REDRIVE: SHAPING REPRESENTATIONS WITH WORLD MODELING FOR END-TO-END DRIVING

Anonymous authors
Paper under double-blind review

ABSTRACT

Driving policies require capabilities of scene understanding and future evolution prediction. To achieve this goal, current end-to-end models typically construct complex perception-planning pipelines or introduce world models that explicitly predict future states, resulting in a complex system architecture. Inspired by the transferability of general-purpose visual representations, we argue that combining sufficiently strong visual representations with representation world modeling can support effective planning without relying on complex inference-time auxiliary modules. Based on this insight, we present ReDrive, an end-to-end driving framework that strengthens planning-oriented visual features via future representation prediction. To achieve this, ReDrive adopts a three-stage training pipeline consisting of driving video pretraining, joint world-modeling and planning training, and planner adaptation. This yields a strong planning-oriented representation and a high-performance planner, while requiring neither auxiliary perception modules nor future prediction at inference time. Experiments on NAVSIM demonstrate strong performance, achieving 91.0 PDMS on NAVSIM v1 and 90.8 EPDMS on NAVSIM v2. These results show that shaping representations with world modeling is sufficient to enable high-performance end-to-end planning while retaining a simple encoder-planner inference pipeline.

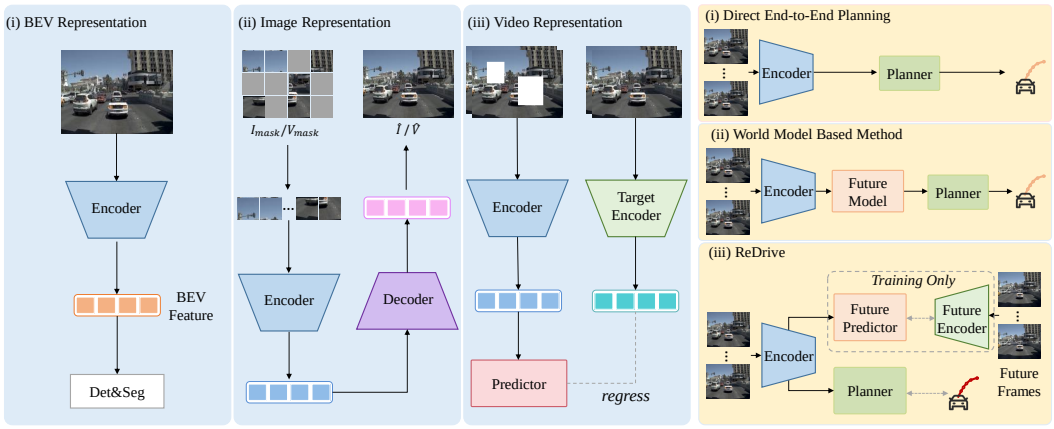
1 INTRODUCTION

End-to-end autonomous driving (Hu et al., 2023; Jiang et al., 2023) requires planning-oriented visual representations that capture scene semantics while reflecting future scene evolution. Scene semantics (Li et al., 2024b; Wang et al., 2026b) provide the essential context for trajectory planning. Future scene evolution (Xia et al., 2026a; Wang et al., 2026a) provides predictive information for guiding future driving behaviors.

Prior work has explored diverse strategies for learning visual representations for planning. BEV-based methods (Li et al., 2024b; Hu et al., 2023; Jiang et al., 2023) leverage task-specific supervision to learn spatially structured BEV features. Inspired by the transferability of self-supervised visual representations trained from large-scale visual data without task-specific annotations (Oquab et al., 2023; Siméoni et al., 2025; Bardes et al., 2024; Assran et al., 2025), recent works transfer these general-purpose representations to driving policies to provide strong visual priors for downstream planning (Wang et al., 2026b; Xing et al., 2026a;b). However, when used alone without auxiliary perception tasks, these general-purpose representations still underperform task-specific BEV representations in downstream planning.

World models provide another paradigm by explicitly modeling future scene evolution. Generative world models (Xia et al., 2026a; Zhang et al., 2025; Li et al., 2026a) predict future observations in pixel space, capturing scene evolution through visual generation. However, modeling fine-grained future geometric details may distract from the scene dynamics most relevant to planning. Recent approaches model dynamics in latent space by predicting future representations, avoiding explicit pixel generation while focusing on higher-level scene evolution (Lu et al., 2026; Zhong et al., 2026). These methods achieve planning performance comparable to perception-based approaches without auxiliary perception tasks. However, inference-time reliance on future representations introduces additional computational cost and pipeline complexity.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



(a) Existing visual representations. (b) Comparison to existing methods.

Figure 1: Overview of existing visual representation learning paradigms and our proposed ReDrive for end-to-end autonomous driving.

Building on these observations, we argue that strong pretrained video representations combined with representation world modeling can support high-performance driving without complex auxiliary modules or inference-time future prediction. We propose ReDrive, an end-to-end driving framework that strengthens planning-oriented visual representations through trajectory-conditioned future prediction. ReDrive adopts a two-branch architecture that predicts ego trajectories and future representations conditioned on the ego trajectories. Jointly optimizing the two branches provides the visual encoder with additional supervision on planning-relevant scene evolution beyond direct trajectory supervision. At inference, ReDrive uses only the learned visual encoder and trajectory generator, with the future predictor completely removed.

ReDrive follows a three-stage training procedure consisting of drive video pretraining, joint training, and planner adaptation. We first initialize the video encoder with a general-purpose video pretrained model (Assran et al., 2025) and further pretrain it on large-scale driving videos through masked visual modeling in latent feature space. We then jointly train the encoder with a trajectory generator and a future representation predictor. The trajectory generator learns to predict the ground-truth ego trajectory, while the future predictor uses the expert trajectory as a condition to predict the corresponding future representation. The shared encoder therefore receives supervision from both tasks, encouraging it to capture information about future scene evolution beyond motion supervision alone. Finally, we freeze the visual encoder and future predictor while adapting the planner with its on-policy rollout. We obtain the corresponding future representation of the rollout trajectory using the frozen future predictor. Therefore, we directly optimize the planner under supervision of the future representation modeling.

We evaluate ReDrive on both NAVSIM v1 (Dauner et al., 2024) and NAVSIM v2 (Cao et al., 2025). After training, the learned planning-oriented visual representations exhibit stronger planning capability, yielding a 3.5-point PDMS improvement over representations learned without our complete training pipeline. ReDrive demonstrates strong trajectory planning performance across both benchmarks, achieving a PDMS of 91.0 on NAVSIM v1 and an EPDMS of 90.8 on NAVSIM v2.

Our contributions are summarized as follows:

- We introduce ReDrive, a simple yet effective driving policy built upon a strong planning-oriented visual representation shaped with world modeling without the requirement of inference-time auxiliary architecture designs, e.g., perception heads and future predictors.
- We develop a three-stage training strategy that progressively incorporates future prediction supervision into trajectory learning and further optimizes the planner on its own rollouts under future representation modeling supervision.
- We evaluate ReDrive on the NAVSIM v1 and NAVSIM v2 benchmarks, demonstrating strong planning performance across both benchmarks.

2 Related Work

2.1 End-to-end Autonomous Driving

End-to-end autonomous driving aims to directly optimize driving decisions from sensor observations. Early approaches (Hu et al., 2022; 2023; Jiang et al., 2023) build scene representations and jointly optimize perception, prediction, and planning to support end-to-end trajectory generation. Subsequent methods (Liao et al., 2025; Zheng et al., 2026; Zou et al., 2025; Xing et al., 2025) adopt generative planners that directly model the trajectory distribution. Recent methods incorporate reinforcement learning (Gao et al., 2025; 2026; Jiang et al., 2025) or vision language model (VLM) prior (Jiang et al., 2024; Song et al., 2026; Li et al., 2026c) to enhance planning performance.

2.2 Driving Representation Learning

Classical autonomous driving systems commonly build bird's-eye-view (BEV) representations from multi-view camera inputs to provide spatially structured features for downstream perception and planning (Li et al., 2024b; Hu et al., 2023; Jiang et al., 2023). Recent advances in self-supervised visual pretraining offer a different paradigm. Several works (Wang et al., 2026a; Kirby et al., 2026) have explored general-purpose image representations, such as DINOv2 features (Oquab et al., 2023), as visual backbones for end-to-end driving. More recently, a series of works (Xing et al., 2026b; Wang et al., 2026b; Xing et al., 2026a) further incorporate pretrained video representations (Bardes et al., 2024; Assran et al., 2025) into driving scene understanding for better temporal semantic extraction. Another direction predicts compact action-oriented representations for driving decisions (Yang et al., 2026a).

2.3 Driving World Action Models

World action models leverage explicit scene evolution prediction to enhance trajectory generation. Existing methods typically forecast future states in BEV or latent representations and use them for trajectory generation, candidate selection, or online evaluation (Li et al., 2025c; Zheng et al., 2025; Hong et al., 2026; Zhang et al., 2026). Other approaches connect world representations with planning policies through future visual generation (Zhang et al., 2025; Xia et al., 2026a;b). These methods mainly use predicted futures or world representations as additional information for planning, making trajectory decisions directly dependent on future prediction. In contrast, we use future dynamics to improve the video representation used for planning, enabling direct trajectory planning with a stronger representation.

3 Method

3.1 Overview

Our goal is to embed predictive information into visual representations for trajectory planning, removing the need for complex future reasoning during inference. To this end, ReDrive incorporates future representation prediction into the training process through three stages: Self-Supervised Pretraining, Joint Training, and Planner Adaptation, as illustrated in Fig. 2.

3.2 Driving-Domain Pretraining

We adopt a video-pretrained backbone from V-JEPA2 (Assran et al., 2025) to capture temporal information across video frames and adapt it to the driving domain through self-supervised pretraining on driving videos. V-JEPA2 (Assran et al., 2025) masks a set of spatiotemporal tokens $\mathcal{M}$ and predicts their latent representations from the visible context. The context encoder extracts representations from visible tokens, while the momentum target encoder provides target representations. The predictor predicts the masked target representations from the visible context. The pretraining objective is defined as

$$\mathcal{L}_{\text{ssl}} = \frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} \left\| \hat{z}_i - z_i^{\text{tgt}} \right\|_1 , \tag{1}$$

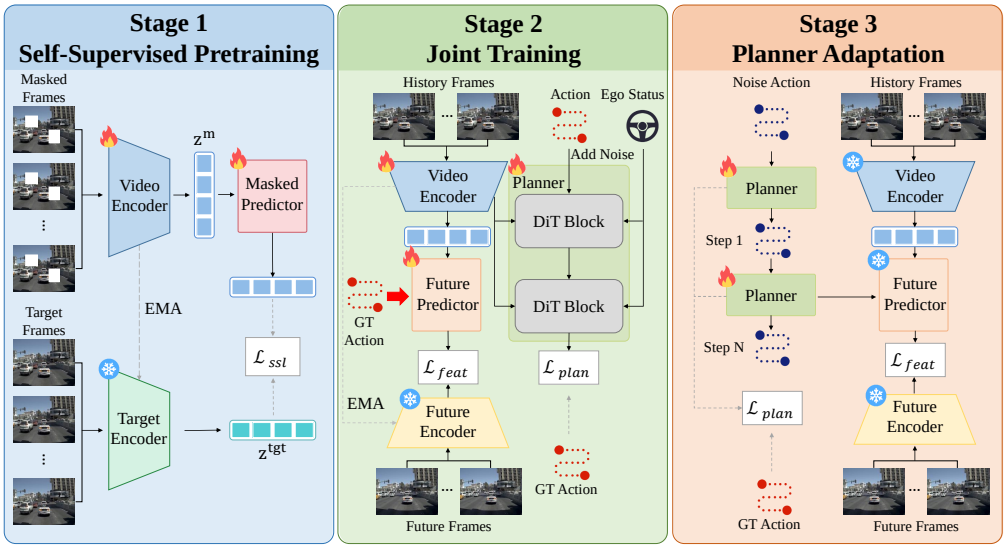


Figure 2: Overview of ReDrive and its three-stage training procedure. Self-Supervised Pretraining adapts the video encoder to the driving domain. Joint Training jointly learns trajectory generation and future representation prediction from the shared history representation. Planner Adaptation freezes the encoder and future predictor and uses planner-generated trajectories as the motion condition for predictive supervision.

where $\hat{z}_i$ and z_i^{tgt} denote the predicted and target representations at masked position i, respectively, and the ℓ_1 distance is averaged over the feature dimension. By optimizing this masked prediction objective on driving videos, the pretrained video representation is adapted to the driving domain. The resulting context encoder is then used to initialize the video encoder for subsequent future feature prediction and trajectory planning.

3.3 Joint Training

We jointly optimize future representation prediction and trajectory planning through a shared video representation. The future prediction branch learns future representations conditioned on the ground-truth ego trajectory, while the planning branch directly generates the ego trajectory from the history representation. This joint training introduces future-aware supervision into the video representation while keeping trajectory generation decoupled from future prediction.

3.3.1 Future Representation Prediction

We divide the video into a history clip $X_{1:T}$ and a non-overlapping future clip $X_{T+1:2T}$. The history clip is encoded by E_θ and the future clip is processed by the EMA target encoder to provide the prediction target. We condition future feature prediction on the ground-truth ego trajectory corresponding to the future clip $X_{T+1:2T}$. The trajectory is encoded into action tokens A, which are incorporated into the predictor through cross-attention with the history representation Z_h to predict the future representation:

$$Z_h = E_\theta(X_{1:T}), \qquad Z_f^{\text{tgt}} = E^{\text{tgt}}(X_{T+1:2T}), \qquad \hat{Z}_f = P_\phi(Z_h, A). \tag{2}$$

We regress the predicted representation $\hat{Z}_f$ toward the stop-gradient target Z_f^{tgt} after feature normalization:

$$\mathcal{L}_{\text{feat}} = \frac{1}{N} \left\| \text{Norm}(\hat{Z}_f) - \text{Norm}(Z_f^{\text{tgt}}) \right\|_1, \tag{3}$$

where N denotes the number of elements in the future representation.

The objective updates both the predictor P_ϕ and the video encoder E_θ. Future representation prediction encourages the encoder to capture scene evolution conditioned on ego motion. This enriches the visual representation with future dynamics.

3.3.2 TRAJECTORY GENERATION

The planning branch adopts an Action DiT (Peebles & Xie, 2023) to generate trajectories directly from the history video representation Z_h. During training, we construct a noisy trajectory from the ground-truth trajectory τ and Gaussian noise $\epsilon \sim \mathcal{N}(0, I)$ at noise level t

$$\tau_t = (1-t)\tau + t\epsilon. \tag{4}$$

The noisy trajectory τ_t, noise level t, and auxiliary driving condition c are embedded into action tokens. The Action DiT incorporates the history representation Z_h through cross-attention as the visual condition and predicts the velocity $\hat{v}_t$. Following the linear noising path in equation 4, we obtain the target velocity v_t and optimize the Action DiT with the flow-matching objective

$$\hat{v}_t = G_\psi(\tau_t, t, c, Z_h), \qquad v_t = \epsilon - \tau, \qquad \mathcal{L}_{\text{plan}} = |\hat{v}_t - v_t|_2^2, \tag{5}$$

where G_ψ predicts the velocity field for trajectory generation.

Joint Optimization. The future prediction and planning branches are jointly optimized through the video encoder E_θ. The overall training objective is

$$\mathcal{L} = \mathcal{L}_{\text{plan}} + \lambda \mathcal{L}_{\text{feat}}, \tag{6}$$

where λ controls the weight of future feature prediction loss. The planning objective updates E_θ and G_ψ, while the future prediction objective updates E_θ and P_ϕ. This decoupling prevents the prediction objective from steering the trajectory policy toward actions that are easier to predict rather than better for planning. The Action DiT is therefore optimized solely by the planning objective, while future prediction improves planning by shaping a stronger video representation.

3.4 PLANNER ADAPTATION

After Stage 2, the trained future predictor provides a fixed mapping from the history representation and an ego trajectory to a predicted future representation. We then freeze the video encoder and future predictor and condition future prediction on planner-generated trajectories. This allows the feature prediction loss to supervise the planner through the fixed predictor, alongside the trajectory supervision.

During this stage, the video encoder E_θ and future predictor P_ϕ are frozen, and only the Action DiT is optimized. Starting from Gaussian noise, the Action DiT performs a multi-step rollout to generate an ego trajectory $\hat{\tau}$. The generated trajectory is encoded into action tokens $\hat{A}$ and fed into the frozen future predictor, replacing the ground-truth trajectory condition in equation 2. The resulting feature prediction loss $\mathcal{L}_{\text{feat}}$ is propagated through the generated trajectory to optimize the planner.

For trajectory supervision, we apply the flow-matching objective at each denoising step and an L1 loss on the final generated trajectory. The overall adaptation objective is

$$\mathcal{L}_{\text{plan}} = \mathcal{L}_{\text{fm}} + \lambda_{\text{traj}} \mathcal{L}_{\text{traj}}, \qquad \mathcal{L}_{\text{adapt}} = \mathcal{L}_{\text{plan}} + \lambda_{\text{feat}} \mathcal{L}_{\text{feat}}. \tag{7}$$

The video encoder and future predictor remain frozen to preserve the future dynamics learned under ground-truth trajectory conditioning. Imperfect planner-generated trajectories could otherwise introduce erroneous supervision into the future predictor and alter the learned correspondence between ego motion and future scene evolution. Consequently, all adaptation objectives update only the Action DiT.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Datasets. We conduct downstream trajectory planning experiments on NAVSIM (Dauner et al., 2024), covering both the v1 and v2 benchmarks, including the challenging NavHard evaluation split of NAVSIM v2. NAVSIM (Dauner et al., 2024) provides approximately 103K training scenes and 12K test scenes for standardized planning evaluation. For driving-domain video pretraining, we use videos from the training splits of nuScenes (Caesar et al., 2020) and nuPlan (Caesar et al., 2021), together with an 80-hour subset of the NVIDIA PhysicalAI-Autonomous-Vehicles Dataset. These

Table 1: Comparison with state-of-the-art methods on NAVSIM v1. The best and second-best results are highlighted in bold and underlined, respectively.

Type	Method	Inputs	NC↑	DAC↑	EP↑	C↑	TTC↑	PDMS↑
Perception-based	Transfuser (Chitta et al., 2022)	C + L	97.7	92.8	79.2	100	92.8	84.0
	VADv2 (Jiang et al., 2026)	Camera	97.2	89.1	91.6	100	76.0	80.9
	UniAD (Hu et al., 2023)	Camera	97.8	91.9	92.9	100	78.8	83.4
	Hydra-MDP (Li et al., 2024a)	C + L	98.4	97.7	85.0	100	94.5	89.9
	Hydra-MDP++ (Li et al., 2025a)	C + L	97.6	96.0	80.4	100	93.1	86.6
	DiffusionDrive (Liao et al., 2025)	C + L	98.2	96.2	82.2	100	94.7	88.1
	GoalFlow (Xing et al., 2025)	C + L	98.4	98.3	85.0	100	94.6	90.3
	DriveDPO (Shang et al., 2026)	C + L	98.5	98.1	84.3	100	94.8	90.0
	DriveSuprim (Yao et al., 2026)	Camera	98.6	98.6	91.3	100	95.5	89.9
Perception-free	LAW (Li et al., 2025b)	C + L	97.4	93.3	78.8	100	91.9	83.8
	World4Drive (Zheng et al., 2025)	C + L	97.4	94.3	79.9	100	92.8	85.1
	Epona (Zhang et al., 2025)	Camera	97.9	95.1	80.4	99.9	93.8	86.2
	Drive-JEPA (Wang et al., 2026b)	Camera	98.7	96.2	82.9	100	95.5	89.0
	DriveLaW (Xia et al., 2026a)	Camera	99.0	97.1	81.3	100	96.7	89.1
	ReWorld (Xia et al., 2026b)	Camera	99.1	98.2	82.0	99.8	97.7	<u>90.4</u>
	DAWN (Lu et al., 2026)	Camera	98.7	95.9	84.3	100	96.0	89.1
	ReDrive (Ours)	Camera	99.1	97.9	84.3	100	97.2	**91.0**

durations refer to the source collection before filtering by dataset split. Only videos from the training splits of nuScenes (Caesar et al., 2020) and nuPlan (Caesar et al., 2021) are used for pretraining. The pretrained video encoder is subsequently used to initialize the predictive adaptation and trajectory planning stages on NAVSIM.

Implementation Details. We initialize the video encoder from the pretrained V-JEPA2 (Assran et al., 2025) model and further pretrain it for 100 epochs on the driving videos using 16-frame clips. We then jointly train the video encoder, future predictor, and Action DiT for 35K steps on the NAVTRAIN split of NAVSIM (Dauner et al., 2024), where four observed frames are used to predict the subsequent four frames. The Planner Adaptation stage further trains the Action DiT for 5K steps with the video encoder and future predictor frozen. We use five denoising steps for trajectory rollout during this stage. The input resolution is 256×512. The future predictor consists of 12 Transformer blocks with 12 attention heads and a hidden dimension of 768. All experiments run on 16 NVIDIA H20 GPUs.

Evaluation Metrics. We evaluate planning performance on both NAVSIM v1 and NAVSIM v2. For NAVSIM v1, we report the Predictive Driver Model Score (PDMS) together with its individual components, including no-at-fault collision (NC), drivable area compliance (DAC), time-to-collision (TTC), ego progress (EP), and comfort. For NAVSIM v2, we report the Extended Predictive Driver Model Score (EPDMS) as the primary aggregate metric following the official evaluation protocol. On NavHard, we report the individual metric components and per-stage scores (S.) for Stage-1 and Stage-2, together with the combined EPDMS.

4.2 Comparison with State-of-the-Art Methods

NAVSIM v1. Tab. 1 compares ReDrive with state-of-the-art methods on NAVSIM v1. Among perception-free approaches, ReDrive achieves the best PDMS of 91.0, outperforming the previous best ReWorld by 0.6 points. Notably, despite relying only on camera inputs and without explicit perception modules, ReDrive surpasses several perception-based methods, including Hydra-MDP (Li et al., 2024a), DiffusionDrive (Liao et al., 2025), GoalFlow (Xing et al., 2025), and DriveDPO (Shang et al., 2026) in terms of PDMS. These results show that strong predictive visual representations can enable a direct planner to achieve competitive performance without relying on explicit perception modules.

NAVSIM v2. Tab. 2 compares ReDrive with state-of-the-art methods on NAVSIM v2. ReDrive achieves the best overall performance and outperforms the second-best SparseDriveV2 by 0.7 points and also exceeds recent world-model and world-action methods such as Latent-WAM (Wang et al., 2026a), DriveFuture (Hong et al., 2026), CoWorld-VLA (Huang et al., 2026), and Dream-

Table 2: Comparison with state-of-the-art methods on NAVSIM v2. The best and second-best results are highlighted in bold and underlined, respectively. NC–EC uniformly report the corrected-evaluator metrics.

Method	NC↑	DAC↑	DDC↑	TLC↑	EP↑	TTC↑	LK↑	HC↑	EC↑	EPDMS↑
DiffusionDrive (Liao et al., 2025)	98.2	95.9	99.4	99.8	87.5	97.3	96.8	98.3	87.7	84.5
DiffusionDriveV2 (Zou et al., 2025)	97.7	96.6	99.2	99.8	88.9	97.2	96.0	97.8	91.0	87.5
WAM-Diff (Xu et al., 2025)	99.0	98.4	99.3	99.9	87.0	98.6	96.2	98.1	78.5	89.7
DreamerAD (Yang et al., 2026b)	98.0	97.2	99.5	99.8	87.8	97.4	97.5	98.3	72.4	87.7
Latent-WAM (Wang et al., 2026a)	98.1	97.3	99.6	99.8	87.7	97.3	97.6	98.1	87.3	89.3
DriveFuture (Hong et al., 2026)	98.8	99.1	99.6	99.9	86.6	98.4	96.4	98.3	74.8	89.9
CoWorld-VLA (Huang et al., 2026)	99.1	97.0	99.6	99.9	87.9	98.5	97.7	98.2	86.2	90.0
SparseDriveV2 (Sun et al., 2026)	98.1	98.1	99.6	99.8	91.1	97.3	96.9	98.2	78.4	<u>90.1</u>
ReDrive (Ours)	99.1	97.8	99.6	99.9	87.6	98.7	98.2	98.3	86.4	**90.8**

Table 3: Comparison with state-of-the-art methods on NAVSIM v2 NavHard. S1 and S2 denote Stage-1 and Stage-2, respectively. S. denotes the per-stage score, and EPDMS reports the combined score. The best and second-best results are highlighted in bold and underlined, respectively.

Method	Stage	NC↑	DAC↑	DDC↑	TLC↑	EP↑	TTC↑	LK↑	HC↑	EC↑	S.↑	EPDMS↑
LTF (Chitta et al., 2022)	S1	97.3	80.2	97.8	99.3	83.4	96.2	92.9	97.8	71.1	61.3	24.4
	S2	79.4	69.0	85.6	98.5	83.8	76.7	47.9	97.0	70.6	39.2	
DiffusionDrive (Liao et al., 2025)	S1	96.8	86.0	98.8	99.3	84.0	95.8	96.7	97.6	79.6	66.7	27.5
	S2	80.1	72.8	84.4	98.4	85.9	76.6	46.4	96.3	72.8	40.5	
GTRS-DP (Li et al., 2025d)	S1	94.7	78.8	96.1	99.5	83.0	94.4	92.0	97.5	72.8	–	23.8
	S2	80.3	74.4	84.9	98.0	81.9	78.8	45.4	96.7	70.1	–	
GuideFlow (Liu et al., 2026)	S1	96.6	80.5	96.3	99.3	82.3	94.9	91.5	97.7	67.8	–	27.1
	S2	87.3	76.7	88.8	99.2	84.3	85.1	49.7	93.1	44.5	–	
ReCogDrive (Li et al., 2026c)	S1	96.4	78.9	98.7	99.8	82.6	95.6	94.4	97.6	74.2	67.7	25.7
	S2	80.2	65.0	82.4	98.7	85.2	76.9	43.8	96.6	71.8	37.6	
SGDrive (Li et al., 2026b)	S1	95.8	87.6	97.8	99.8	84.4	94.7	92.9	97.8	28.9	71.1	25.5
	S2	79.4	65.4	79.1	98.9	88.9	75.3	42.7	96.4	29.6	35.2	
Metis (Li et al., 2026a)	S1	96.6	87.8	99.0	99.3	84.5	95.6	97.8	97.8	77.8	<u>75.8</u>	<u>32.2</u>
	S2	79.6	73.3	84.9	97.8	85.8	76.6	47.7	95.4	75.3	<u>41.7</u>	
ReDrive (Ours)	S1	97.9	93.6	99.4	99.6	84.2	96.9	97.8	97.8	73.3	**82.3**	**34.4**
	S2	82.2	75.8	84.3	98.2	87.3	77.9	47.8	95.9	54.2	**42.3**	

erAD (Yang et al., 2026b). These results further show that predictive visual representations can support strong planning performance on NAVSIM v2 without relying on an explicit future model during inference.

NAVSIM v2 NavHard. Tab. 3 compares ReDrive with state-of-the-art methods on NAVSIM v2 NavHard. ReDrive outperforms Metis (Li et al., 2026a) and DiffusionDrive (Liao et al., 2025) in both per-stage scores and combined EPDMS. Its strengths are most pronounced in Stage-1, where it leads the compared methods in collision avoidance, drivable-area compliance, driving-direction compliance, TTC, and lane keeping. In Stage-2, ReDrive improves collision avoidance and drivable-area compliance over Metis (Li et al., 2026a). This pattern suggests stronger safety and road compliance, with room to improve motion comfort under the more challenging evaluation conditions.

4.3 Ablation Studies

Effect of Visual Pretraining. We compare different pretraining strategies and temporal lengths in Tab. 4. For a fair comparison, DINOv2 (Oquab et al., 2023), MAE (He et al., 2022), and our Stage 1 pretraining use encoders with comparable parameter counts initialized from their respective pretrained checkpoints. All encoders are further pretrained on the same driving data for the same number of epochs and evaluated after the same Stage 2 training. DINOv2 outperforms MAE, sug-

Table 4: Ablation on visual pretraining. We compare image- and video-based pretraining and study the temporal length used for self-supervised pretraining. All results are evaluated after Stage 2 training.

Pretraining	DINOv2	MAE	V-JEPA2			
Frames	–	–	4	8	12	16
PDMS↑	84.4	83.9	89.9	90.4	90.7	**90.8**

Table 5: Ablation on encoder fine-tuning and future prediction in Stage 2, with the encoder pretrained on 4-frame driving videos.

Encoder	Future Predictor	PDMS↑
Frozen	✗	85.7
Trainable	✗	89.2
Trainable	✓	**89.9**

Table 6: Ablation on the three-stage training pipeline. The Encoder and Planner are jointly trained for evaluation, with the Future Predictor introduced in Stage 2 and Planner-only adaptation in Stage 3.

Stage 1	Stage 2	Stage 3	PDMS↑
✗	✗	✗	88.9
✓	✗	✗	89.2
✓	✓	✗	89.9
✓	✓	✓	**90.2**

Table 7: Planning evaluation of frozen visual encoders with newly trained planners. Stage 2 joint training is included for reference. All models use 8-frame pretraining.

Encoder	Planner	PDMS↑
8-frame Pretrained	Reinitialized	86.7
Stage 2 Trained	Reinitialized	90.2
Stage 2 Joint Training (reference)		90.4

gesting that stronger visual representations benefit trajectory planning. Our video-based pretraining further improves over both image-based baselines, demonstrating the additional benefit of temporal representation learning. Increasing the number of frames consistently improves performance, further highlighting the importance of temporal context for planning.

Effect of Encoder Adaptation and Future Prediction. During Stage 2, starting from the same encoder pretrained on 4-frame driving videos, we ablate whether the encoder is frozen and whether the Future Predictor is jointly trained with the planner in Tab. 5. Fine-tuning the encoder with the planning objective improves PDMS from 85.7 to 89.2. Jointly training the Future Predictor further improves PDMS to 89.9, showing that future representation prediction provides complementary supervision beyond the planning objective for learning planning-oriented visual representations.

Effect of Training Stages. We evaluate each stage of the training pipeline in Tab. 6. Starting from the original V-JEPA2 model, Stage 1 pretrains the encoder on 4-frame driving videos, improving PDMS from 88.9 to 89.2. Stage 2 jointly trains the Encoder, Future Predictor, and Planner, improving PDMS from 89.2 to 89.9. Stage 3 further conditions future prediction on planner-generated trajectories, improving PDMS to 90.2. These results demonstrate the complementary contributions of the three stages.

4.4 Experimental Analysis

ReDrive Indeed Improves Representations. We further evaluate whether Stage 2 improves the visual representation itself. With the encoder frozen and a newly initialized planner, the Stage 2 encoder achieves 90.2 PDMS, substantially outperforming the pretrained encoder at 86.7 and approaching the 90.4 of Stage 2 joint training. This shows that the improvement from Stage 2 is largely retained in the visual representation and can be directly utilized for planning. All variants are trained to convergence.

Qualitative Results. Fig. 3 compares the human trajectory, Drive-JEPA (Wang et al., 2026b), and ReDrive in representative driving scenarios using both front-camera and BEV visualizations. Compared with Drive-JEPA, ReDrive produces trajectories that more closely follow the human trajectory while maintaining better consistency with the road geometry and surrounding traffic. The visual comparison provides qualitative evidence that the learned representation supports more accurate and scene-aware planning.

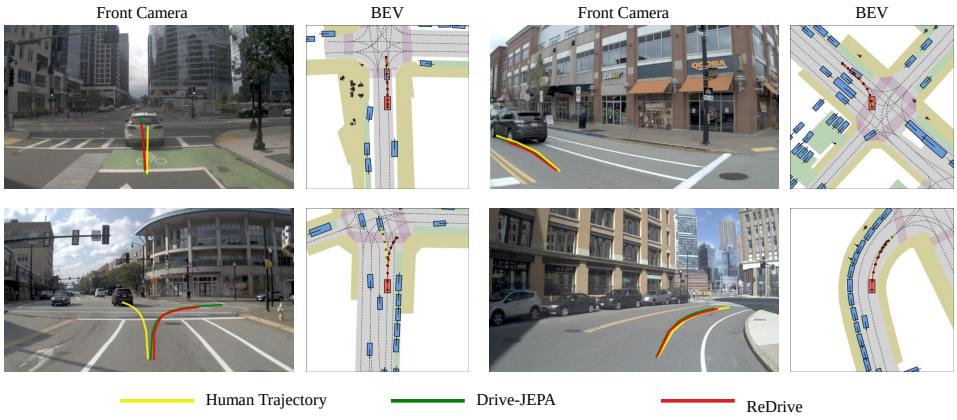


Figure 3: Qualitative comparison of planning results using front-camera and bird's-eye-view (BEV) visualizations. We compare the human trajectory, Drive-JEPA (Wang et al., 2026b), and ReDrive across representative driving scenarios.

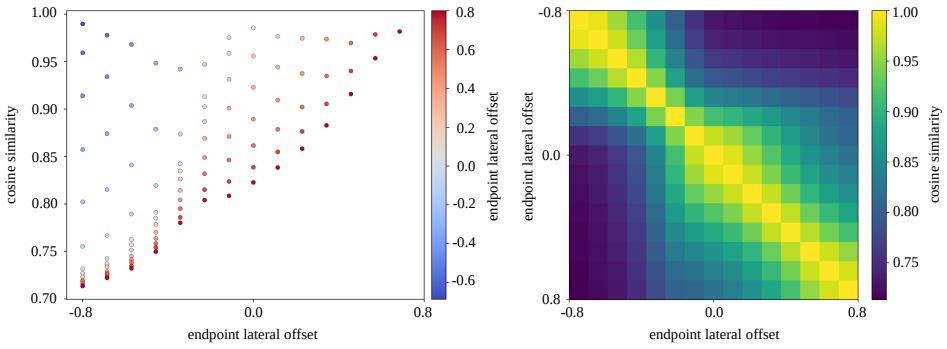


Figure 4: Sensitivity of the Future Predictor to trajectory conditions with the history representation held fixed. (a) Cosine similarities between future representations predicted under different lateral trajectory offsets. The horizontal axis and point color indicate the offsets of the two trajectory conditions being compared. (b) Pairwise cosine similarity matrix for all trajectory conditions.

Future Predictor Analysis. We analyze how the Future Predictor responds to different action conditions under the same scene context. For each scene, we keep the history representation fixed and laterally shift the ground-truth trajectory using 15 offsets uniformly spaced from -0.8 to 0.8 in the normalized action space. Each shifted trajectory is fed into the Future Predictor together with the fixed history representation to obtain the corresponding predicted future representation. Fig. 4(a) plots the cosine similarities between representations predicted under different trajectory conditions, while Fig. 4(b) presents the corresponding pairwise similarity matrix. The cosine similarity generally decreases as the difference between lateral offsets increases, indicating that the predicted representations are sensitive to the trajectory condition even when the scene context remains unchanged.

5 Conclusion

We presented ReDrive, an end-to-end driving framework that learns planning-oriented visual representations through trajectory-conditioned future representation prediction. By jointly learning trajectory generation and action-conditioned future representation prediction, ReDrive incorporates supervision on future scene evolution into visual representation learning. The planner is further adapted using its own generated trajectories with supervision from the frozen future predictor. Experiments on NAVSIM v1 and v2, including NavHard, demonstrate strong planning performance. The future predictor is removed at inference, enabling direct planning without additional modules.

AI USE STATEMENT

We used generative AI tools to assist with identifying relevant literature and improving the readability of the manuscript. All suggested references were manually verified, and all AI-assisted edits were reviewed and revised by the authors. We take full responsibility for the final content of this work.

REFERENCES

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, et al. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mahmoud Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video. *arXiv preprint arXiv:2404.08471*, 2024.

Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11621–11631, 2020.

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

Wei Cao, Marcel Hallgarten, Tianyu Li, Daniel Dauner, Xunjiang Gu, Caojun Wang, Yakov Miron, Marco Aiello, Hongyang Li, Igor Gilitschenski, Boris Ivanovic, Marco Pavone, Andreas Geiger, and Kashyap Chitta. Pseudo-simulation for autonomous driving. In *9th Annual Conference on Robot Learning*, 2025.

Kashyap Chitta, Aditya Prakash, Bernhard Jaeger, Zehao Yu, Katrin Renz, and Andreas Geiger. Transfuser: Imitation with transformer-based sensor fusion for autonomous driving. *IEEE transactions on pattern analysis and machine intelligence*, 45(11):12878–12895, 2022.

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

Hao Gao, Shaoyu Chen, Bo Jiang, Bencheng Liao, Yiang Shi, Xiaoyang Guo, Yuechuan Pu, Haoran Yin, Xiangyu Li, Xinbang Zhang, Ying Zhang, Wenyu Liu, Qian Zhang, and Xinggang Wang. RAD: Training an end-to-end driving policy via large-scale 3DGS-based reinforcement learning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Hao Gao, Shaoyu Chen, Yifan Zhu, Yuehao Song, Wenyu Liu, Qian Zhang, and Xinggang Wang. Rad-2: Scaling reinforcement learning in a generator-discriminator framework. *arXiv preprint arXiv:2604.15308*, 2026.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15979–15988, 2022.

Yufeng Hong, Xiaotian Zhou, Yingyan Li, Xiangpo Zhou, Lin Liu, Yadan Luo, Shaoqing Xu, Lei Yang, and Ziying Song. Drivefuture: Future-aware latent world models for autonomous driving. *arXiv preprint arXiv:2605.09701*, 2026.

Shengchao Hu, Li Chen, Penghao Wu, Hongyang Li, Junchi Yan, and Dacheng Tao. 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.

Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhai Wang, et al. Planning-oriented autonomous driving. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17853–17862. IEEE, 2023.

Minqing Huang, Yujiao Xiang, Zihan Liang, Jiajie Huang, Jingqi Wang, Zhi Xu, Feiyang Tan, Hangning Zhou, Mu Yang, and Gong Che. Coworld-vla: Thinking in a multi-expert world model for autonomous driving. *arXiv preprint arXiv:2605.10426*, 2026.

Bo Jiang, Shaoyu Chen, Qing Xu, Bencheng Liao, Jiajie Chen, Helong Zhou, Qian Zhang, Wenyu Liu, Chang Huang, and Xinggang Wang. Vad: Vectorized scene representation for efficient autonomous driving. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8306–8316. IEEE, 2023.

Bo Jiang, Shaoyu Chen, Bencheng Liao, Xingyu Zhang, Wei Yin, Qian Zhang, Chang Huang, Wenyu Liu, and Xinggang Wang. Senna: Bridging large vision-language models and end-to-end autonomous driving. *arXiv preprint arXiv:2410.22313*, 2024.

Bo Jiang, Shaoyu Chen, Qian Zhang, Wenyu Liu, and Xinggang Wang. Alphadrive: Unleashing the power of vlms in autonomous driving via reinforcement learning and reasoning. *arXiv preprint arXiv:2503.07608*, 2025.

Bo Jiang, Shaoyu Chen, Hao Gao, Bencheng Liao, Qian Zhang, Wenyu Liu, and Xinggang Wang. VADv2: End-to-end autonomous driving via probabilistic planning. In *The Fourteenth International Conference on Learning Representations*, 2026.

Ellington Kirby, Alexandre Boulch, Yihong Xu, Yuan Yin, Gilles Puy, Éloi Zablocki, Andrei Bursuc, Spyros Gidaris, Renaud Marlet, Florent Bartoccioni, Anh-Quan Cao, Nermin Samet, Tuan-Hung VU, and Matthieu Cord. Driving on registers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 32058–32069, June 2026.

Jingyu Li, Zhe Liu, Dongnan Hu, Junjie Wu, Zipei Ma, Wenxiao Wu, Chao Han, Zhihui Hao, Zhikang Liu, Kun Zhan, et al. Metis: A generalizable and efficient world-action model for autonomous driving and urban navigation. *arXiv preprint arXiv:2606.15869*, 2026a.

Jingyu Li, Junjie Wu, Dongnan Hu, Xiangkai Huang, Bin Sun, Zhihui Hao, Xianpeng Lang, Xiatian Zhu, and Li Zhang. Sgdrive: Scene-to-goal hierarchical world cognition for autonomous driving. *arXiv preprint arXiv:2601.05640*, 2026b.

Kailin Li, Zhenxin Li, Shiyi Lan, Yuan Xie, Zhizhong Zhang, Jiayi Liu, Zuxuan Wu, Zhiding Yu, and Jose M Alvarez. Hydra-mdp++: Advancing end-to-end driving via expert-guided hydra-distillation. *arXiv preprint arXiv:2503.12820*, 2025a.

Yingyan Li, Lue Fan, Jiawei He, Yuqi Wang, Yuntao Chen, Zhaoxiang Zhang, and Tieniu Tan. Enhancing end-to-end autonomous driving with latent world model. In *International Conference on Learning Representations*, volume 2025, pp. 42942–42959, 2025b.

Yingyan Li, Yuqi Wang, Yang Liu, Jiawei He, Lue Fan, and Zhaoxiang Zhang. End-to-end driving with online trajectory evaluation via bev world model. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 27137–27146. IEEE, 2025c.

Yongkang Li, Kaixin Xiong, Xiangyu Guo, Fang Li, Sixu Yan, Gangwei Xu, Lijun Zhou, Long Chen, Haiyang Sun, Bing Wang, Kun Ma, et al. Recogdrive: A reinforced cognitive framework for end-to-end autonomous driving. In *International Conference on Learning Representations*, volume 2026, pp. 157518–157556, 2026c.

Zhenxin Li, Kailin Li, Shihao Wang, Shiyi Lan, Zhiding Yu, Yishen Ji, Zhiqi Li, Ziyue Zhu, Jan Kautz, Zuxuan Wu, et al. Hydra-mdp: End-to-end multimodal planning with multi-target hydra-distillation. *arXiv preprint arXiv:2406.06978*, 2024a.

Zhenxin Li, Wenhao Yao, Zi Wang, Xinglong Sun, Joshua Chen, Nadine Chang, Maying Shen, Zuxuan Wu, Shiyi Lan, and Jose M Alvarez. Generalized trajectory scoring for end-to-end multimodal planning. *arXiv preprint arXiv:2506.06664*, 2025d.

Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Qiao Yu, and Jifeng Dai. Bevformer: learning bird's-eye-view representation from lidar-camera via spatiotemporal transformers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(3):2020–2036, 2024b.

Bencheng Liao, Shaoyu Chen, Haoran Yin, Bo Jiang, Cheng Wang, Sixu Yan, Xinbang Zhang, Xiangyu Li, Ying Zhang, Qian Zhang, et al. Diffusiondrive: Truncated diffusion model for end-to-end autonomous driving. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12037–12047. IEEE, 2025.

Lin Liu, Caiyan Jia, Guanyi Yu, Ziying Song, JunQiao Li, Feiyang Jia, Peiliang Wu, Xiaoshuai Hao, and Yadan Luo. Guideflow: Constraint-guided flow matching for planning in end-to-end autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3719–3728, 2026.

Hongbo Lu, Liang Yao, Chenghao He, Haoyu Wang, Xiang Gu, Xianfei Li, Wenlong Liao, Tao He, and Pai Peng. The dawn of world-action interactive models. *arXiv preprint arXiv:2605.11550*, 2026.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4195–4205, 2023.

Shuyao Shang, Yuntao Chen, Yuqi Wang, Yingyan Li, and ZHAO-XIANG ZHANG. Drivedpo: Policy learning via safety dpo for end-to-end autonomous driving. *Advances in Neural Information Processing Systems*, 38:81565–81585, 2026.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Yuehao Song, Shaoyu Chen, Hao Gao, Yifan Zhu, Weixiang Yue, Jialv Zou, Bo Jiang, Zihao Lu, Yu Wang, Qian Zhang, and Xinggang Wang. Senna-2: Aligning vlm and end-to-end driving policy for consistent decision making and planning. *arXiv preprint arXiv:2603.11219*, 2026.

Wenchao Sun, Xuewu Lin, Keyu Chen, Zixiang Pei, Xiang Li, Yining Shi, and Sifa Zheng. Sparsedrivev2: Scoring is all you need for end-to-end autonomous driving. In *European Conference on Computer Vision*, pp. 446–463. Springer, 2026.

Linbo Wang, Yupeng Zheng, Qiang Chen, Shiwei Li, Yichen Zhang, Zebin Xing, Qichao Zhang, Xiang Li, Deheng Qian, Pengxuan Yang, et al. Latent-wam: Latent world action modeling for end-to-end autonomous driving. *arXiv preprint arXiv:2603.24581*, 2026a.

Linhan Wang, Zichong Yang, Chen Bai, Guoxiang Zhang, Xiaotong Liu, Xiaoyin Zheng, Xiao-Xiao Long, Chang-Tien Lu, and Cheng Lu. Drive-jepa: Video jepa meets multimodal trajectory distillation for end-to-end driving. *arXiv preprint arXiv:2601.22032*, 2026b.

Tianze Xia, Yongkang Li, Lijun Zhou, Jingfeng Yao, Kaixin Xiong, Haiyang Sun, Bing Wang, Kun Ma, Guang Chen, Hangjun Ye, et al. Drivelaw: Unifying planning and video generation in a latent driving world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 39701–39712, 2026a.

Tianze Xia, Lijun Zhou, Kaixin Xiong, Jingfeng Yao, Yu Zhu, Zhenxin Zhu, Bing Wang, Guang Chen, Hangjun Ye, Wenyu Liu, et al. Reworld: Learning better representations for world action models. *arXiv preprint arXiv:2606.27504*, 2026b.

Yining Xing, Zehong Ke, Zhiyuan Liu, Yanbo Jiang, Wenhao Yu, and Jianqiang Wang. Clear: Cognition and latent evaluation for adaptive routing in end-to-end autonomous driving. *arXiv preprint arXiv:2606.06219*, 2026a.

Yining Xing, Zhiyuan Liu, Zehong Ke, Wenhao Yu, and Jianqiang Wang. Drift: Drift and aggregation for motion planning. *arXiv preprint arXiv:2607.14507*, 2026b.

Zebin Xing, Xingyu Zhang, Yang Hu, Bo Jiang, Tong He, Qian Zhang, Xiaoxiao Long, and Wei Yin. Goalflow: Goal-driven flow matching for multimodal trajectories generation in end-to-end autonomous driving. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1602–1611. IEEE, 2025.

Mingwang Xu, Jiahao Cui, Feipeng Cai, Hanlin Shang, Zhihao Zhu, Shan Luan, Yifang Xu, Neng Zhang, Yaoyi Li, Jia Cai, et al. Wam-diff: A masked diffusion vla framework with moe and online reinforcement learning for autonomous driving. *arXiv preprint arXiv:2512.11872*, 2025.

Jiwei Yang, Zhengxian Chen, Chaosheng Huang, and Jun Li. Auto-jepa: A latent world model of continuous intent for end-to-end autonomous driving. *arXiv preprint arXiv:2607.29031*, 2026a.

Pengxuan Yang, Yupeng Zheng, Deheng Qian, Zebin Xing, Qichao Zhang, Linbo Wang, Yichen Zhang, Shaoyu Guo, Zhongpu Xia, Qiang Chen, et al. Dreamerad: Efficient reinforcement learning via latent world model for autonomous driving. *arXiv preprint arXiv:2603.24587*, 2026b.

Wenhao Yao, Zhenxin Li, Shiyi Lan, Zi Wang, Xinglong Sun, Jose M Alvarez, and Zuxuan Wu. Drivesuprim: Towards precise trajectory selection for end-to-end planning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 11910–11918, 2026.

Bozhou Zhang, Nan Song, Xiatian Zhu, Jiankang Deng, Li Zhang, et al. Future-aware end-to-end driving: Bidirectional modeling of trajectory planning and scene evolution. *Advances in Neural Information Processing Systems*, 38:10204–10229, 2026.

Kaiwen Zhang, Zhenyu Tang, Xiaotao Hu, Xingang Pan, Xiaoyang Guo, Yuan Liu, Jingwei Huang, Li Yuan, Qian Zhang, Xiao-Xiao Long, et al. Epona: Autoregressive diffusion world model for autonomous driving. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 27220–27230. IEEE, 2025.

Yupeng Zheng, Pengxuan Yang, Zebin Xing, Qichao Zhang, Yuhang Zheng, Yinfeng Gao, Pengfei Li, Teng Zhang, Zhongpu Xia, Peng Jia, et al. World4drive: End-to-end autonomous driving via intention-aware physical latent world model. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 28632–28642. IEEE, 2025.

Zhiyu Zheng, Shaoyu Chen, Haoran Yin, Xinbang Zhang, Jialv Zou, Xinggang Wang, Qian Zhang, and Lefei Zhang. Resad: Normalized residual trajectory modeling for end-to-end autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3729–3739, June 2026.

Ruiguo Zhong, Benshan Ma, Xiaolong Chen, Lang Zhang, Mingyue Feng, Yaonong Wang, Pei Liu, and Jun Ma. Da-wam: Decision-aligned future latents for driving world models. *arXiv preprint arXiv:2608.19085*, 2026.

Jialv Zou, Shaoyu Chen, Bencheng Liao, Zhiyu Zheng, Yuehao Song, Lefei Zhang, Qian Zhang, Wenyu Liu, and Xinggang Wang. Diffusiondrivev2: Reinforcement learning-constrained truncated diffusion modeling in end-to-end autonomous driving. *arXiv preprint arXiv:2512.07745*, 2025.

A Appendix

A.1 Training Details

A.1.1 Training Data

nuPlan. nuPlan (Caesar et al., 2021) is a large-scale autonomous driving dataset and planning benchmark containing approximately 1,200 hours of real-world driving data collected in Boston, Pittsburgh, Las Vegas, and Singapore. Among them, 120 hours are released with the full sensor suite, including eight cameras, five LiDARs, IMU, and GPS, together with detailed HD maps and automatically generated 3D annotations. The dataset covers more than 30 types of driving scenarios, including lane changes, unprotected turns, and interactions with pedestrians. For pretraining, we use the front-view camera sequences from the NAVTRAIN sensor data derived from nuPlan.

nuScenes. nuScenes (Caesar et al., 2020) is a large-scale multimodal autonomous driving dataset collected in Boston and Singapore. It contains 1,000 driving scenes, each approximately 20 seconds long, with 1.4 million camera images and 390,000 LiDAR sweeps. The sensor suite consists of six cameras, one LiDAR, five radars, IMU, and GPS, providing complete 360-degree observations of the surrounding environment. The official dataset is divided into 700 training scenes, 150 validation scenes, and 150 test scenes. For pretraining, we use the `CAM_FRONT` stream from all 700 training scenes.

PhysicalAI-Autonomous-Vehicles Dataset. PhysicalAI-Autonomous-Vehicles Dataset is a large-scale multi-sensor autonomous driving dataset released by NVIDIA. The full dataset contains approximately 1,700 hours of driving data and 306,152 clips, where each clip has a duration of 20 seconds. The data are collected across 25 countries and more than 2,500 cities, covering diverse traffic, weather, road, and geographic conditions. All clips provide seven-camera observations, while 298,326 clips additionally contain LiDAR data and 160,761 clips contain radar data. The complete dataset occupies approximately 133 TB. For pretraining, we use an 80-hour subset and retain the front wide-angle camera with a 120° field of view.

Camera Alignment. The camera configurations of PhysicalAI-Autonomous-Vehicles Dataset and nuPlan (Caesar et al., 2021) differ substantially in field of view and projection model. To reduce this domain discrepancy, we geometrically align the PhysicalAI-Autonomous-Vehicles Dataset front wide-angle videos to the nuPlan front-camera view before pretraining. Specifically, we use the per-clip f-theta camera calibration provided by PhysicalAI-Autonomous-Vehicles Dataset to map each target nuPlan pixel ray back to the corresponding source image location, followed by bilinear resampling. The resulting videos approximately match the field of view and camera intrinsics of the nuPlan front camera. We additionally discard clips whose calibrated field of view is insufficient to cover the target view, avoiding extrapolation during remapping.

A.1.2 Training Procedure

Pretraining. We initialize the visual encoder from the publicly released V-JEPA2 ViT-L (Assran et al., 2025) checkpoint and further pretrain it on a mixture of driving videos from nuPlan (Caesar et al., 2021), nuScenes (Caesar et al., 2020), and PhysicalAI-Autonomous-Vehicles Dataset. The three datasets are uniformly sampled during training. Each training clip contains 16 frames sampled at 10 Hz with a spatial resolution of 256×512. We follow the V-JEPA2 pretraining recipe and adopt a ViT-L encoder together with a lightweight predictor and an EMA target encoder. A multi-block masking strategy is applied over the full temporal extent, and the predictor learns to reconstruct the masked latent representations produced by the target encoder. After pretraining, only the visual encoder is retained for subsequent training, while the pretraining predictor and target encoder are discarded. The detailed pretraining configuration is summarized in Tab. 8.

Joint Training. After pretraining, we retain the visual encoder and jointly optimize it with a newly initialized Future Predictor and Action DiT. Each sample contains four observed frames at 256×512 resolution. The Future Predictor takes the history representation and ego trajectory as conditions to predict the target representations of the subsequent four frames, while the Action DiT predicts an 8-step trajectory covering 4 s at 2 Hz. A target encoder is maintained as an exponential moving average (EMA) of the visual encoder with a decay of 0.9999 and remains gradient-free throughout

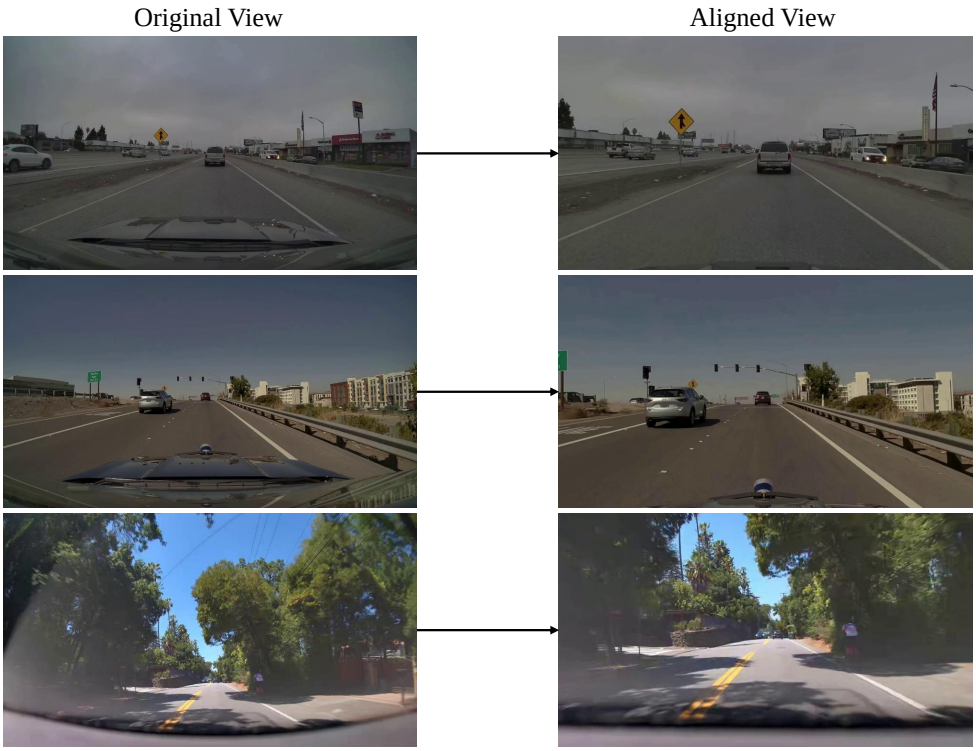


Figure 5: Illustration of camera alignment for PhysicalAI-Autonomous-Vehicles Dataset. The original wide-angle camera views are geometrically transformed using per-clip calibration to obtain aligned views with nuPlan-style camera geometry. The alignment reduces the discrepancy in camera projection and view distribution before self-supervised pretraining.

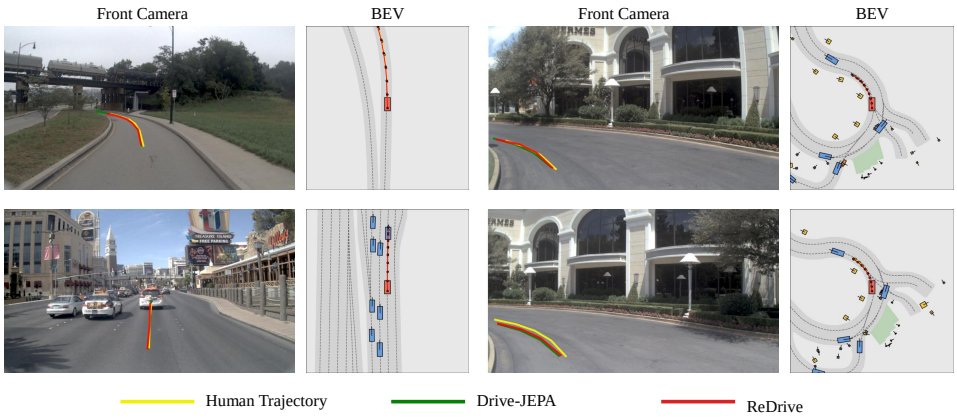


Figure 6: Additional qualitative comparison of the human trajectory, Drive-JEPA, and ReDrive using front-camera and bird's-eye-view (BEV) visualizations.

training. The model architecture and optimization settings are summarized in Tab. 9 and Tab. 10, respectively.

A.2 More Qualitative Results

We provide additional qualitative comparisons between the human trajectory, Drive-JEPA (Wang et al., 2026b), and ReDrive in Fig. 6. The examples cover diverse road geometries and traffic in-

Table 8: Pretraining configuration.

Module	Configuration	Setting
Encoder	Architecture	ViT-L
	Depth	24
	Hidden dimension	1024
	Attention heads	16
	Parameters	304M
Predictor	Depth	12
	Hidden dimension	384
	Attention heads	12
	Parameters	22M
Input	Frames	16
	Frame rate	10 Hz
	Resolution	256×512
	Patch / Tubelet size	16×16 / 2
Masking	Small blocks	8, scale 0.15
	Large blocks	2, scale 0.70
Optimization	Optimizer	AdamW
	Learning rate	5.25×10^{-4}
	Weight decay	0.04
	Precision	bfloat16
	Epochs	100
Target Encoder	Update	EMA
	Decay	0.99925

Table 9: Model configuration for joint training.

Module	Configuration	Setting
Encoder	Architecture	ViT-L
	Depth	24
	Hidden dim.	1024
	Attention heads	16
	Parameters	304M
Future Predictor	Depth	12
	Hidden dim.	768
	Attention heads	12
	MLP ratio	4
	Parameters	153M
Action DiT	Depth	28
	Hidden dim.	512
	Attention heads	16
	Parameters	103M
Target Encoder	Update	EMA
	Decay	0.9999

Table 10: Optimization configuration for joint training.

Configuration	Setting
Observed frames	4
Predicted frames	4
Resolution	256×512
Optimizer	AdamW
Learning rate	3×10^{-5}
β_1, β_2	0.9, 0.95
Weight decay	1×10^{-5}
LR schedule	Constant w/ warmup
Warmup steps	1,000
Gradient clipping	1.0
Planning loss	Flow-matching MSE
Prediction loss	L1
Prediction weight	0.1
Diffusion timesteps	1,000
Inference steps	5
Precision	bfloat16

teractions, with both front-camera and bird's-eye-view (BEV) visualizations. These results further illustrate the planning behavior of ReDrive across different driving scenarios.

A.3 Additional Future Predictor Visualizations

We provide additional visualizations of the Future Predictor across diverse driving scenes. We keep the history representation fixed and vary the lateral trajectory condition to examine the corresponding

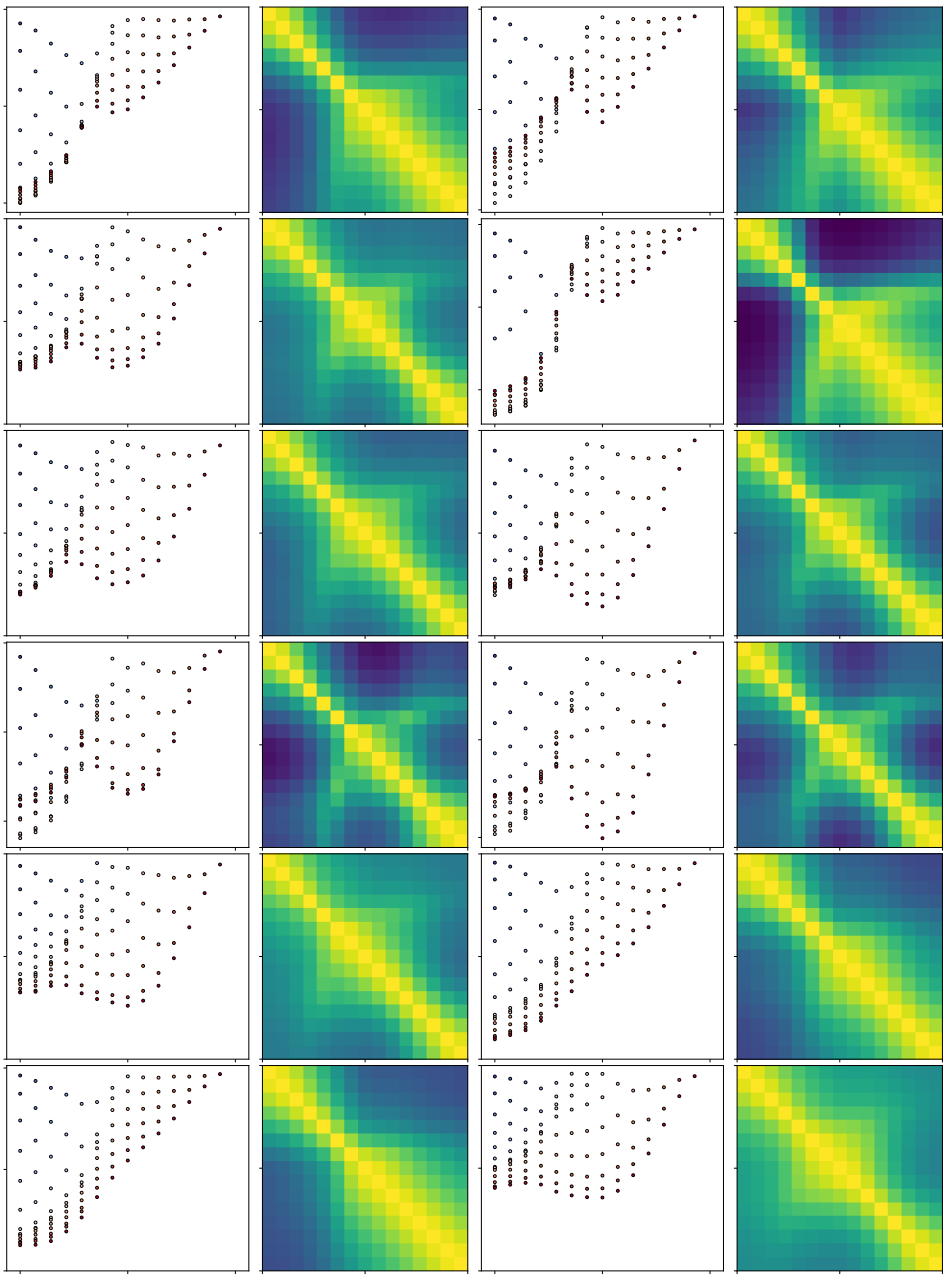


Figure 7: Additional Future Predictor visualizations. Predicted future representations under different lateral trajectory conditions across diverse driving scenes. The representations exhibit consistent action-dependent variations across different scenarios.

changes in predicted future representations. As shown in Fig. 7, the representations consistently vary with the trajectory condition across different scenes. The pairwise similarity matrices also exhibit a clear decay as the difference between action conditions increases. These results further show that the Future Predictor captures action-dependent future scene evolution across diverse driving scenarios.