction

Human cognition is widely regarded as a task-scheduling system composed of modular processes, working memory, and a central executive [2,3,35]. This structure enables humans to decompose complex goals, selectively activate relevant sub-processes, and reason over extended temporal contexts by recalling past

* Corresponding author.

experiences. In driving, such cognition manifests as drivers integrate past experiences with current observations to decide whether to follow or overtake, then selects the proper operational expert to execute the maneuver. Inspired by this principle, we argue that E2E-AD systems should embody two fundamental cognitive mechanisms: (1) Experiential reasoning, the ability to recall historical experiences, extract temporal regularities, and anticipate future behavior; (2) Functional specialization, the formation of distinct yet complementary modules that handle different sub-tasks through specialized reasoning and representation. Together, these mechanisms allow an E2E-AD system to exhibit modularity, memory-driven reasoning, and dynamic scheduling analogous to cognition.

Existing research explores these two mechanisms separately. For experiential reasoning, early studies use RNN-based modeling [14, 17], which treats all frames equally and fails to emphasize critical temporal transitions or sudden events, lacking adaptability across time scales. Later approaches employ Transformer [25, 38] and Diffusion [32, 51] to capture long-range dependencies and model temporal continuity, as shown in Fig. 1 (upper). However, these methods suffer from high computational cost, unstable temporal consistency, and limited interpretability or control over frame importance. For functional specialization, mainstream methods integrate Mixture-of-Experts (MoE) into E2E-AD models [37, 41], as shown in Fig. 1 (middle). However, conventional MoE frameworks rely on lightweight gating networks with limited discriminative capacity, making it difficult to distinguish driving semantics. Expert assignment behaves as a black box, lacking interpretability and consistency across scenes. Some works attempt to mitigate this by adding supervision to the gating network, yet such reliance on labeled data weakens adaptability in open environments [45, 49]. Moreover, most MoE-based methods assign experts from short-term features, ignoring temporal dependencies across frames, causing unstable attention and action shifts in long-horizon or interactive scenarios. These observations suggest that temporal perception and expert specialization are often developed as separate components, with limited interaction between temporal reasoning and expert routing in current E2E-AD systems.

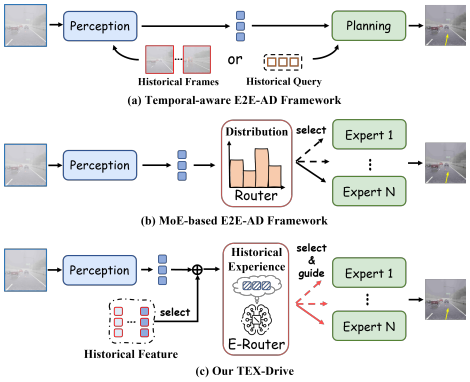


**Fig. 1:** Comparison between our method and previous E2E-AD paradigms, highlighting the differences in how historical information is integrated and functional specialization is achieved. Notably, our approach integrates temporal perception with an experience router (E-Router).

To address these limitations while enhancing structural interpretability, we propose TEX-Drive, an E2E-AD framework where Temporal perception meets EXperience-guided mixture-of-experts. The proposed key-frame guided temporal perception unit computes explicit cross-temporal relevance scores and deterministically selects maneuver-critical frames, transforming implicit temporal aggregation into an observable and analyzable routing signal. Instead of uniformly encoding historical frames, the model explicitly identifies which past observations contribute most to the current driving state, yielding structured temporal semantics for planning. These selected key-frame features are progressively integrated into long-term driving memory, which serves as a structured experience representation for the planner. The experience-driven MoE module then performs expert routing based on this temporally grounded memory, enabling more consistent and interpretable expert assignment across scenes. Unlike conventional gating mechanisms, routing decisions are guided by explicit temporal relevance and memory signals, making expert activation traceable over time. Furthermore, planning feedback dynamically refines temporal attention and memory update, forming a closed-loop interaction between perception and expert reasoning. This mutually reinforced design improves decision stability while preserving interpretability across both temporal reasoning and expert specialization. On Bench2Drive, TEX-Drive achieves a Driving Score of 68.22 and a Success Rate of 40.62%, outperforming recent state-of-the-art baselines and validating the effectiveness of unified temporal–experience modeling.

Our main contributions are summarized as follows:

- **A unified temporal–experience MoE framework.** TEX-Drive jointly models experiential reasoning and functional specialization, bridging temporal perception and expert routing within a single end-to-end architecture.
- **An interpretable key-frame temporal perception mechanism.** The proposed key-frame guided unit produces explicit frame-level routing scores and deterministic key-frame selection, rendering temporal reasoning structurally traceable and dynamically phase-aware across driving stages.
- **An experience-driven expert routing mechanism.** The planner performs expert assignment based on temporally grounded long-term memory signals, enabling stable, context-aware, and semantically consistent specialization in dynamic environments.

## 2 Related Works

### 2.1 End-to-End Autonomous Driving

E2E-AD directly maps sensory inputs to control outputs within a unified differentiable framework [5, 11]. Early imitation learning methods [4, 11] learned human-like behaviors from demonstrations but often suffered from distributional shift and limited generalization in long-tail cases. Reinforcement learning–based frameworks [28, 43] optimized long-term rewards but required large-scale simulation and were difficult to stabilize. Recent works emphasize planning-centric

E2E-AD by unifying perception and trajectory prediction. Transformer-based architectures such as DriveTransformer [25] and ThinkTwice [23] leverage cross-modal attention between camera and LiDAR to enhance spatial reasoning, while diffusion-based approaches [27, 51] generate multi-modal trajectories to capture temporal smoothness. However, most temporal models stack or average frames, lacking mechanisms to explicitly evaluate the importance of each observation. Consequently, they fail to recognize critical moments, leading to delayed reactions and inconsistent long-horizon reasoning. Attempts to address this via recurrent models [9, 14] improve sequential continuity but treat all frames uniformly, whereas attention-based designs [38,39] increase computation and exhibit unstable temporal alignment. This motivates explicit modeling of key moments that shape driving decisions.

### 2.2 General Mixture-of-Experts

The MoE framework [13,40,48] introduces conditional computation by activating sparse subset of experts, improving scalability without linear cost growth. Classical designs such as GShard [29] and Switch Transformer [15] employ lightweight gating networks for routing, while variants enhance load balancing [30, 52] and dynamic specialization [8, 21]. In autonomous driving, MoE-based architectures promote functional specialization by decomposing complex behaviors into modular experts. GEMINUS [45] employs scene- and goal-aware experts for adaptive planning, and MoSE [49] learns distinct driving skills such as merging and overtaking. These studies demonstrate that MoE can improve decision modularity and interpretability in behavior planning [6, 34]. However, most gating networks rely on shallow architectures and short-term features, limiting their ability to discriminate fine-grained semantics and maintain temporal consistency. Expert activation often behaves as a black box, lacking interpretability and stability over long-horizon scenarios. These limitations highlight the need for a temporally consistent and experience-guided expert selection mechanism capable of long-term adaptation and stable decision-making in dynamic environments.

## 3 Methodology

### 3.1 Overview

TEX-Drive is an E2E-AD framework that integrates temporal perception with experience routing, as illustrated in Fig. 2. The overall architecture consists of three core modules: a perception encoder, a key-frame guided temporal perception unit (KTP), and an experience-driven MoE planning unit (E-MoE).

Specifically, the perception encoder extracts multi-scale spatial features from multi-view camera inputs and ego-state information. The temporal perception module performs explicit key-frame selection and temporal fusion to model dynamic scene evolution, generating temporally fused representations for the current frame. The motion planning module further incorporates driving goals and

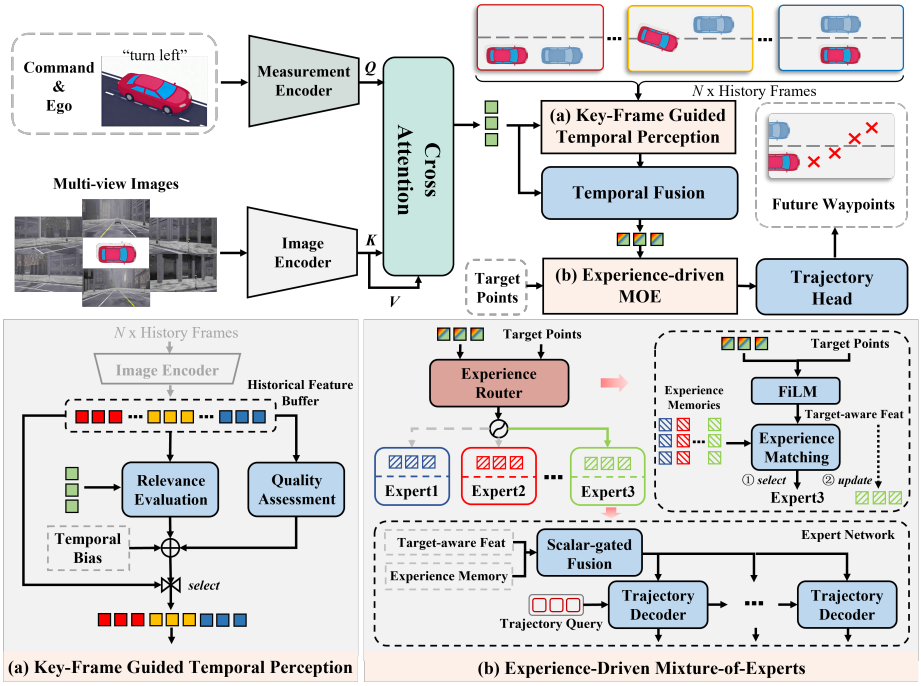


**Fig. 2:** Overview of the TEX-Drive framework. The perception encoder first extracts multi-scale spatial features from multi-view images and ego states. (a) The key-frame guided temporal perception unit maintains a historical feature buffer and adaptively selects key frames to capture dynamic scene evolution. (b) The experience-gated planning unit fuses goal-conditioned temporal features with long-term memory through an experience router and expert network, enabling adaptive expert routing and trajectory generation. These modules form a unified temporal–experience modeling pipeline achieving stable and interpretable E2E motion planning in complex environments.

long-term experience to select and route experts, producing future trajectories. Through these three stages, TEX-Drive achieves a unified modeling process that spans temporal understanding, knowledge scheduling, and action generation.

### 3.2 Key-Frame Guided Temporal Perception

The temporal perception module aims to capture the dynamic evolution of driving scenes. Existing E2E-AD models typically use simple multi-frame concatenation or averaging, treating all frames equally and introducing redundant or noisy temporal information. In contrast, we introduce a key-frame selection mechanism that adaptively identifies and retains the most informative frames for the current driving decision. The overall structure is shown in Fig. 2(a).

**Historical Feature Buffer.** At each time step $t$, the perception encoders receive multi-view camera images and ego measurements (e.g., speed, heading,

steering angle). The image encoder outputs spatial features $F_t^{\text{vis}} \in \mathbb{R}^{d_m}$, while the measurement encoder produces low-dimensional semantic embeddings $F_t^{mea} \in \mathbb{R}^{d_m}$. To model cross-modal interactions between visual observations and driving states, we adopt a cross-attention fusion that produces the current joint representation $F_t$. Meanwhile, the system maintains a historical feature buffer that stores previously encoded features corresponding to the same driving route. Formally, features from the most recent $K_1$ sampled frames are stored in a first-in-first-out (FIFO) queue to form the temporal sequence $\mathcal{H}_t = \{F_{t-K_1+1}, \ldots, F_{t-1}\}$, where $K_1$ denotes the buffer size.

This mechanism ensures that the model can always access multi-scale semantic representations from recent frames along the same route, providing a foundation for subsequent key-frame selection.

**Historical Frame Feature Selection.** Building on the historical feature sequence provided by the buffer module, we further introduce an explicit key-frame selection mechanism to identify frames that contribute most to the current driving decision. Unlike prior implicit weighting strategies, our module jointly models frame quality, decision relevance, and temporal proximity bias, achieving interpretable temporal filtering and dynamic enhancement.

For each historical candidate frame $F_i$, KTP first computes the current-to-history matching logits $L_{t,i} = ((W_q F_t)(W_k F_i)^\top)/\sqrt{d}$ with respect to the current feature $F_t$. Different from self-attention within a single historical feature, the Softmax operation is applied over the historical-frame candidate dimension, i.e., $A_{r,i}^{(m,n)} = \exp(L_{t,i}^{(m,n)}) / \sum_{j=1}^{K_1} \exp(L_{t,j}^{(m,n)})$, where $(m, n)$ denotes a pair of spatial tokens between the current feature and a historical candidate. Thus, $A_{r,i}$ measures the relative cross-temporal relevance between the current feature $F_t$ and the $i$-th historical candidate $F_i$, rather than the intra-frame attention within $F_i$. The scalar routing score is then obtained by averaging over all token pairs, i.e., $s_r^i = \frac{1}{|A_{r,i}|} \sum_{m,n} A_{r,i}^{(m,n)}$. The scalar routing score $s_r^i$ quantifies the semantic contribution of each historical frame to the current driving state, helping identify frames that are highly relevant to maneuvers such as braking or lane changing.

Then, we integrate frame quality assessment, decision relevance evaluation, and temporal bias into a unified correlation score:

$$s_i = s_q^i + s_r^i + \lambda \cdot \text{TemporalBias}(i), \ \text{TemporalBias}(i) = \exp(-(K_1 - i)), \quad (1)$$

where $\lambda$ is a global weighting coefficient that balances the temporal decay effect. Frames that are temporally closer to the present receive higher bias scores, encouraging the model to focus on more recent and contextually relevant cues. Finally, the model selects the top-$K_2$ key frames according to the aggregated relevance scores, i.e., $\mathcal{H} * t' = F_j \mid j \in \text{Top-}K_2(s_i * i = 1^{K_1})$.

After obtaining the selected key-frame set, the model performs temporal fusion to construct a unified representation of the current driving state. Each key frame is assigned a normalized attention weight for temporal fusion, as detailed below:

$$\begin{aligned} w_i &= \exp(-\lambda \cdot (1 - \text{TemporalBias}(i))), \\ \hat{F}_t &= f_{\text{fusion}}([F_t, w_1 F_{t_1}, \ldots, w_{K_2} F_{t_{K_2}}]), \end{aligned} \quad (2)$$

where $f_{\text{fusion}}(\cdot)$ denotes a hierarchical fusion network that aggregates multi-scale temporal cues. This explicit key-frame selection mechanism operationalizes temporal reasoning as structured dependency filtering, which not only preserves temporal continuity but also provides an interpretable criterion for key-frame selection, enhancing feature discriminability and decision consistency.

### 3.3 Experience-Driven MoE Planning

To enable adaptive and knowledge-driven motion planning in complex driving scenes, TEX-Drive introduces an experience-gated planning module, as shown in Fig. 2(b). The module comprises two core components, the experience router and the expert network, which perform dynamic expert selection and goal-conditioned adaptation based on long-term memory. By maintaining both decision adaptivity and experience continuity, this design enhances the stability and interpretability of motion planning.

**Experience Router.** The system maintains a set of $N$ experts $\{E_1, E_2, \ldots, E_N\}$. Each expert $E_i$ owns an experience memory vector $\mathcal{M}_i \in \mathbb{R}^{d_m}$, which stores its long-term scene priors and strategy tendencies. To enable goal-oriented expert selection, we introduce a target-guided modulation mechanism before expert matching. This design is based on the observation that the activation of driving experience depends not only on the perceived scene state but also on the motion intent defined by the target position. Therefore, we employ FiLM-based conditional modulation to embed the goal into the temporally fused feature representation, obtaining the target-aware feature that more precisely aligns the driving context with the most relevant expert experiences. Specifically, the goal feature $G_t$ is processed through a FiLM network [36] to generate channel-wise scaling and shifting parameters $\gamma_i$ and $\beta_i$:

$$[\gamma_i \,\|\, \beta_i] = f_{\text{FiLM}}(G_t), \tilde{Z}_t = \gamma_i \odot \hat{F}_t + \beta_i, \tag{3}$$

where $f_{\text{FiLM}}(\cdot)$ denotes a multi-layer perceptron that projects the goal features into channel-wise modulation parameters, and $\odot$ represents element-wise multiplication. Subsequently, to identify which expert best matches the current driving context, TEX-Drive computes the similarity between the target-modulated temporal feature $\tilde{Z}_t$ and each expert memory $\mathcal{M}_i$ for experience matching:

$$s_t^i = \frac{(\tilde{Z}_t W_q) \cdot (\mathcal{M}_i W_k)^\top}{\sqrt{d_m}}, \tag{4}$$

where $W_q$ and $W_k$ are learnable projection matrices, and $\sqrt{d_m}$ normalizes the similarity by feature dimension. The routing weights $R_t^i = \text{softmax}(s_t^i)$ indicate the confidence of assigning the current state to expert $E_i$'s experience domain.

The router then selects the top-$K_3$ experts with the highest routing weights to form the active expert set, i.e., $\mathcal{E}_t = \{E_i \mid i \in \text{Top}-K_3(R_t, k)\}$.

When an expert $E_i$ is selected, its memory $\mathcal{M}_i$ is not updated via back-propagation but through momentum updating during inference, allowing the expert to gradually adapt while preserving long-term consistency. Although the

router activates the top-$K_3$ experts for inference, only the highest-ranked expert receives the momentum update [18] to prevent homogenization of the experience of different experts. The update rule is:

$$\mathcal{M}_i \leftarrow (1-\mu)\mathcal{M}_i + \mu f_{\text{proj}}(\tilde{Z}_t), \tag{5}$$

where $f_{\text{proj}}(\cdot)$ is a linear projection function and $\mu \in [0,1]$ controls the momentum strength. This update gradually incorporates recent information while preserving temporal stability, mitigating catastrophic forgetting via a working-memory rehearsal mechanism. Updating only the top-ranked expert further prevents representation convergence and expert collapse.

**Expert Network.** After the experience router selects the most relevant experts $\mathcal{E}_t$, TEX-Drive activates these experts to generate motion trajectories. Each expert's long-term memory $\mathcal{M}_i$ not only stores its past behavioral patterns but also encodes adaptive strategies for specific driving contexts. These accumulated experiences are therefore crucial for producing trajectories aligned with the current situation. In this stage, each expert first performs a scalar-gated fusion between the goal-modulated feature $\tilde{Z}_t$ and its memory $\mathcal{M}_i$:

$$\begin{aligned} \alpha &= \sigma\left(W_{\text{g}}\left[\tilde{Z}_t \,\|\, \mathcal{M}_i\right] + b_{\text{g}}\right), \\ H_i &= \text{LN}\left(\tilde{Z}_t + \alpha\left(\mathcal{M}_i - \tilde{Z}_t\right)\right), \end{aligned} \tag{6}$$

where $W_{\text{g}}$ and $b_{\text{g}}$ are learnable parameters of the scalar gating layer. $\sigma(\cdot)$ denotes the sigmoid activation that controls the fusion ratio between $\tilde{Z}_t$ and $\mathcal{M}_i$, and $\text{LN}(\cdot)$ represents layer normalization for stabilizing the fused representation. This scalar gating effectively interpolates between $\mathcal{M}_i$ and $\tilde{Z}_t$ to ensure expert-specific adaptation. Each expert performs autoregressive decoding based on its memory-enhanced representation $H_i$, ensuring that the generated trajectories remain consistent with its experience domain. Instead of direct regression, the decoder predicts a sequence of $T$ discrete time steps $\{y^i_{t+1}, y^i_{t+2}, \ldots, y^i_{t+T}\}$. The decoder is initialized with $H_i$ for the first prediction step, and then proceeds autoregressively as:

$$\begin{aligned} h^i_{t+\tau} &= \text{TrajDecoder}_i\left(H_i, \left\{y^i_{t+k}\right\}_{k=1}^{\tau-1}\right), \\ y^i_{t+\tau} &= W_{\text{out},i} h^i_{t+\tau} + b_{\text{out},i}, \end{aligned} \tag{7}$$

where $W_{\text{out},i}$ and $b_{\text{out},i}$ are the output projection weights and biases of expert $E_i$, mapping the decoder's hidden state $h^i_{t+\tau}$ into the waypoint space. $\text{TrajDecoder}_i$ is a lightweight autoregressive Transformer decoder [44], ensuring temporal consistency and semantic stability. Each expert outputs a candidate trajectory $Y^i_t = \{y^i_{t+1}, \ldots, y^i_{t+T}\}$. After all activated experts complete autoregressive predictions, the system fuses their trajectories through weighted aggregation according to the routing confidence from the experience router. The aggregated results are refined by a trajectory head to produce a feasible path.

# 4 Experiments

## 4.1 Experimental Settings

**Datasets.** While real-world datasets typically focus on open-loop evaluation, which mainly measures prediction accuracy rather than interactive decision quality, we conduct our primary experiments on the challenging multi-ability closed-loop benchmark Bench2Drive [24] to comprehensively assess end-to-end driving performance. Bench2Drive contains 44 driving scenarios automatically collected by the reinforcement learning expert Think2Drive [31]. Compared with real-world datasets such as nuScenes, Bench2Drive repeatedly samples each scenario under diverse locations and weather conditions, resulting in a more balanced perception–behavior distribution. Following standard practice, we use the base subset (1,000 clips), with 950 for training and 50 for validation. Each clip includes six camera views per frame.

**Evaluation Metrics.** For closed-loop evaluation, we follow the official protocol of the Bench2Drive benchmark and adopt two core metrics: Driving Score and Success Rate. Driving Score combines route completion with weighted penalties for collisions, red-light violations, and timeouts, while Success Rate measures the proportion of routes completed without violations. In addition, Bench2Drive provides five ability-level metrics including Merging, Overtaking, Emergency Brake, Give Way, and Traffic Sign Recognition, which assess the model's adaptability across interactive behaviors. For open-loop evaluation, we use the standard Bench2Drive metric L2 Error to measure accuracy of predicted trajectories.

**Implementation Details.** The future prediction steps are set to 4, with a prediction frequency of 2 Hz. The FIFO buffer size $K_1$ is set to 20, and the frame sampling interval is one frame every five frames. The future prediction horizon is set to K = 4 steps with a 2 Hz prediction frequency. For low-level control, we follow the well-tuned PID parameters proposed in TransFuser [10]. Specifically, the longitudinal controller parameters are set to $K_P = 5.0$, $K_I = 0.5$, $K_D = 1.0$, and the lateral controller parameters are set to $K_P = 0.75$, $K_I = 0.75$, $K_D = 0.3$, with a control buffer length of 40. The maximum throttle and braking values are limited to 0.75 and 0.4, respectively, with a braking ratio of 1.1 to ensure smooth acceleration and deceleration. All experiments are conducted on 4 NVIDIA RTX A800 GPUs.

## 4.2 Main Results

To comprehensively evaluate the driving performance of the proposed method, we conduct both closed-loop and open-loop tests on the Bench2Drive benchmark and compare our TEX-Drive with eleven representative methods from 2022 to 2025. The baselines cover a wide range of E2E-AD frameworks, including three multimodal models incorporating LiDAR information and the recently introduced mixture-of-experts model GEMINUS. As summarized in Table 1, TEX-Drive achieves the best overall performance across all metrics even with camera-only input. In closed-loop evaluation, TEX-Drive achieves a Driving Score of

**Table 1:** Comparison of open-loop and closed-loop performance of different models on the Bench2Drive benchmark. "↓" indicate that lower values are better, while "↑" indicate that higher values are better. "C" denotes camera-only, while "C+L" denotes both camera and LiDAR. Compared with existing state-of-the-art methods, our approach achieves the best performance in both open-loop and closed-loop evaluations.

| Method | Modality | Open-loop Metric | Closed-loop Metric | |
|---|---|---|---|---|
| | | Avg. L2 (m)↓ | Driving Score ↑ | Success Rate (%) ↑ |
| TCP [47] | C | 1.70 | 40.70 | 15.00 |
| LeTFuser [1] | C+L | 0.84 | 52.53 | 18.18 |
| UniAD-Base [20] | C | 0.73 | 45.81 | 16.36 |
| ThinkTwice [23] | C+L | 0.95 | 58.79 | 29.54 |
| VAD [26] | C | 0.91 | 42.35 | 15.00 |
| DriveAdapter [22] | C | 1.01 | 64.22 | 33.08 |
| EATNet [7] | C+L | 1.12 | 42.97 | 15.91 |
| DriveTrans [25] | C | 0.62 | 63.46 | 35.01 |
| SparseDrive [42] | C | 0.83 | 42.12 | 15.00 |
| TTOG [33] | C | 0.74 | 45.23 | 16.36 |
| GEMINUS [45] | C | 1.60 | 65.39 | 37.73 |
| TEX-Drive (ours) | C | **0.60** | **68.22** | **40.62** |

68.22 and a Success Rate of 40.62%, surpassing the best-performing GEMINUS by 2.83 and 2.89, respectively. Despite using only monocular visual input, TEX-Drive also exceeds multimodal baselines such as LeTFuser and EATNet, demonstrating the effectiveness and robustness of temporal–experience fusion under single-sensor settings. This consistent improvement primarily stems from the synergy between key-frame temporal perception and experience-guided routing. The former explicitly captures dynamic driving phases, while the latter leverages long-term memory to maintain stable expert activation, leading to smoother maneuvers and more reliable long-horizon reasoning.

In the multi-ability evaluation shown in Table 2, TEX-Drive achieves the highest mean ability score of 45.09%, surpassing DriveAdapter by 3.01. It reaches or matches the best results in Overtaking, Emergency Brake, Give Way, and Traffic Sign, and notably improves the most challenging Merging ability. These results demonstrate that TEX-Drive effectively balances specialization and generalization, maintaining stable and interpretable behavior across diverse driving scenarios. Overall, coupling explicit key-frame temporal modeling with experience-driven expert routing significantly enhances trajectory accuracy, control stability, and decision robustness, validating the effectiveness of the proposed temporal–experience interaction in E2E-AD.

### 4.3 Ablation Study

**Component-wise Ablation.** Table 4 reports ablations of TEX-Drive on the Bench2Drive benchmark, analyzing each component's contribution. The upper

**Table 2:** Comparison of success rates across five driving abilities on the Bench2Drive benchmark, including Merging, Overtaking, Emergency Brake, Give Way, and Traffic Sign. Bold numbers indicate the best performance for each task. Compared with other methods, TEX-Drive achieves the highest mean ability score and leads in most tasks, demonstrating superior generalization and decision stability.

| Method | Multi-Ability (%)↑ | | | | | |
|---|---|---|---|---|---|---|
| | Merging | Overtaking | Em-Brake | Give Way | Traffic Sign | Mean |
| TCP [47] | 16.18 | 20.00 | 20.00 | 10.00 | 6.99 | 14.63 |
| LeTFuser [1] | 0.00 | 20.00 | 18.18 | 0.00 | 47.22 | 17.08 |
| UniAD [20] | 14.10 | 17.78 | 21.67 | 10.00 | 14.21 | 15.55 |
| ThinkTwice [23] | 27.38 | 18.42 | 35.82 | **50.00** | 54.23 | 37.17 |
| VAD [26] | 8.11 | 24.44 | 18.64 | 20.00 | 19.15 | 18.07 |
| DriveAdapter [22] | 28.82 | 26.38 | 48.76 | **50.00** | **56.43** | **42.08** |
| EATNet [7] | 0.00 | 13.33 | 18.18 | 0.00 | 41.66 | 14.63 |
| DriveTrans [25] | 17.57 | 35.00 | 48.36 | 40.00 | 52.10 | 38.60 |
| SparseDrive [42] | 12.50 | 17.50 | 20.00 | 23.03 | 18.60 | 18.60 |
| TTOG [33] | 16.18 | 24.29 | 20.00 | 21.50 | 23.03 | 21.12 |
| GEMINUS [45] | 11.11 | **37.50** | **55.00** | 40.00 | 45.26 | 37.77 |
| TEX-Drive (ours) | **32.22** | **37.50** | 50.00 | **50.00** | 55.77 | **45.09** |

part shows a baseline without temporal modeling or experts, which performs poorly on Success Rate and Multi-Ability Score, highlighting the need for explicit temporal perception and experience-based specialization for stable decisions. Adding KTP or E-MoE yields gains, with KTP increasing Driving Score by 2.6 and Success Rate by 3.4, while E-MoE improves consistency by 3.1/3.8.

However, when both modules are combined, the complete model KTP + E-MoE achieves a remarkable improvement, surpassing the TP + MoE variant by 5.2 in Driving Score and 4.9 in Success Rate. This significant improvement demonstrates that KTP effectively extracts key temporal frames while suppressing redundant information, and E-MoE leverages accumulated experiential knowledge to generate more stable and semantically consistent expert routing. More importantly, when the two modules operate jointly, KTP provides high-quality temporal semantics that deliver clear phase cues for E-MoE's experience routing, while E-MoE, in turn, stabilizes KTP's temporal focus through experience-driven feedback. This mutually reinforcing mechanism yields a synergistic effect greater than the sum of their individual contributions, substantially enhancing temporal consistency and decision stability, and ultimately underpinning TEX-Drive's superior closed-loop driving performance.

**Hyperparameter Sensitivity.** We introduce two key hyperparameters in TEX-Drive, where $K_2$ controls the number of selected historical frames in KTP, and $K_3$ determines the number of simultaneously activated experts in E-MoE. To evaluate their impact, we conduct ablation experiments, and the results are shown in Fig. 3. The red curve (for $K_2$) exhibits a mild downward trend as

more historical frames are used, while the blue curve (for $K_3$) declines more sharply as the number of activated experts increases. This indicates that both excessive temporal input and dense expert activation can degrade performance. The decline of the blue curve occurs because activating too many experts weakens functional specialization and amplifies interference among experts, leading to unstable routing [12, 16]. The slight decrease in the red curve suggests that additional historical frames do not provide more useful temporal cues but instead introduce redundant or noisy information. From the final performance perspective, we select $K_3 = 2$. Although $K_2 = 3$ is not the absolute optimum, the improvement from using more frames is limited while computational cost increases significantly, making this setting a more efficient and balanced choice.

### 4.4 Expert Activation Analysis

To test whether experience guidance produces meaningful specialization rather than expert averaging, we analyze two dynamic scenarios in Fig. 4. The heatmaps show the normalized routing probabilities derived from Eq. (4). We divide each sequence into behavioral phases according to its ground-truth trajectory and compare whether expert activation changes coherently with these phases.

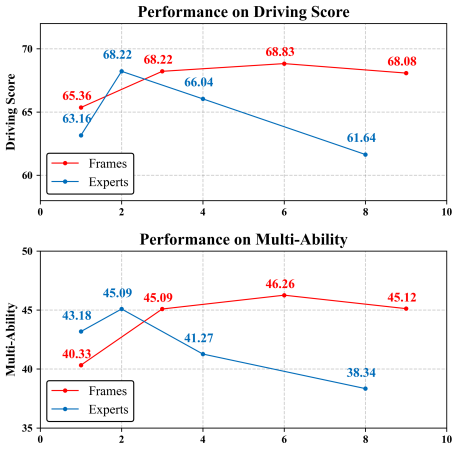


**Fig. 3:** Ablation results of TEX-Drive with respect to the number of key frames ($K_2$) and experts ($K_3$). The red curve indicates varying $K_2$ while fixing $K_3 = 2$, and the blue curve represents varying $K_3$ with $K_2 = 3$. The results highlight the correctness of our chosen hyperparameters.

In the narrow-turn case (first row), the trajectory comprises straight (steps 0–6), turning (7–16), and straight (17–19) phases. Expert 3 dominates the turn but is suppressed in both straight segments, whereas Expert 7 exhibits the opposite pattern. Thus, the repeated straight-driving phase recovers the same expert preference after the turn, providing stronger evidence of phase-dependent specialization than a single activation peak. TEX-Drive completes the maneuver, while TP+MoE exhibits fragmented switching, mistimed steering, and a large turning radius that causes lane departure.

This temporal structure is important for interpreting specialization. A high response from one expert at an isolated step could be caused by transient visual variation, but sustained activation over the complete turning interval, followed by recovery of the preceding pattern when straight driving resumes, is consistent with the evolution of the maneuver itself. Moreover, Experts 3 and 7 are not

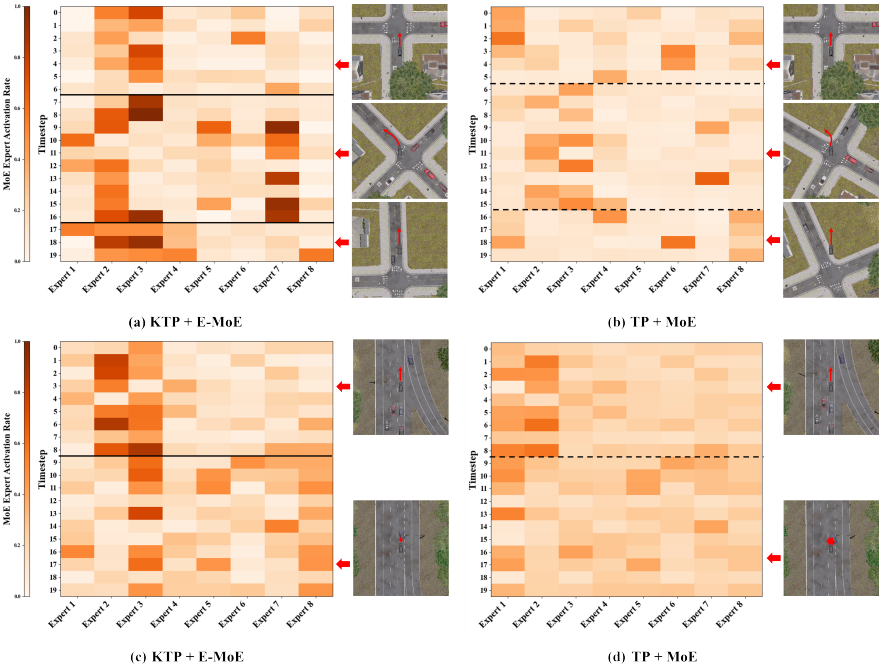


**Fig. 4:** A visual comparison of expert activations between KTP + E-MoE and the ablated TP + MoE under two representative driving scenarios. The first row shows a narrow-turn case with three sequential phases, while the second row presents an interactive merging scenario requiring timely deceleration. The heatmaps depict the normalized expert activation rates over time during closed-loop driving, where darker colors indicate higher activation probabilities. Solid lines (left) mark clear phase boundaries, whereas dashed lines (right) indicate blurred transitions. The snapshots marked by red arrows illustrate corresponding scene moments.

globally active or inactive: their relative importance reverses at the two phase boundaries. The heatmap therefore reveals a complementary division of labor conditioned on the current driving stage, rather than a fixed preference of the router for a single expert. The ablated TP+MoE lacks both the sustained turning response and the recovery pattern, making its routing less temporally aligned with the executed trajectory.

The merging case (second row) contains normal driving (steps 1–8) followed by braking (9–19) as another vehicle enters the ego lane. Expert 2 is stronger in normal driving, while Expert 8 becomes persistently active throughout braking. This switch is aligned with the required deceleration: TEX-Drive slows down and follows safely, whereas TP+MoE reacts late and collides.

The two examples also test different aspects of routing. The narrow turn is a structured maneuver with a return to an earlier phase, whereas merging requires a one-way transition triggered by another road user. Nevertheless, TEX-Drive

**Table 3:** Quantitative analysis of expert specialization. Route Ent.: mean routing entropy (lower is sharper); Temp. Persist.: mean duration of an unchanged top-1 expert (higher is more stable); Util. Ent.: entropy of aggregate expert usage.

| Method | Route Ent. ↓ | Temp. Persist. ↑ | Util. Ent. |
|---|---|---|---|
| TP + MoE | 0.78 | 1.9 | **0.83** |
| KTP + E-MoE | **0.49** | **4.8** | 0.61 |

changes its dominant expert near the behaviorally meaningful boundary in both cases. In the merging sequence, Expert 8 remains active after braking begins instead of firing only when the other vehicle first appears. This persistence indicates that the router represents the ongoing control requirement, rather than merely detecting a salient object. Conversely, the stronger response of Expert 2 before braking shows that Expert 8 is selectively recruited when deceleration becomes necessary. These observations connect expert activation to distinct control stages without claiming that an expert encodes only one universal semantic function across every scene.

To complement the qualitative observations above, we report three specialization metrics. *Routing entropy* is the normalized entropy of expert probabilities averaged across all time steps and measures decision uncertainty. *Temporal persistence* averages the lengths of consecutive runs sharing the same top-ranked expert and measures routing continuity. *Utilization entropy* is the normalized entropy of expert activation frequencies over the evaluation set and measures overall usage balance.

Table 3 shows that KTP + E-MoE reduces routing entropy from 0.78 to 0.49 and increases temporal persistence from 1.9 to 4.8 steps. The router therefore makes sharper decisions and maintains them for about 2.5× longer, rather than producing confident but rapidly oscillating assignments. Its lower utilization entropy (0.61 versus 0.83) indicates non-uniform expert usage, which is expected because driving maneuvers occur with unequal frequencies and does not alone imply capacity collapse. Such collapse would instead be reflected by persistently inactive experts or degraded task performance. In contrast, Fig. 4 shows different experts activated across behavioral phases, and the complete model substantially improves the driving results in Table 4. Taken together, these results support selective, temporally stable, and scenario-conditioned specialization rather than random routing or uniform averaging.

### 4.5 Real-World Evaluation

To evaluate our method in real-world settings, we conduct open-loop trajectory prediction experiments on the nuScenes dataset. Since several methods in Table 1 are evaluated only in simulation, we compare against representative approaches that report results on nuScenes. As shown in Table 5, we report L2 displacement error and collision rate under 1–3 second prediction horizons, together with inference efficiency measured in frames per second. Overall, our method achieves

**Table 4:** Ablation study on TEX-Drive. We report Driving Score, Success Rate, and Multi-Ability. "TP" denotes the variant using unfiltered historical features, while "MoE" indicates the model with a standard Mixture-of-Experts structure.

| KTP | E-MoE | DS | SR | MA |
|---|---|---|---|---|
| ✗ | ✗ | 53.77 | 21.81 | 19.59 |
| ✗ | MoE | 55.15 | 26.81 | 21.37 |
| ✗ | ✓ | 61.72 | 33.18 | 37.14 |
| TP | ✗ | 57.86 | 25.00 | 20.83 |
| ✓ | ✗ | 62.74 | 34.38 | 40.09 |
| TP | MoE | 58.27 | 31.25 | 33.59 |
| ✓ | ✓ | **68.22** | **40.62** | **45.09** |

**Table 5:** Real-world generalization results on the nuScenes dataset. We report open-loop L2 displacement error and collision rate at 1–3 s horizons. These results confirm the practical effectiveness of our method in real-world driving scenarios.

| Method | L2 (m) ↓ | | | | Collision (%) ↓ | | | | FPS |
|---|---|---|---|---|---|---|---|---|---|
| | 1s | 2s | 3s | Avg. | 1s | 2s | 3s | Avg. | |
| ST-P3 [19] | 1.33 | 2.11 | 2.90 | 2.11 | 0.23 | 0.62 | 1.27 | 0.71 | 2.7 |
| UniAD [20] | 0.48 | 0.96 | 1.65 | 1.03 | 0.05 | 0.17 | 0.71 | 0.31 | 1.7 |
| GenAD [50] | 0.36 | 0.83 | 1.55 | 0.91 | 0.06 | 0.23 | 1.00 | 0.43 | 2.4 |
| VAD [26] | 0.41 | 0.70 | 1.05 | 0.72 | 0.07 | 0.17 | 0.41 | 0.22 | 7.6 |
| DWM [46] | 0.43 | 0.77 | 1.20 | 0.80 | 0.10 | 0.21 | 0.48 | 0.26 | 0.3 |
| Ours | 0.45 | **0.59** | **0.91** | **0.65** | 0.08 | **0.13** | **0.39** | **0.20** | 3.3 |

competitive performance across all time scales and obtains the lowest average collision rate, indicating strong safety in real-world scenarios. In addition, TEX-Drive maintains competitive inference efficiency. Although the proposed framework incorporates a mixture-of-experts architecture that increases the overall model capacity, the sparse routing mechanism activates only a small subset of experts during inference, resulting in only a minor impact on runtime latency.

More importantly, as the prediction horizon increases from 1s to 3s, most baseline methods exhibit substantial performance degradation. In contrast, our method demonstrates the smallest increase in both L2 error and collision rate, suggesting stronger long-horizon prediction stability. These results indicate that the proposed temporal perception and experience-driven expert routing remain effective under extended prediction horizons, leading to improved robustness and safety in realistic driving environments. Qualitative visualizations of this experiment are provided in the supplementary material.

## 5 Conclusion

We presented TEX-Drive, which combines key-frame temporal perception with experience-guided MoE planning for E2E autonomous driving. Experiments on Bench2Drive and nuScenes show consistent improvements over strong baselines, with ablations validating both KTP and E-MoE. Although closed-loop evaluation is currently limited to simulation, future work will explore richer interaction modeling and extend closed-loop testing to real-world deployments.

## Acknowledgements

This work was supported in part by the New Generation Artificial Intelligence-National Science and Technology Major Project under Grant 2025ZD0122404, the National Natural Science Foundation of China under Grant U24A20291, and the Sanqin Talent Special Support Plan under Grant 2024STZZK09.

## References


1. Agand, P., Mahdavian, M., Savva, M., Chen, M.: Letfuser: Light-weight end-to-end transformer-based sensor fusion for autonomous driving with multi-task learning. arXiv preprint arXiv:2310.13135 (2023)
2. Anderson, J.R.: Act: A simple theory of complex cognition. American psychologist **51**(4), 355 (1996)
3. Baddeley, A.: The episodic buffer: a new component of working memory? Trends in cognitive sciences **4**(11), 417–423 (2000)
4. Bansal, M., Krizhevsky, A., Ogale, A.: Chauffeurnet: Learning to drive by imitating the best and synthesizing the worst. arXiv preprint arXiv:1812.03079 (2018)
5. Bojarski, M., Del Testa, D., Dworakowski, D., Firner, B., Flepp, B., Goyal, P., Jackel, L.D., Monfort, M., Muller, U., Zhang, J., et al.: End to end learning for self-driving cars. arXiv preprint arXiv:1604.07316 (2016)
6. Cai, W., Jiang, J., Wang, F., Tang, J., Kim, S., Huang, J.: A survey on mixture of experts in large language models. IEEE Transactions on Knowledge and Data Engineering (2025)
7. Chen, W., Kong, F., Chen, L., Wang, S., Wang, Z., Sun, H.: Eatnet: Efficient axial transformer network for end-to-end autonomous driving. In: 2024 IEEE 27th International Conference on Intelligent Transportation Systems (ITSC). pp. 762–767. IEEE (2024)
8. Chen, Z., Shen, Y., Ding, M., Chen, Z., Zhao, H., Learned-Miller, E.G., Gan, C.: Mod-squad: Designing mixtures of experts as modular multi-task learners. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11828–11837 (2023)
9. Chi, L., Mu, Y.: Deep steering: Learning end-to-end driving model from spatial and temporal visual cues. arXiv preprint arXiv:1708.03798 (2017)
10. Chitta, K., Prakash, A., Jaeger, B., Yu, Z., Renz, K., Geiger, A.: Transfuser: Imitation with transformer-based sensor fusion for autonomous driving. IEEE transactions on pattern analysis and machine intelligence **45**(11), 12878–12895 (2022)
11. Codevilla, F., Müller, M., López, A., Koltun, V., Dosovitskiy, A.: End-to-end driving via conditional imitation learning. In: 2018 IEEE international conference on robotics and automation (ICRA). pp. 4693–4700. IEEE (2018)
12. Dai, D., Deng, C., Zhao, C., Xu, R., Gao, H., Chen, D., Li, J., Zeng, W., Yu, X., Wu, Y., et al.: Deepseekmoe: Towards ultimate expert specialization in mixture-of-experts language models. arXiv preprint arXiv:2401.06066 (2024)
13. Du, N., Huang, Y., Dai, A.M., Tong, S., Lepikhin, D., Xu, Y., Krikun, M., Zhou, Y., Yu, A.W., Firat, O., et al.: Glam: Efficient scaling of language models with mixture-of-experts. In: International conference on machine learning. pp. 5547–5569. PMLR (2022)
14. Eraqi, H.M., Moustafa, M.N., Honer, J.: End-to-end deep learning for steering autonomous vehicles considering temporal dependencies. arXiv preprint arXiv:1710.03804 (2017)
15. Fedus, W., Zoph, B., Shazeer, N.: Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. Journal of Machine Learning Research **23**(120), 1–39 (2022)
16. Gu, N., Zhang, Z., Feng, Y., Chen, Y., Fu, P., Lin, Z., Wang, S., Sun, Y., Wu, H., Wang, W., et al.: Elastic moe: Unlocking the inference-time scalability of mixture-of-experts. arXiv preprint arXiv:2509.21892 (2025)

17. Gu, Z., Li, Z., Di, X., Shi, R.: An lstm-based autonomous driving model using a waymo open dataset. Applied Sciences **10**(6), 2046 (2020)
18. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9729–9738 (2020)
19. Hu, S., Chen, L., Wu, P., Li, H., Yan, J., Tao, D.: St-p3: End-to-end vision-based autonomous driving via spatial-temporal feature learning. In: European Conference on Computer Vision. pp. 533–549. Springer (2022)
20. 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: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 17853–17862 (2023)
21. Huang, Q., An, Z., Zhuang, N., Tao, M., Zhang, C., Jin, Y., Xu, K., Chen, L., Huang, S., Feng, Y.: Harder tasks need more experts: Dynamic routing in moe models. arXiv preprint arXiv:2403.07652 (2024)
22. Jia, X., Gao, Y., Chen, L., Yan, J., Liu, P.L., Li, H.: Driveadapter: Breaking the coupling barrier of perception and planning in end-to-end autonomous driving. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7953–7963 (2023)
23. Jia, X., Wu, P., Chen, L., Xie, J., He, C., Yan, J., Li, H.: Think twice before driving: Towards scalable decoders for end-to-end autonomous driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21983–21994 (2023)
24. Jia, X., Yang, Z., Li, Q., Zhang, Z., Yan, J.: Bench2drive: Towards multi-ability benchmarking of closed-loop end-to-end autonomous driving. Advances in Neural Information Processing Systems **37**, 819–844 (2024)
25. Jia, X., You, J., Zhang, Z., Yan, J.: Drivetransformer: Unified transformer for scalable end-to-end autonomous driving. In: International Conference on Learning Representations (ICLR) (2025)
26. Jiang, B., Chen, S., Xu, Q., Liao, B., Chen, J., Zhou, H., Zhang, Q., Liu, W., Huang, C., Wang, X.: Vad: Vectorized scene representation for efficient autonomous driving. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8340–8350 (2023)
27. Jiang, C., Cornman, A., Park, C., Sapp, B., Zhou, Y., Anguelov, D., et al.: Motiondiffuser: Controllable multi-agent motion prediction using diffusion. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9644–9653 (2023)
28. Kendall, A., Hawke, J., Janz, D., Mazur, P., Reda, D., Allen, J.M., Lam, V.D., Bewley, A., Shah, A.: Learning to drive in a day. In: 2019 international conference on robotics and automation (ICRA). pp. 8248–8254. IEEE (2019)
29. Lepikhin, D., Lee, H., Xu, Y., Chen, D., Firat, O., Huang, Y., Krikun, M., Shazeer, N., Chen, Z.: Gshard: Scaling giant models with conditional computation and automatic sharding. arXiv preprint arXiv:2006.16668 (2020)
30. Lewis, M., Bhosale, S., Dettmers, T., Goyal, N., Zettlemoyer, L.: Base layers: Simplifying training of large, sparse models. In: International Conference on Machine Learning. pp. 6265–6274. PMLR (2021)
31. Li, Q., Jia, X., Wang, S., Yan, J.: Think2drive: Efficient reinforcement learning by thinking with latent world model for autonomous driving (in carla-v2). In: European Conference on Computer Vision. pp. 142–158. Springer (2024)

32. 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: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 12037–12047 (2025)
33. Liu, L., Song, Z., Pan, H., Yang, L., Jia, C.: Two tasks, one goal: Uniting motion and planning for excellent end to end autonomous driving performance. arXiv preprint arXiv:2504.12667 (2025)
34. Mu, S., Lin, S.: A comprehensive survey of mixture-of-experts: Algorithms, theory, and applications. arXiv preprint arXiv:2503.07137 (2025)
35. Norman, D.A., Shallice, T.: Attention to action: Willed and automatic control of behavior. In: Consciousness and self-regulation: Advances in research and theory volume 4, pp. 1–18. Springer (1986)
36. Perez, E., Strub, F., De Vries, H., Dumoulin, V., Courville, A.: Film: Visual reasoning with a general conditioning layer. In: Proceedings of the AAAI conference on artificial intelligence. vol. 32 (2018)
37. Pini, S., Perone, C.S., Ahuja, A., Ferreira, A.S.R., Niendorf, M., Zagoruyko, S.: Safe real-world autonomous driving by learning to predict and plan with a mixture of experts. arXiv preprint arXiv:2211.02131 (2022)
38. Prakash, A., Chitta, K., Geiger, A.: Multi-modal fusion transformer for end-to-end autonomous driving. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7077–7087 (2021)
39. Renz, K., Chitta, K., Mercea, O.B., Koepke, A., Akata, Z., Geiger, A.: Plant: Explainable planning transformers via object-level representations. arXiv preprint arXiv:2210.14222 (2022)
40. Shen, S., Hou, L., Zhou, Y., Du, N., Longpre, S., Wei, J., Chung, H.W., Zoph, B., Fedus, W., Chen, X., et al.: Mixture-of-experts meets instruction tuning: A winning combination for large language models. arXiv preprint arXiv:2305.14705 (2023)
41. Sun, Q., Wang, H., Zhan, J., Nie, F., Wen, X., Xu, L., Zhan, K., Jia, P., Lang, X., Zhao, H.: Generalizing motion planners with mixture of experts for autonomous driving. In: 2025 IEEE International Conference on Robotics and Automation (ICRA). pp. 6033–6039. IEEE (2025)
42. Sun, W., Lin, X., Shi, Y., Zhang, C., Wu, H., Zheng, S.: Sparsedrive: End-to-end autonomous driving via sparse scene representation. In: 2025 IEEE International Conference on Robotics and Automation (ICRA). pp. 8795–8801. IEEE (2025)
43. Toromanoff, M., Wirbel, E., Moutarde, F.: End-to-end model-free reinforcement learning for urban driving using implicit affordances. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7153–7162 (2020)
44. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
45. Wan, C., Cui, Y., Du, J., Yang, S., Bai, Y., Yi, P., Li, N., Huang, Y.: Geminus: Dual-aware global and scene-adaptive mixture-of-experts for end-to-end autonomous driving. arXiv preprint arXiv:2507.14456 (2025)
46. Wang, Y., He, J., Fan, L., Li, H., Chen, Y., Zhang, Z.: Driving into the future: Multiview visual forecasting and planning with world model for autonomous driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 14749–14759 (June 2024)

47. Wu, P., Jia, X., Chen, L., Yan, J., Li, H., Qiao, Y.: Trajectory-guided control prediction for end-to-end autonomous driving: A simple yet strong baseline. Advances in Neural Information Processing Systems **35**, 6119–6132 (2022)
48. Wu, X., Huang, S., Wang, W., Ma, S., Dong, L., Wei, F.: Multi-head mixture-of-experts. Advances in Neural Information Processing Systems **37**, 94073–94096 (2024)
49. Xu, L., Yu, J., Peng, X., Chen, Y., Li, W., Yoo, J., Chunag, S., Lee, D., Ji, D., Zhang, C.: Mose: Skill-by-skill mixture-of-expert learning for autonomous driving. arXiv e-prints pp. arXiv–2507 (2025)
50. Zheng, W., Song, R., Guo, X., Zhang, C., Chen, L.: Genad: Generative end-to-end autonomous driving. In: European Conference on Computer Vision. pp. 87–104. Springer (2024)
51. Zheng, Y., Liang, R., Zheng, K., Zheng, J., Mao, L., Li, J., Gu, W., Ai, R., Li, S.E., Zhan, X., Liu, J.: Diffusion-based planning for autonomous driving with flexible guidance. In: ICLR (2025), https://openreview.net/forum?id=wM2sfVgMDH
52. Zhou, Y., Lei, T., Liu, H., Du, N., Huang, Y., Zhao, V., Dai, A.M., Le, Q.V., Laudon, J., et al.: Mixture-of-experts with expert choice routing. Advances in Neural Information Processing Systems **35**, 7103–7114 (2022)