PriorEye: Geospatial Visual Priors for End-to-End Autonomous Driving

Kyuhwan Yeon, Benjamin Ramtoula, and Daniele De Martini

Mobile Robotics Group, University of Oxford, United Kingdom
{kyuhwan, benjamin, daniele}@robots.ox.ac.uk

Abstract. Most end-to-end autonomous driving methods rely solely on instantaneous sensor observations, limiting them to reactive behavior without the anticipatory foresight human drivers employ through prior experience. We introduce geospatial visual priors, street-level visual context anchored to the intended driving route, providing visual-spatial foresight independent of real-time sensors. We propose a memory augmentation module featuring a dual-memory architecture and an adaptive memory gate, which can be easily integrated into existing end-to-end approaches. This design pairs a contextual memory for retrieved priors with a persistent fallback memory, and dynamically regulates the influence of memories based on current state compatibility. Evaluated on the NAVSIM-v2 benchmark, our approach consistently improves performance across diverse end-to-end baselines. Furthermore, because these priors are independent of onboard sensors, our method inherently improves robustness against sensor corruption, while the dual-memory design ensures safe fallback when the retrieved priors themselves become unreliable. Our project page is available at `https://ori-mrg.github.io/PriorEye`.

Keywords: End-to-End Autonomous Driving · Priors · Memory

1 Introduction

End-to-end (E2E) autonomous driving has emerged as a compelling paradigm due to its scalability and conceptual simplicity, directly mapping raw sensor observations to driving actions or trajectories [6,53]. By enabling joint optimization and avoiding hand-crafted intermediate representations, E2E approaches mitigate cascading errors inherent to traditional modular pipelines [8,19,25,26,30,38, 48, 51, 59, 70]. Recent advances in vision-language models have further improved performance and enhanced semantic interpretability [20, 27, 41, 43], establishing E2E driving as a prominent approach for autonomous driving.

Despite these advances, a fundamental limitation remains: most existing E2E driving policies are short-horizon reactive, relying only on a few seconds of sensor observations as input [20]. However, effective driving requires reasoning over extended time horizons. Human drivers naturally achieve this by utilizing visual and spatial foresight derived from prior experience [13, 40, 46]. For instance,

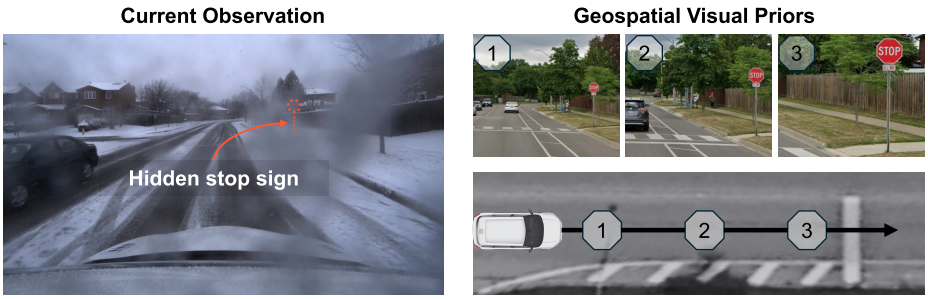


Fig. 1: Motivation for geospatial visual priors. *Left:* A current onboard observation from the Boreas dataset [2], where the camera is occluded by snow and critical features such as stop signs become invisible. Methods that rely solely on instantaneous observations are vulnerable in such conditions. *Right:* Geospatial visual priors, street-level images collected offline along the intended route. Because these priors are independent of current sensor conditions, the stop sign remains clearly visible in the retrieved imagery, enabling anticipatory driving even when onboard perception fails.

guided by driving memory, a human driver can proactively merge into an adjacent lane well before a lane ends, even when the lane drop is not yet visible. Similarly, they can decelerate in anticipation of a known sharp curve hidden behind obstacles. This prior information remains uncorrupted by severe weather or visual occlusions. In contrast, current E2E systems, constrained by their exclusive focus on instantaneous sensor observations, lack such anticipatory behavior in complex scenarios. Moreover, this over-reliance on real-time visual cues renders them highly susceptible to sensor corruption, as they lack the persistent contextual evidence necessary to stabilize perception against temporary disturbances.

Inspired by these observations, we propose a mechanism to augment existing end-to-end driving solutions with geospatial visual priors. Acting as a form of driving memory, these priors explicitly encode visual and spatial foresight. In our formulation, spatial priors represent the intended driving route and its associated spatial anchors along that route, while visual priors correspond to offline-collected street-level images that capture the visual context of the driving environment. Geospatial visual priors are constructed by anchoring visual priors to spatial priors. As illustrated in Fig. 1, these priors provide contextual information about the upcoming driving environment, supplying critical road elements (*e.g.*, stop signs and road markings) that instantaneous onboard sensors may miss due to occlusions or sensor degradation. By coupling street-level visual context with the spatial structure of the intended route, the proposed design enables anticipatory, efficient, and robust driving even under occlusions and limited visibility.

Inspired by memory-augmented Large Language Models (LLMs) [1, 56, 60], we formulate geospatial visual priors as long-term memory. Specifically, we introduce a model-agnostic memory augmentation framework where these priors

constitute the contextual memory, providing the driving policy with extended visual-spatial context beyond the immediate sensor view. To ensure robustness against potential retrieval errors, our design complements this with a persistent memory, acting as a fallback reference. An adaptive memory gate explicitly balances the reliance between these memory sources and real-time observations based on similarity cues. Importantly, the proposed module can be seamlessly integrated into a wide range of E2E planners without modifying their core architectures. Our contributions are threefold:

- We introduce geospatial visual priors that provide spatially grounded visual foresight for end-to-end driving.
- We propose a model-agnostic dual memory architecture with an adaptive memory gate that effectively fuses geospatial visual priors while preventing over-reliance on misaligned or corrupted memory.
- We validate that by leveraging priors that are independent of instantaneous onboard sensor observations, our method significantly improves driving performance across diverse E2E planners and enhances robustness under sensor degradation.

2 Related Work

End-to-End Autonomous Driving. Existing E2E driving methods can be broadly categorized by their planning formulations. Regression-based approaches directly predict future trajectories from sensor inputs, offering simplicity and efficiency [4,8,18,19,22,51]. Generative planners, including diffusion-based models, model the distribution of feasible trajectories to improve diversity and robustness [24,37,38,64]. Trajectory scoring methods employ a learned scorer to select the most suitable trajectory from a set of candidate trajectories, enabling effective handling of multi-modal and non-convex planning uncertainty [30,31,34–36, 67]. Recent works further incorporate vision-language models to enhance semantic reasoning [11, 20, 27, 43, 45, 47, 66], and widely explore 3D Gaussian Splatting and world models for data augmentation, model evaluation, and performance improvement [3, 14, 16, 21, 32, 33, 54, 57, 69]. Despite these advances, a common limitation persists: most E2E methods operate within a short input horizon of only a few seconds, restricting them to purely reactive behavior without the ability to reason about upcoming road context beyond the immediate sensor view.

Priors for Autonomous Driving. Prior knowledge in autonomous driving typically refers to information that is not directly observable from online sensors and is obtained offline or from prior experience. Driving experience or memory has been leveraged for high-level reasoning, enabling models to reuse past knowledge from previously encountered scenarios and improve robustness in complex situations [39,58,68]. Spatial priors such as maps, aerial images, and route information encode road topology and traffic rules, primarily used for upstream tasks

like online map construction to provide long-range structural guidance beyond the current field of view [28, 42, 65]. Visual priors such as street-level images exploit offline visual data to capture appearance-level cues and scene context of driving environments [23].

Although prior information for autonomous driving has been widely explored, existing methods primarily leverage priors for upstream tasks such as online HD map construction or high-level reasoning. In contrast, the direct integration of priors into E2E planning remains largely underexplored. Furthermore, spatial and visual priors are typically considered in isolation, leaving the planner unable to access visual context along the intended route. To address these gaps, our work introduces geospatial visual priors that couple offline visual context with route-level spatial information and inject them directly into E2E driving policies. As a result, our method explicitly provides visual-spatial foresight derived from priors to the E2E planner.

3 Methodology

We consider E2E driving systems that receive raw sensor observations, the ego-vehicle state, and a high-level navigational command to produce a planned trajectory, and assume access to a lane connectivity graph of the deployment area.

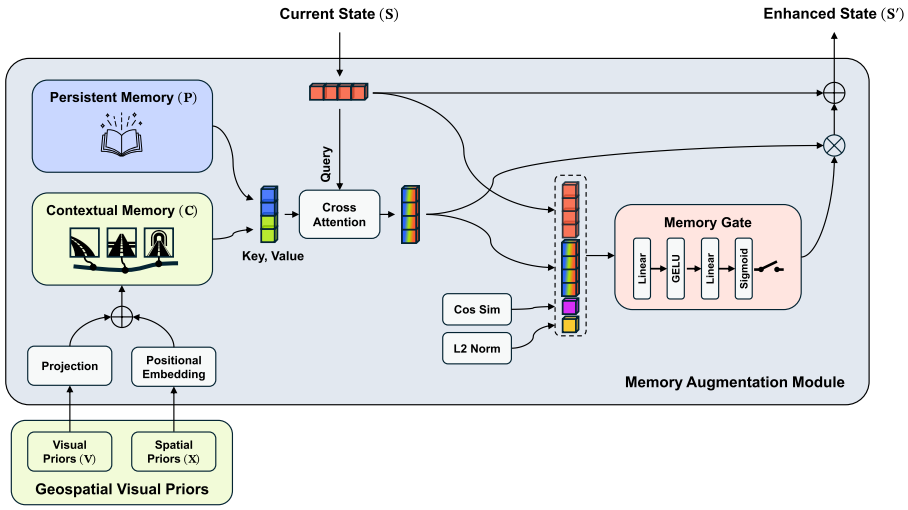
To enable such systems to leverage geospatial visual priors, we first construct a global offline repository of street-level visual embeddings spanning the target area. For a given task, we retrieve relevant embeddings and anchor them to the intended route to form geospatial visual priors (Section 3.1).

To incorporate these priors, we introduce the memory augmentation module (Section 3.2; Fig. 2a), a generic state-editing mechanism that takes the intermediate state of an E2E model along with the retrieved geospatial visual priors and produces an enhanced state. Fig. 2b illustrates the general integration scheme into existing E2E pipelines.

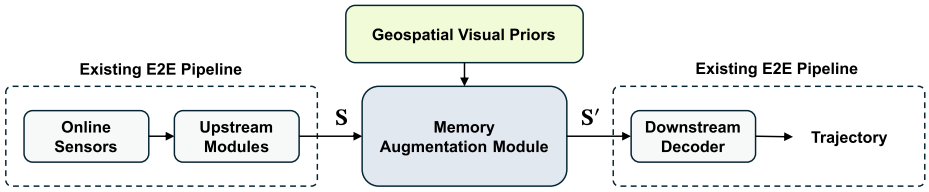
3.1 Geospatial Visual Priors and Memory Bank

Geospatial visual priors are visual-spatial representations that anchor street-level visual context to the intended driving route, providing the E2E model with foresight of the upcoming environment. To construct these priors, we introduce the memory bank, which stores pairs of embeddings of street-level images and their corresponding locations across the entire map of the target deployment area.

Construction of the Memory Bank. Given a target deployment area, we traverse each lane segment and sample street-level images along the lane centerline at a fixed spatial interval. Each sampled image is encoded by a frozen visual backbone [55] and stored as a lane-associated memory. The proposed framework only requires coarsely distributed street-level images and is robust to noise, as the goal is to provide semantically meaningful visual cues for navigation rather than exact geometric reconstruction.



(a) Architecture of the Memory Augmentation Module



(b) Integration into existing E2E pipelines

Fig. 2: Overview of the proposed framework. (a) Detailed architecture of the memory augmentation module. (b) The module takes the current driving state $\mathbf{S}$ from upstream modules and geospatial visual priors as input, and produces an enhanced state $\mathbf{S}'$ for the downstream decoder. Dashed boxes indicate existing components that remain unmodified.

Querying Geospatial Visual Priors from the Memory Bank. During both training and inference, only a subset of visual embeddings is relevant to the current driving context and needs to be retrieved from the memory bank. Given the current vehicle state, we first identify the lane on which the vehicle is currently located. Starting from the current lane, we perform a depth-first search (DFS) over the lane connectivity graph to generate candidate lane sequences extending forward.

Next, based on the high-level navigational intention (`Turn Left`, `Turn Right`, and `Straight`), we select the corresponding lane sequence from the candidate set and retrieve the embeddings stored along it from the memory bank. Each retrieved visual embedding $\mathbf{v}_i \in \mathbb{R}^{D_m}$ is anchored to a relative (x, y) coordinate $\mathbf{x}_i \in \mathbb{R}^2$ along the lane, where D_m is the dimensionality of the visual embedding space. Stacking N such pairs yields the visual priors $\mathbf{V} \in \mathbb{R}^{N \times D_m}$ and spatial

priors $\mathbf{X} \in \mathbb{R}^{N\times 2}$, which together constitute the queried geospatial visual priors:

$$\{(\mathbf{v}_i, \mathbf{x}_i)\}_{i=1}^{N}. \tag{1}$$

3.2 Memory Augmentation Module

We propose a memory augmentation module (Fig. 2a) to integrate the queried geospatial visual priors into the E2E system. It is composed of two key components: a dual memory architecture that encodes the priors into contextual memory, complemented by a persistent memory, and a memory fusion mechanism that selectively integrates these memories into the current driving state via cross-attention and adaptive gating.

Dual Memory Architecture. We adopt a dual memory architecture inspired by Titans [1], allowing us to treat priors as retrievable memories. The proposed design decomposes memory into two complementary components: contextual memory, which provides relevant geospatial visual priors, and persistent memory, which serves as a default reference when contextual memory is not helpful.

Contextual Memory. Contextual memory is an input-dependent representation that integrates geospatial visual priors. It is constructed as

$$\mathbf{C} = \phi(\mathbf{V}) + \psi(\mathbf{X}), \tag{2}$$

where $\phi(\cdot)$ projects the visual priors into the E2E model's internal embedding space, and $\psi(\cdot)$ denotes a 2D sinusoidal positional embedding that encodes spatial priors in the form of relative coordinates into the same embedding space. Their additive combination enables the contextual memory to jointly capture semantic appearance information and coarse spatial layout.

Persistent Memory. However, contextual memory derived from street-level imagery can be corrupted or misaligned due to seasonal changes or geometric mismatches [23]. In the absence of alternative references, the attention mechanism is inevitably forced to attend to these corrupted tokens. This issue mirrors the attention sink problem in Transformer architectures [12, 62], where models require dedicated tokens to absorb attention mass when informative features are missing. Inspired by this, we include persistent memory [1, 50], a set of learnable tokens designed to serve as a robust fallback against contextual corruption:

$$\mathbf{P} \in \mathbb{R}^{K\times D}, \tag{3}$$

where K denotes the number of persistent tokens and D is the embedding dimension. These tokens are independent of the current input and shared across all samples.

Memory Fusion. To selectively integrate the dual memories into the current driving state, we employ an attention-based mechanism with adaptive gating (Fig. 2a). We define the current driving state as a set of query tokens

$$\mathbf{S} \in \mathbb{R}^{Q \times D}, \tag{4}$$

where Q denotes the number of state tokens produced by the E2E driving model and D is the model's embedding dimension. These tokens encode intermediate driving state features, including perceptual representations and ego-vehicle information.

We model memory retrieval as an associative process using multi-head cross-attention [1, 52, 61, 71], with the current driving state as queries and the memory representations as keys and values. Specifically, we concatenate persistent memory and contextual memory along the token dimension and apply cross-attention as follows:

$$\mathbf{M} = [\mathbf{P}; \mathbf{C}], \tag{5}$$

$$\tilde{\mathbf{S}} = \mathrm{Attn}(\mathbf{S}, \mathbf{M}, \mathbf{M}), \tag{6}$$

where $\tilde{\mathbf{S}}$ denotes the memory-attended driving state. This operation allows the model to dynamically attend to both contextual and persistent memory.

Memory Gate. We introduce an adaptive gate to regulate memory influence. Drawing from local inference enhancement techniques [7, 9], we explicitly model the compatibility between the current state $\mathbf{S}$ and the memory-attended state $\tilde{\mathbf{S}}$ to further improve performance. To measure complementary semantic consistency, we compute the Euclidean distance d [17] and cosine similarity c [58, 68]:

$$d = \frac{\|\mathbf{S} - \tilde{\mathbf{S}}\|_2}{\sqrt{D}}, \quad c = \cos(\mathbf{S}, \tilde{\mathbf{S}}), \tag{7}$$

$$\mathbf{G} = \sigma\Big(f_g\Big([\mathbf{S}; \tilde{\mathbf{S}}; d; c]\Big)\Big), \tag{8}$$

where $\mathbf{G}$ denotes the gate, $\sigma(\cdot)$ is the sigmoid function, and $f_g(\cdot)$ is a lightweight MLP. The gate adaptively modulates memory contribution based on the estimated consistency. To stabilize training, we initialize the gate bias toward low values so that it starts in a near-closed state [15, 49].

Finally, the fused driving state is obtained via a gated residual update:

$$\mathbf{S}' = \mathbf{S} + \mathbf{G} \odot \tilde{\mathbf{S}}, \tag{9}$$

where $\odot$ denotes element-wise multiplication. The resulting updated state $\mathbf{S}'$ is then fed back into the downstream decoder of the E2E model, effectively replacing the original state $\mathbf{S}$ with a memory-informed representation for final trajectory prediction. The proposed module is jointly trained with each baseline model using their original training objectives, without introducing any additional loss terms.

Table 1: Performance on the NAVSIM-v2 `navhard-two-stage` benchmark [3, 10]. S1 and S2 denote Stage 1 and Stage 2, respectively. The "+ PriorEye" suffix indicates models augmented with our proposed module.

Method	Stg.	NC↑	DAC↑	DDC↑	TLC↑	EP↑	TTC↑	LK↑	HC↑	EC↑	EPDMS↑
LTF [8]	S 1	95.1	80.7	98.8	99.6	84.3	94.2	92.7	97.6	79.6	24.7
	S 2	79.4	68.3	83.4	98.1	86.6	77.0	43.4	95.9	73.2	
LTF + PriorEye	S 1	95.6	86.4	99.4	99.3	83.6	95.1	96.4	97.6	77.3	**32.4** (+31.2%)
	S 2	78.8	76.6	86.0	98.3	85.2	75.5	50.7	96.2	72.1	
GTRS-DP [36]	S 1	95.0	80.7	97.0	99.9	83.1	94.1	93.1	97.6	69.9	26.3
	S 2	81.3	72.6	83.3	98.5	83.2	78.7	45.7	96.3	62.2	
GTRS-DP + PriorEye	S 1	95.3	82.0	96.4	99.8	81.5	95.1	93.1	97.1	68.0	**30.1** (+14.4%)
	S 2	83.9	75.3	85.7	98.5	82.3	80.8	49.6	96.3	66.9	
GTRS-Dense [36]	S 1	98.4	95.8	99.4	99.3	72.8	98.9	94.9	96.7	40.4	44.9
	S 2	91.2	89.2	94.6	98.8	68.8	90.1	55.0	93.6	49.6	
GTRS-Dense + PriorEye	S 1	97.6	97.1	100.0	99.8	76.1	97.8	97.3	97.8	52.4	**48.6** (+8.2%)
	S 2	89.5	90.0	95.0	99.0	77.1	87.1	55.8	95.3	51.3	
DrivoR [30]	S 1	99.1	98.2	99.3	100.0	69.7	98.9	93.6	97.6	63.6	48.9
	S 2	91.8	89.4	94.1	99.3	59.0	89.9	51.3	98.7	76.5	
DrivoR + PriorEye	S 1	98.7	99.6	99.6	99.8	74.5	98.4	95.6	97.6	73.8	**49.6** (+1.4%)
	S 2	87.2	89.9	93.3	99.1	70.5	85.8	52.3	98.8	74.5	

4 Experiments

4.1 Experimental Setup

Dataset. We train and evaluate our method on the NAVSIM v2 benchmark [3, 10], a large-scale real-world end-to-end autonomous driving benchmark built upon nuPlan [29]. All models are trained on the full `navtrain` split and evaluated on the `navhard-two-stage` and `navtest` splits, following the official benchmark protocol. For ablation studies, we reserve a subset of `navtrain` for validation, training models on the remaining data and evaluating them on the held-out validation split.

Metrics. We report performance using the Extended Predictive Driver Model Score (EPDMS), the official composite metric of the NAVSIM v2 benchmark [3, 10]. EPDMS aggregates several safety and comfort sub-metrics: No-at-fault Collision (NC), Drivable Area Compliance (DAC), Driving Direction Compliance (DDC), Traffic Light Compliance (TLC), Ego Progress (EP), Time-to-Collision (TTC), Lane Keeping (LK), History Comfort (HC), and Extended Comfort (EC). The score for `navhard-two-stage` is computed via a two-stage pseudo-simulation protocol: Stage 1 evaluates planned trajectories on real-world observations, while Stage 2 assesses robustness under controlled perturbations using synthetic observations. In contrast, `navtest` follows only the Stage 1 evaluation.

Baselines and Terminology. We evaluate the model-agnostic nature of our proposed framework by integrating it into four representative E2E driving base-

Table 2: Performance on the NAVSIM-v2 `navtest` benchmark [3, 10].

Method	NC↑	DAC↑	DDC↑	TLC↑	EP↑	TTC↑	LK↑	HC↑	EC↑	EPDMS↑
LTF [8]	97.8	92.5	99.1	99.8	87.8	97.4	96.3	98.3	87.0	84.3
LTF + PriorEye	97.6	95.1	99.5	99.8	87.8	97.4	97.4	98.3	87.0	**86.8** (+2.5)
GTRS-DP [36]	97.3	92.3	98.6	99.7	86.4	97.0	95.3	98.1	79.4	82.2
GTRS-DP + PriorEye	97.5	92.7	98.7	99.8	85.9	97.1	95.6	98.2	78.3	**82.5** (+0.3)
GTRS-Dense [36]	99.1	98.3	99.6	99.9	82.8	99.2	94.8	98.0	47.5	85.4
GTRS-Dense + PriorEye	98.7	99.1	99.8	99.9	85.0	98.8	96.7	98.2	67.5	**88.8** (+3.4)
DrivoR [30]	99.4	99.2	99.8	99.9	76.6	99.4	94.9	98.3	71.7	87.2
DrivoR + PriorEye	99.1	99.6	99.8	99.9	81.5	99.1	95.9	98.3	82.5	**89.9** (+2.7)

lines spanning diverse paradigms: regression-based (LTF [8]), diffusion-based (GTRS-DP [36]), and scoring-based (GTRS-Dense [36] and DrivoR [30]). Models augmented with our proposed module are denoted with the "+ PriorEye" suffix. For qualitative results and robustness evaluations, we default to GTRS-Dense as the baseline and GTRS-Dense + PriorEye as our method. For ablation studies, we use LTF as the baseline, as its lightweight training enables efficient exploration of the many design configurations.

Street-Level Imagery. We leverage the Google Street View (GSV) Image API to query street-level imagery for the covered regions, which serves as the source for constructing the memory bank. GSV imagery is not available at every location and does not guarantee precise spatial alignment. As a result, some queried locations may return no image, while others may suffer from vertical (z-axis) misalignment or spatial mislocalization, causing multiple distinct positions along a lane to be mapped to the same street-view image [23]. In our setting, such missing or redundant observations are not problematic, as discussed in Section 3.2. Although we choose GSV as it provides a convenient and scalable source of street-level imagery that covers the regions associated with our dataset, the proposed framework is not tied to a specific API or data provider. Any sparsely sampled collection of location-anchored driving images can serve the same role.

Implementation Details. For the memory bank, we encode street-level imagery using SigLIP2 [55] (base-patch16-256) at a fixed spatial interval of 5 meters, resulting in a total size of 939 MB covering approximately 6.5 km^2 of the target area [29].

At inference, $N = 20$ geospatial visual priors are retrieved, corresponding to a look-ahead range of approximately 100 meters. The memory augmentation module introduces only 713K additional parameters, representing a negligible overhead of 1.3%, 0.6%, 0.9%, and 1.7% relative to LTF (56.7M), GTRS-DP (116M), GTRS-Dense (83.6M), and DrivoR (41.5M), respectively.

For each baseline, the memory augmentation module edits the intermediate state formed by concatenating the encoded ego-status with perceptual features specific to each architecture: BEV features for LTF and GTRS-DP, image features for GTRS-Dense, and scene tokens for DrivoR.

For training, we follow the existing protocols of each baseline [8, 30, 36]. All models are trained on 8 NVIDIA RTX 5090 GPUs with per-GPU batch sizes of 64, 16, 16, and 12 for LTF, GTRS-DP, GTRS-Dense, and DrivoR, trained for 100, 80, 40, and 20 epochs, respectively.

4.2 Main Results

Table 1 reports the results on the NAVSIM-v2 `navhard-two-stage` benchmark. Across all four end-to-end driving baselines, incorporating the proposed module consistently improves performance, demonstrating the model-agnostic effectiveness of our approach. Specifically, memory augmentation yields relative EPDMS improvements of 31.2%, 14.4%, 8.2%, and 1.4% over LTF, GTRS-DP, GTRS-Dense, and DrivoR, respectively.

We observe a consistent improvement on the `navtest` split (Table 2), where PriorEye again improves all four baselines, with EPDMS gains of 2.5, 0.3, 3.4, and 2.7 points. This consistency across both splits confirms that the benefit of geospatial visual priors is not specific to a single evaluation setting.

Beyond quantitative gains, geospatial visual priors effectively complement instantaneous perception under partial observability, such as occluded intersections. A qualitative example is provided in Fig. 3. By anticipating upcoming road features before they become visible, the model can maintain a more conservative speed, allowing for smoother deceleration when potential hazards arise.

4.3 Robustness Analysis

We further evaluate robustness under (i) instantaneous sensor corruption and (ii) corrupted geospatial visual priors.

Sensor Robustness. Sensor corruption is a common and unavoidable challenge in real-world autonomous driving [5]. This is particularly challenging for current E2E systems relying on instantaneous data, as the only source of information about the environment then becomes unreliable. We hypothesize that geospatial visual priors can improve robustness in such conditions, as retrieved driving memory remains unaffected by onboard sensor degradation. To verify this, we synthetically apply five different corruptions to camera observations. This is achieved by using off-the-shelf transparent overlays captured from a real camera and representing different sensor degradations. Each overlay is combined with the original images by using additive blending (Linear Dodge) for handprints, or alpha compositing for all other corruptions. Figure 4 shows examples of each synthetic corruption applied to the front-center camera.

Fig. 3: Qualitative visualization of planned trajectories in a left-turn scenario. *Top:* Projected future trajectories overlaid on the camera view. The average speed over the prediction horizon is shown in the top-right corner. *Bottom-Left:* BEV map showing the predicted trajectories and spatial priors sampled at 5-meter intervals. *Bottom-Right:* Retrieved street-level visual priors indexed along the intended route. In this scenario, the downstream road contains a pedestrian crosswalk not yet visible to onboard sensors. Relying solely on instantaneous observations, the baseline (red) maintains an average speed of 24.2 km/h. In contrast, our method (green) leverages geospatial visual priors to anticipate the upcoming crosswalk, reflected in the retrieved memories (indices 10–15), and drives at a more cautious 18.2 km/h. This demonstrates how geospatial priors effectively complement instantaneous perception to enable anticipatory driving behavior.

Table 3 reports the quantitative robustness results. Our method consistently outperforms the baseline across all corruption types. The improvement is most pronounced under severe mud corruption, where the baseline suffers a 53.2% performance drop while our method reduces this to 32.5%. Even under mild corruptions such as fingerprints and handprints, our method achieves notably smaller degradation. On average, our method reduces the relative performance degrada-

tion from 20.6% to 13.9% across all corruption types. These results demonstrate that geospatial visual priors provide a complementary source of information, improving robustness across diverse corruption scenarios.

Fig. 4: Example of synthetic sensor corruptions applied to the front-center camera. From left to right, top to bottom: Nominal, Fingerprints, Handprints, Frost, Mud (Mild), Mud (Heavy).

Table 3: Sensor robustness evaluation on `navhard-two-stage` (EPDMS). Degradation relative to nominal is shown in parentheses.

	Nominal	Fingerprints	Handprints	Frost	Mud (Mild)	Mud (Heavy)
Baseline	44.9	40.0 (-10.9%)	40.3 (-10.2%)	34.6 (-22.9%)	42.3 (-5.8%)	21.0 (-53.2%)
Ours	**48.6**	**45.1 (-7.2%)**	**45.4 (-6.6%)**	**38.1 (-21.6%)**	**47.9 (-1.4%)**	**32.8 (-32.5%)**

Geospatial Visual Prior Corruption. In real-world deployment, geospatial visual priors may become unreliable due to outdated imagery, appearance drift, spatial misalignment, or missing observations. Since contextual memory ($\mathbf{C}$) is constructed from retrieved visual–spatial pairs $(\mathbf{V}, \mathbf{X})$, such failures correspond to corruption of $\mathbf{C}$. As discussed in Section 3.2, our framework incorporates persistent memory ($\mathbf{P}$) to mitigate this. To evaluate this design choice, we simulate three corruption scenarios. For visual corruption, we replace the retrieved visual embeddings with those queried from a randomly sampled location 500 meters away, introducing semantically irrelevant appearance features. For spatial corruption, we replace the associated spatial coordinates with random positions at the same 500-meter offset, while keeping the visual embeddings intact. The combined corruption applies both perturbations simultaneously.

Table 4 reports the attention behavior of the full dual-memory model and the driving performance across ablation variant models under corruption. The

model adaptively reallocates attention towards the fallback persistent memory **P** as contextual memory becomes unreliable, reaching 0.97 under visual corruption and 0.99 under combined corruption, effectively suppressing misleading contextual cues.

Under normal conditions, both the context-only model and the full dual-memory model outperform the baseline. Under corruption, however, the context-only model suffers severe degradation, whereas the dual-memory model remains robust, reducing performance drops from 19.6% to 6.8% under visual corruption and from 18.6% to 6.6% under combined corruption. Notably, even under corrupted priors, our full model consistently outperforms the baseline. These results highlight the critical role of persistent memory **P** as a reliable fallback when contextual information becomes unreliable, demonstrating that the proposed dual-memory framework is inherently robust to geospatial visual prior corruption.

Table 4: Robustness under corrupted geospatial visual priors on `navhard-two-stage`. Left: average attention weights measured in the full dual-memory model (**P** + **C**). Right: EPDMS for the baseline, the model with contextual memory only (**C** only), and the full model (**P** + **C**). Degradation relative to the uncorrupted condition is shown in parentheses.

	Attention		**EPDMS**		
Corruption	**to P**	**to C**	**Baseline**	**C only**	**P + C**
None	0.29	**0.71**	44.9	47.9	**48.6**
Visual	**0.97**	0.03	–	38.5 (-19.6%)	**45.3 (-6.8%)**
Spatial	**0.65**	0.35	–	45.2 (-5.6%)	**46.6 (-4.1%)**
Both	**0.99**	0.01	–	39.0 (-18.6%)	**45.4 (-6.6%)**

4.4 Ablation Studies

Finally, we validate all design details of our proposed approach through extensive ablation studies, as summarized in Table 5. All ablations use LTF as the baseline and are evaluated on the validation split.

Embedding backbone. All evaluated backbones for encoding geospatial visual priors improve over the baseline, confirming the value of priors regardless of the encoder. Among the evaluated models, SigLIP2 [55] outperforms both DINOv2 [44] and SegFormer [63], suggesting that its richer semantic representations are better suited for encoding street-level visual priors.

Retrieval strategy. The retrieval strategy for geospatial visual priors plays a crucial role. A naive proximity-based strategy [23] that retrieves the N nearest prior nodes yields only a marginal gain over the baseline, suggesting that

Table 5: Ablation studies on the validation split. For reference, baseline (LTF) scores 81.7 EPDMS.

Description	Configuration	EPDMS ↑
Embedding backbone	DINOv2	83.4
	SegFormer	82.9
	SigLIP2	**84.4**
Retrieval strategy	Proximity-based	82.1
	Intention-guided	**84.4**
Memory components	$\mathbf{P}$ only	82.2
	$\mathbf{C}$ only	83.9
	$\mathbf{P} + \mathbf{C}$	**84.4**
Contextual memory components	$\mathbf{C} = \psi(\mathbf{X})$	82.7
	$\mathbf{C} = \phi(\mathbf{V})$	84.0
	$\mathbf{C} = \phi(\mathbf{V}) + \psi(\mathbf{X})$	**84.4**
Memory gate components	$\mathbf{G} = \sigma\Big(f_g\Big([\mathbf{S}; \tilde{\mathbf{S}}]\Big)\Big)$	83.0
	$\mathbf{G} = \sigma\Big(f_g\Big([\mathbf{S}; \tilde{\mathbf{S}}; d]\Big)\Big)$	83.4
	$\mathbf{G} = \sigma\Big(f_g\Big([\mathbf{S}; \tilde{\mathbf{S}}; c]\Big)\Big)$	83.4
	$\mathbf{G} = \sigma\Big(f_g\Big([\mathbf{S}; \tilde{\mathbf{S}}; d; c]\Big)\Big)$	**84.4**

spatially close but maneuver-irrelevant priors provide little useful signal. In contrast, our intention-guided retrieval selects priors conditioned on the planned maneuver, yielding a substantially larger improvement.

Memory components. While contextual memory $\mathbf{C}$ alone already provides a clear gain over the baseline, persistent memory $\mathbf{P}$ alone yields only a marginal improvement, suggesting that adding learnable parameters alone is not sufficient without meaningful contextual input. Combining both yields the best performance, supporting the dual-memory formulation that benefits not only robustness, as discussed in Section 4.3, but also nominal performance.

Contextual memory components. Spatial encodings $\psi(\mathbf{X})$ provide geometric guidance via the centerline of the intention-guided target lane, yielding a modest gain of 1.0 EPDMS over the baseline. Visual encodings $\phi(\mathbf{V})$ contribute more substantially with a gain of 2.3 EPDMS, offering rich appearance details of the upcoming path. Combining both yields the largest gain of 2.7 EPDMS, demonstrating that structural geometry and semantic appearance are complementary.

Memory gate components. Incorporating similarity cues, namely cosine similarity c and normalized Euclidean distance d, individually improves EPDMS, and using both cues together achieves the best performance. This confirms that the two metrics capture complementary aspects of compatibility between the current state and the retrieved memory.

4.5 Runtime Analysis

We analyze the inference overhead introduced by PriorEye on top of the GTRS-Dense baseline as summarized in Tab. 6. All measurements use a single sample (batch size 1) on an NVIDIA RTX PRO 6000, with CPU-side feature computation on an AMD Ryzen Threadripper 7960X. The memory augmentation module adds only 5.6 ms of GPU latency, and the CPU-side memory retrieval adds 3.7 ms to feature computation. The total inference time remains 67.2 ms, comfortably within a 10 Hz runtime budget. Since the SigLIP2 encoding of street-level imagery is performed entirely offline during memory-bank construction, it incurs no overhead at inference time.

Table 6: Inference runtime of PriorEye on GTRS-Dense. Values are averaged over 495 GPU inference runs (after 5 warm-up runs) and 200 CPU feature-computation runs.

	CPU feat. (ms)	GPU (ms)	Total (ms)
GTRS-Dense	15.5	42.4	57.9
+ PriorEye	19.2	48.0	67.2
Overhead	+3.7	+5.6	+9.3

5 Conclusion

We introduced geospatial visual priors to equip E2E driving policies with visual-spatial foresight beyond the immediate sensor view. Our model-agnostic memory augmentation module, featuring dual memory and adaptive gating, consistently improves performance on the NAVSIM-v2 benchmark. In addition, the proposed method provides robustness under sensor degradation, while ensuring resilience when retrieved priors become unreliable.

Limitations. Our current design has two main limitations that suggest promising future directions. First, prior retrieval relies solely on ego-vehicle location. Incorporating appearance-based matching could enable more robust retrieval, allowing the system to recall routes using both spatial and visual cues. Second, the memory bank is constructed from static street-level imagery. With repeated driving logs over the same area, the framework could be extended with memory update and forgetting mechanisms, enabling the system to continuously adapt to long-term environmental changes.

Acknowledgements

The work was supported by the EPSRC Programme Grant "From Sensing to Collaboration" (EP/V000748/1).

References

1. Behrouz, A., Zhong, P., Mirrokni, V.: Titans: Learning to memorize at test time. In: NeurIPS (2025)
2. Burnett, K., Yoon, D.J., Wu, Y., Li, A.Z., Zhang, H., Lu, S., Qian, J., Tseng, W.K., Lambert, A., Leung, K.Y., Schoellig, A.P., Barfoot, T.D.: Boreas: A multi-season autonomous driving dataset. The International Journal of Robotics Research **42**(1-2) (2023)
3. Cao, W., Hallgarten, M., Li, T., Dauner, D., Gu, X., Wang, C., Miron, Y., Aiello, M., Li, H., Gilitschenski, I., Ivanovic, B., Pavone, M., Geiger, A., Chitta, K.: Pseudo-simulation for autonomous driving. In: CoRL (2025)
4. Casas, S., Sadat, A., Urtasun, R.: MP3: A unified model to map, perceive, predict and plan. In: CVPR (2021)
5. Ceccarelli, A., Secci, F.: RGB cameras failures and their effects in autonomous driving applications. IEEE Trans. Dependable Secur. Comput. **20**(4) (2023)
6. Chen, L., Wu, P., Chitta, K., Jaeger, B., Geiger, A., Li, H.: End-to-end autonomous driving: Challenges and frontiers. IEEE TPAMI **46**(12) (2024)
7. Chen, Q., Zhu, X., Ling, Z., Wei, S., Jiang, H., Inkpen, D.: Enhanced LSTM for natural language inference. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers (2017)
8. Chitta, K., Prakash, A., Jaeger, B., Yu, Z., Renz, K., Geiger, A.: Transfuser: Imitation with transformer-based sensor fusion for autonomous driving. IEEE TPAMI **45**(11) (2023)
9. Conneau, A., Kiela, D., Schwenk, H., Barrault, L., Bordes, A.: Supervised learning of universal sentence representations from natural language inference data. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017 (2017)
10. Dauner, D., Hallgarten, M., Li, T., Weng, X., Huang, Z., Yang, Z., Li, H., Gilitschenski, I., Ivanovic, B., Pavone, M., Geiger, A., Chitta, K.: NAVSIM: data-driven non-reactive autonomous vehicle simulation and benchmarking. In: NeurIPS (2024)
11. Ding, K., Chen, B., Su, Y., Gao, H., Jin, B., Sima, C., Li, X., Zhang, W., Barsch, P., Li, H., Zhao, H.: Hint-ad: Holistically aligned interpretability in end-to-end autonomous driving. In: CoRL (2024)
12. Dong, X., Fu, Y., Diao, S., Byeon, W., Chen, Z., Mahabaleshwarkar, A.S., Liu, S., Keirsbilck, M.V., Chen, M., Suhara, Y., Lin, Y.C., Kautz, J., Molchanov, P.: Hymba: A hybrid-head architecture for small language models. In: ICLR (2025)
13. Epstein, R.A., Patai, E.Z., Julian, J.B., Spiers, H.J.: The cognitive map in humans: Spatial navigation and beyond. Nature Neuroscience **20**(11) (2017)
14. Gao, S., Yang, J., Chen, L., Chitta, K., Qiu, Y., Geiger, A., Zhang, J., Li, H.: Vista: A generalizable driving world model with high fidelity and versatile controllability. In: NeurIPS (2024)
15. Gers, F.A., Schmidhuber, J., Cummins, F.A.: Learning to forget: Continual prediction with LSTM. Neural Comput. **12**(10) (2000)
16. Hassan, M., Stapf, S., Rahimi, A., Rezende, P.M.B., Haghighi, Y., Brüggemann, D., Katircioglu, I., Zhang, L., Chen, X., Saha, S., Cannici, M., Aljalbout, E., Ye, B., Wang, X., Davtyan, A., Salzmann, M., Scaramuzza, D., Pollefeys, M., Favaro, P., Alahi, A.: GEM: A generalizable ego-vision multimodal world model for fine-grained ego-motion, object dynamics, and scene composition control. In: CVPR (2025)

17. Hoffer, E., Ailon, N.: Deep metric learning using triplet network. In: Similarity-Based Pattern Recognition - Third International Workshop, SIMBAD 2015, Copenhagen, Denmark, October 12-14, 2015, Proceedings (2015)
18. 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: ECCV (2022)
19. Hu, Y., Yang, J., Chen, L., Li, K., Sima, C., Zhu, X., Chai, S., Du, S., Lin, T., Wang, W., Lu, L., Jia, X., Liu, Q., Dai, J., Qiao, Y., Li, H.: Planning-oriented autonomous driving. In: CVPR (2023)
20. Hwang, J., Xu, R., Lin, H., Hung, W., Ji, J., Choi, K., Huang, D., He, T., Covington, P., Sapp, B., Zhou, Y., Guo, J., Anguelov, D., Tan, M.: EMMA: end-to-end multimodal model for autonomous driving. TMLR (2025)
21. Jia, X., Yang, Z., Li, Q., Zhang, Z., Yan, J.: Bench2drive: Towards multi-ability benchmarking of closed-loop end-to-end autonomous driving. In: NeurIPS (2024)
22. Jia, X., You, J., Zhang, Z., Yan, J.: Drivetransformer: Unified transformer for scalable end-to-end autonomous driving. In: ICLR (2025)
23. Jia, X., Zhang, C., Jiang, Y., Wong, S., Zhang, Z., Chen, C., Zhang, S., Zhou, X., Yang, X., Yan, J., Jiang, Y.: Spatial retrieval augmented autonomous driving. arXiv preprint arXiv:2512.06865 (2025)
24. Jiang, A., Gao, Y., Sun, Z., Wang, Y., Wang, J., Chai, J., Cao, Q., Heng, Y., Jiang, H., Zhang, Z., Guo, X., Sun, H., Zhao, H.: Diffvla: Vision-language guided diffusion planning for autonomous driving. arXiv preprint arXiv:2505.19381 (2025)
25. Jiang, B., Chen, S., Gao, H., Liao, B., Zhang, Q., Liu, W., Wang, X.: VADv2: End-to-end autonomous driving via probabilistic planning. In: ICLR (2026)
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: ICCV (2023)
27. Jiang, S., Huang, Z., Qian, K., Luo, Z., Zhu, T., Zhong, Y., Tang, Y., Kong, M., Wang, Y., Jiao, S., Ye, H., Sheng, Z., Zhao, X., Wen, T., Fu, Z., Chen, S., Jiang, K., Yang, D., Choi, S., Sun, L.: A survey on vision-language-action models for autonomous driving. arXiv preprint arXiv:2506.24044 (2025)
28. Jiang, Z., Zhu, Z., Li, P., Gao, H., Yuan, T., Shi, Y., Zhao, H., Zhao, H.: P-mapnet: Far-seeing map generator enhanced by both sdmap and hdmap priors. IEEE Robotics Autom. Lett. **9**(10) (2024)
29. Karnchanachari, N., Geromichalos, D., Tan, K.S., Li, N., Eriksen, C., Yaghoubi, S., Mehdipour, N., Bernasconi, G., Fong, W.K., Guo, Y., Caesar, H.: Towards learning-based planning: The nuplan benchmark for real-world autonomous driving. In: ICRA (2024)
30. Kirby, E., Boulch, A., Xu, Y., Yin, Y., Puy, G., Zablocki, É., Bursuc, A., Gidaris, S., Marlet, R., Bartoccioni, F., Cao, A., Samet, N., Vu, T., Cord, M.: Driving on registers. arXiv preprint arXiv:2601.05083 (2026)
31. Li, K., Li, Z., Lan, S., Xie, Y., Zhang, Z., Liu, J., Wu, Z., Yu, Z., Álvarez, J.M.: Hydra-mdp++: Advancing end-to-end driving via expert-guided hydra-distillation. arXiv preprint arXiv:2503.12820 (2025)
32. 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: ECCV (2024)
33. Li, Y., Wang, Y., Liu, Y., He, J., Fan, L., Zhang, Z.: End-to-end driving with online trajectory evaluation via BEV world model. arXiv preprint arXiv:2504.01941 (2025)

34. Li, Z., Li, K., Wang, S., Lan, S., Yu, Z., Ji, Y., Li, Z., Zhu, Z., Kautz, J., Wu, Z., Jiang, Y., Álvarez, J.M.: Hydra-mdp: End-to-end multimodal planning with multi-target hydra-distillation. arXiv preprint arXiv:2406.06978 (2024)
35. Li, Z., Yao, W., Wang, Z., Sun, X., Chen, J., Chang, N., Shen, M., Song, J., Wu, Z., Lan, S., Álvarez, J.M.: ZTRS: zero-imitation end-to-end autonomous driving with trajectory scoring. arXiv preprint arXiv:2510.24108 (2025)
36. Li, Z., Yao, W., Wang, Z., Sun, X., Chen, J., Chang, N., Shen, M., Wu, Z., Lan, S., Álvarez, J.M.: Generalized trajectory scoring for end-to-end multimodal planning. arXiv preprint arXiv:2506.06664 (2025)
37. Liao, B., Chen, S., Yin, H., Jiang, B., Wang, C., Yan, S., Zhang, X., Li, X., Zhang, Y., Zhang, Q., Wang, X.: Diffusiondrive: Truncated diffusion model for end-to-end autonomous driving. In: CVPR (2025)
38. Liu, L., Jia, C., Yu, G., Song, Z., Li, J., Jia, F., Wu, P., Hao, X., Luo, Y.: Guideflow: Constraint-guided flow matching for planning in end-to-end autonomous driving. arXiv preprint arXiv:2511.18729 (2025)
39. Luo, Z., Qian, K., Wang, J., Luo, Y., Miao, J., Fu, Z., Wang, Y., Jiang, S., Huang, Z., Hu, Y., Yang, Y., Ye, H., Yang, M., Dong, X., Jiang, K., Yang, D.: Mtrdrive: Memory-tool synergistic reasoning for robust autonomous driving in corner cases. arXiv preprint arXiv:2509.20843 (2025)
40. Maguire, E.A., Gadian, D.G., Johnsrude, I.S., Good, C.D., Ashburner, J., Frackowiak, R.S.J., Frith, C.D.: Navigation-related structural change in the hippocampi of taxi drivers. Proceedings of the National Academy of Sciences **97**(8) (2000)
41. Marcu, A., Chen, L., Hünermann, J., Karnsund, A., Hanotte, B., Chidananda, P., Nair, S., Badrinarayanan, V., Kendall, A., Shotton, J., Arani, E., Sinavski, O.: Lingoqa: Visual question answering for autonomous driving. In: ECCV (2024)
42. Mazumder, K., Flohr, F.B.: Satmap: Revisiting satellite maps as prior for online HD map construction. arXiv preprint arXiv:2601.10512 (2026)
43. NVIDIA: Alpamayo-r1: Bridging reasoning and action prediction for generalizable autonomous driving in the long tail. arXiv preprint arXiv:2511.00088 (2025)
44. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P., Li, S., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jégou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: Dinov2: Learning robust visual features without supervision. TMLR (2024)
45. Renz, K., Chen, L., Arani, E., Sinavski, O.: Simlingo: Vision-only closed-loop autonomous driving with language-action alignment. In: CVPR (2025)
46. Rosenbaum, R.S., Ziegler, M., Winocur, G., Grady, C.L., Moscovitch, M.: "I have often walked down this street before": fMRI Studies on the hippocampus and other structures during mental navigation of an old environment. Hippocampus **14**(7) (2004)
47. Shao, H., Hu, Y., Wang, L., Song, G., Waslander, S.L., Liu, Y., Li, H.: Lmdrive: Closed-loop end-to-end driving with large language models. In: CVPR (2024)
48. Song, Z., Jia, C., Liu, L., Pan, H., Zhang, Y., Wang, J., Zhang, X., Xu, S., Yang, L., Luo, Y.: Don't shake the wheel: Momentum-aware planning in end-to-end autonomous driving. In: CVPR (2025)
49. Srivastava, R.K., Greff, K., Schmidhuber, J.: Training very deep networks. In: NeurIPS (2015)
50. Sukhbaatar, S., Grave, E., Lample, G., Jégou, H., Joulin, A.: Augmenting self-attention with persistent memory. arXiv preprint arXiv:1907.01470 (2019)
51. Sun, W., Lin, X., Shi, Y., Zhang, C., Wu, H., Zheng, S.: Sparsedrive: End-to-end autonomous driving via sparse scene representation. In: ICRA (2025)

52. Sun, Y., Ochiai, H., Wu, Z., Lin, S., Kanai, R.: Associative transformer. In: CVPR (2025)
53. Tampuu, A., Matiisen, T., Semikin, M., Fishman, D., Naveed, M.: A survey of end-to-end driving: Architectures and training methods. IEEE Trans. Neural Networks Learn. Syst. **33**(4) (2022)
54. Tian, H., Li, T., Liu, H., Yang, J., Qiu, Y., Li, G., Wang, J., Gao, Y., Zhang, Z., Wang, L., Ye, H., Tan, T., Chen, L., Li, H.: Simscale: Learning to drive via real-world simulation at scale. arXiv preprint arXiv:2511.23369 (2025)
55. Tschannen, M., Gritsenko, A.A., Wang, X., Naeem, M.F., Alabdulmohsin, I., Parthasarathy, N., Evans, T., Beyer, L., Xia, Y., Mustafa, B., Hénaff, O.J., Harmsen, J., Steiner, A., Zhai, X.: Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. arXiv preprint arXiv:2502.14786 (2025)
56. Wang, W., Dong, L., Cheng, H., Liu, X., Yan, X., Gao, J., Wei, F.: Augmenting language models with long-term memory. In: NeurIPS (2023)
57. Wang, X., Zhu, Z., Huang, G., Chen, X., Zhu, J., Lu, J.: Drivedreamer: Towards real-world-drive world models for autonomous driving. In: ECCV (2024)
58. Wang, Y., Liu, Q., Jiang, Z., Wang, T., Jiao, J., Chu, H., Gao, B., Chen, H.: RAD: retrieval-augmented decision-making of meta-actions with vision-language models in autonomous driving. In: CVPR (2025)
59. 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. In: NeurIPS (2022)
60. Wu, Y., Liang, S., Zhang, C., Wang, Y., Zhang, Y., Guo, H., Tang, R., Liu, Y.: From human memory to AI memory: A survey on memory mechanisms in the era of llms. arXiv preprint arXiv:2504.15965 (2025)
61. Wu, Y., Rabe, M.N., Hutchins, D., Szegedy, C.: Memorizing transformers. In: ICLR (2022)
62. Xiao, G., Tian, Y., Chen, B., Han, S., Lewis, M.: Efficient streaming language models with attention sinks. In: ICLR (2024)
63. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Álvarez, J.M., Luo, P.: Segformer: Simple and efficient design for semantic segmentation with transformers. In: NeurIPS (2021)
64. Xing, Z., Zhang, X., Hu, Y., Jiang, B., He, T., Zhang, Q., Long, X., Yin, W.: Goalflow: Goal-driven flow matching for multimodal trajectories generation in end-to-end autonomous driving. In: CVPR (2025)
65. Xiong, X., Liu, Y., Yuan, T., Wang, Y., Wang, Y., Zhao, H.: Neural map prior for autonomous driving. In: CVPR (2023)
66. Xu, Z., Zhang, Y., Xie, E., Zhao, Z., Guo, Y., Wong, K.K., Li, Z., Zhao, H.: Drivegpt4: Interpretable end-to-end autonomous driving via large language model. IEEE Robotics Autom. Lett. **9**(10) (2024)
67. Yao, W., Li, Z., Lan, S., Wang, Z., Sun, X., Álvarez, J.M., Wu, Z.: Drivesuprim: Towards precise trajectory selection for end-to-end planning. arXiv preprint arXiv:2506.06659 (2025)
68. Yuan, J., Sun, S., Omeiza, D., Zhao, B., Newman, P., Kunze, L., Gadd, M.: Rag-driver: Generalisable driving explanations with retrieval-augmented in-context multi-modal large language model learning. In: RSS (2024)
69. Zeng, S., Chang, X., Xie, M., Liu, X., Bai, Y., Pan, Z., Xu, M., Wei, X.: FutureSightDrive: Thinking visually with spatio-temporal CoT for autonomous driving. In: NeurIPS (2025)

70. Zheng, W., Song, R., Guo, X., Zhang, C., Chen, L.: Genad: Generative end-to-end autonomous driving. In: ECCV (2024)
71. Zhong, S., Xu, M., Ao, T., Shi, G.: Understanding transformer from the perspective of associative memory. arXiv preprint arXiv:2505.19488 (2025)