DRIVEWEAVE-POLICY: WEAVING SPARSE FUTURE IMAGINATION AND TRAJECTORY PLANNING THROUGH ALIGNED DENOISING

Anonymous authors
Paper under double-blind review

ABSTRACT

A central challenge for driving world-action models (WAMs) is how to effectively leverage visual generative priors for trajectory planning. Yet driving scenes are highly dynamic, and densely generating future visual observation imposes a substantial modeling burden on trajectory planning. Separate diffusion branches offer a straightforward way to reuse pretrained generative models in driving WAMs. However, coordinating these branches through LLM-mediated feature exchange alone may be insufficient to ensure consistency between imagined futures and planned actions. In fact, imagination and planning reflect the same underlying scene dynamics, and we accordingly introduce DriveWeave. Unlike existing frame-by-frame autoregressive unified WAMs, DriveWeave is designed to better harness visual generative priors at a lower sampling frequency. Within this framework, we compare three training paradigms for learning and transferring cross-modal priors: Imagine-then-Plan, Plan-then-Imagine, and Coupled Denoising. We find that sparse anchored imagination and coupled denoising best match the tight perception–action loop of driving. Furthermore, we introduce a learnable trajectory encoding method that brings the distributions of low-dimensional trajectory features and high-dimensional visual features closer together in a shared latent space. We coordinate visual and trajectory denoising through two complementary mechanisms: cross-modal information exchange and dynamic alignment. For information exchange, shared attention enables bidirectional conditioning at every denoising step. For dynamic alignment, we reformulate the reverse-diffusion dynamics with a consistency-driven coupling term that adjusts modality-specific updates based on agreement between predicted futures. On NAVSIM, DriveWeave achieves a PDMS of 91.3, the best performance among the evaluated methods.

1 INTRODUCTION

Driving is a coupling between imagination and action. A skilled driver sees "what the scene will look like if I go this way" and "what I should do now" as two intertwined, mutually corrective thoughts. Imagination and action share a single internal belief about how the scene evolves (Zhang et al., 2025; Xia et al., 2026; Chen et al., 2025). This motivates a central question for driving world-action models (WAMs): how can future visual imagination effectively support trajectory planning?

Pretrained video diffusion models provide rich visual and temporal priors for planning (Ho et al., 2022; Yang et al., 2024; Hu et al., 2023). But existing driving WAMs largely inherit embodied-agent designs without explicitly addressing the temporal coupling between ego motion and future observations. Ego motion shapes future viewpoints, requiring predicted observations to be consistent with the planned trajectory and its terminal state. These WAMs typically use multi-branch designs that denoise modalities separately, limiting cross-modal interaction to LLM-mediated feature exchange without explicitly coordinating generative updates. This may yield inconsistent imagined futures and planned actions. However, jointly denoising still faces the substantial burden of dense future-frame generation in dynamic driving scenes, which may compromise trajectory precision.

Since imagination and planning reflect the same scene dynamics, their interaction should not be confined to the LLM but rather under a Diffusion Transformer (DiT) (Peebles & Xie, 2023; Lipman

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.

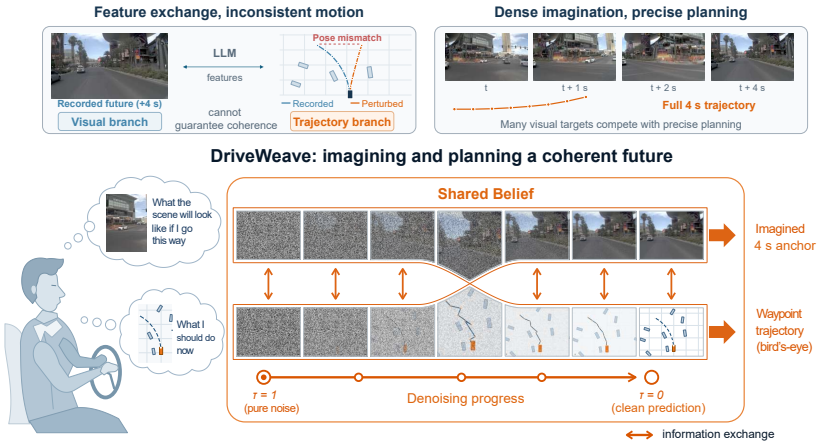


Figure 1: Why couple imagination and planning? Separate visual and trajectory predictions can imply different ego motion, while dense visual targets increase the generation burden. DriveWeave combines bidirectional information exchange with pose-consistent denoising and sparse future anchors.

et al., 2023). We introduce **DriveWeave**, a unified WAM with controllable cross-modal information flow, supporting three paradigms: *Imagine-then-Plan*, *Plan-then-Imagine*, and *Coupled Denoising* (Fig. 3). The first two learn the visual or planning objective first, whereas the third learns them jointly through bidirectional interaction between their intermediate representations. This formulation provides a common architectural basis for studying how cross-modal interaction affects planning.

Shared attention enables cross-modal interaction but does not ensure consistent predictions. Trajectory and visual representations differ in structure, scale, and precision requirements, so sharing a diffusion timestep alone does not coordinate their updates. We therefore address joint generation through two complementary mechanisms: information exchange and dynamic alignment. For **information exchange**, a learnable trajectory codec brings trajectory and visual latent distributions closer in a shared space, while shared attention enables bidirectional conditioning. For **dynamic alignment**, we introduce pose-consistency coupling into the reverse-diffusion formulation, shaping the learned velocity field and coordinating modality-specific updates. This encourages coherent imagined observations and planned trajectories beyond latent distribution matching.

To reduce the visual-generation burden, we introduce **Sparse Anchored Future Imagination**. Rather than densely generating every future frame, DriveWeave predicts visual observations at selected timestamps within the planning horizon while producing the full trajectory. This reduces temporal sampling density without shortening the planning horizon, concentrating generation on a compact set of future visual targets. Together with coupled denoising, these anchors allow us to study how much future visual generation benefits planning and at what computational cost.

We evaluate DriveWeave on NAVSIM, comparing the three paradigms under specified training and inference settings, with ablations assessing information exchange, consistency training, sampling correction, and visual target density. Consistency training improves unguided planning and further improves guided planning at comparable inference cost. Our main contributions are as follows:

- We introduce DriveWeave, a unified WAM that models imagination and planning within a shared DiT, systematically comparing three cross-modal interaction paradigms.
- We distinguish information exchange from dynamic alignment, combining codec-supported bidirectional attention with pose-consistency coupling that shapes the learned velocity field and reverse-diffusion updates.
- We propose Sparse Anchored Future Imagination to reduce dense future-generation demands while preserving the planning horizon.
- DriveWeave achieves the best performance among the evaluated methods on the **NAVSIM** benchmark.

2 Related Work

2.1 End-to-End Driving and NAVSIM

End-to-end driving learns policies from sensor observations to trajectories or control commands, using jointly learned or implicit intermediate representations (Hu et al., 2023; Weng et al., 2024; Chitta et al., 2023). By learning these components from driving data, end-to-end approaches can potentially exploit correlations that are difficult to capture in modular pipelines. However, most end-to-end driving systems focus primarily on predicting the ego trajectory, while the visual evolution of the scene is either omitted or modeled as an auxiliary task (Wu et al., 2022; Li et al., 2024; Sun et al., 2025). This limits their ability to exploit rich visual priors learned from large-scale video data. Recent benchmarks such as NAVSIM provide a standardized setting for evaluating vision-based driving policies with open-loop trajectory metrics and realistic driving constraints (Dauner et al., 2024; Li et al., 2024; Shang et al., 2025). Our work focuses on how future visual generation can support planning through coordinated visual and trajectory predictions.

2.2 Diffusion Models for Planning

Diffusion models support multimodal trajectory generation and visuomotor policy learning (Janner et al., 2022; Ajay et al., 2023; Chi et al., 2023). In driving, MotionDiffuser models multi-agent motion distributions (Jiang et al., 2023), while DiffusionDrive uses truncated diffusion to improve trajectory generation efficiency (Liao et al., 2025). Beyond conditional generation, energy-guided methods steer diffusion and flow-matching dynamics toward desired objectives (Feng et al., 2025). DriveWeave applies this to cross-modal physical consistency: a shared pose-consistency energy coordinates the motion implied by visual, trajectory, and future ego-state predictions in a joint generative process.

2.3 World Models and World-Action Models

World models learn environment dynamics for prediction and control (Ha & Schmidhuber, 2018; Hafner et al., 2020; 2025), including model-based driving policies (Hu et al., 2022). Large-scale video models provide visual and temporal priors motivating joint modeling of future observations and actions (Ho et al., 2022; Brooks et al., 2024; NVIDIA, 2025). Driving WAMs organize this interaction differently. Epona combines a multimodal spatiotemporal context model with separate visual and trajectory diffusion modules (Zhang et al., 2025). DrivingGPT unifies visual and action tokens via autoregressive prediction (Chen et al., 2025), while DriveWAM adapts a pretrained video DiT with a unified video–action sequence and joint flow matching (Shi et al., 2026). These approaches establish diverse cross-modal conditioning. DriveWeave studies complementary questions: how training order affects visual-prior use, how pose-consistency coupling coordinates heterogeneous predictions, and how sparse visual anchors support planning within a fixed horizon.

3 Method

Future imagination can support planning when the generated observation and the trajectory describe a compatible ego motion. DriveWeave builds this connection within a pretrained video Diffusion Transformer (DiT), as shown in Fig. 2. Shared attention lets each prediction inform the other; pose-consistency coupling uses their disagreement to shape the learned velocity field and its reverse-time updates. Sparse visual anchors focus this interaction on selected future outcomes within the full planning horizon.

3.1 Task and Joint Representation

Given current left/front/right camera observations and ego measurements c, we predict a four-second trajectory $a = \{(p_i, \psi_i)\}_{i=1}^{H}$, with $H = 8$ and waypoint times $t_i = 0.5i\,\mathrm{s}$. Positions $p_i = (x_i, y_i)$ and headings ψ_i are expressed in the current ego frame. Visual targets lie at selected timestamps $\mathcal{A} = \{t_{k_1}, \ldots, t_{k_K}\}$, with $t_{k_K} = 4\,\mathrm{s}$. Our default uses only the terminal visual anchor.

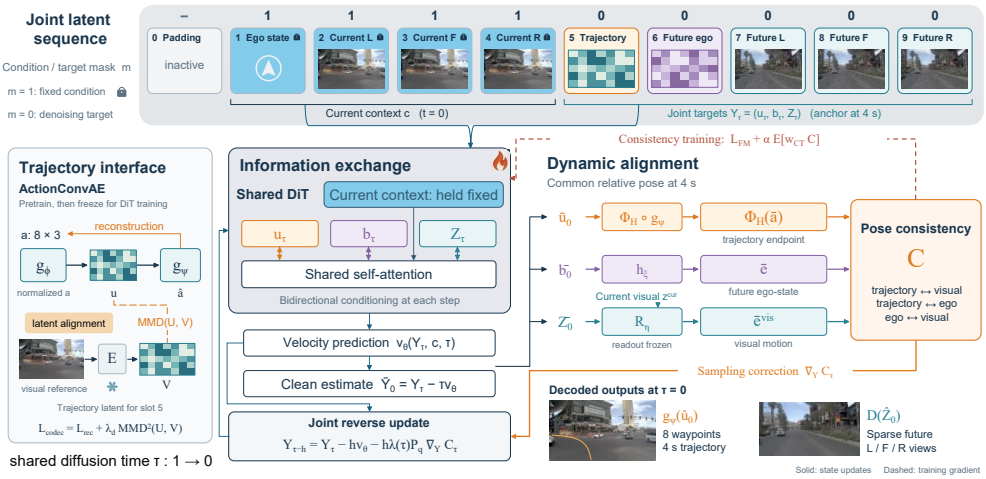


Figure 2: DriveWeave overview. Current observations and ego state form fixed conditions. Trajectory, future ego-state, and visual latents share a DiT. ActionConvAE provides the trajectory interface; a common pose-consistency energy shapes the learned velocity field and corrects reverse-time updates. The condition mask separates observed context from generated targets.

Let $\Phi_i(a) = (p_i, \psi_i)$. We also predict the terminal ego pose $e = \Phi_H(a)$ as an auxiliary target. The clean joint representation is

$$Y_0 = (u_0, b_0, Z_0), \qquad u_0 = g_\phi(a), \qquad b_0 = h_\omega(e), \tag{1}$$

where Z_0 contains the future multi-view latents produced by the frozen visual tokenizer. Dedicated decoders recover $\hat{a} = g_\psi(u)$ and $\hat{e} = h_\xi(b)$. The trajectory is encoded and decoded directly by its dedicated codec. Predicting b supplies a separate terminal-pose estimate that can be compared with both the trajectory endpoint and the imagined observation.

All modalities are arranged as semantic slots with modality, camera, and physical-time embeddings. A slot can contain multiple Transformer tokens. For one anchor, the sequence comprises padding, current ego state, three current images, trajectory, future ego state, and three future images. Participation r_i and condition indicator m_i define the target mask $q_i = r_i(1 - m_i)$. Inactive slots are excluded from attention and loss, and current conditions remain fixed throughout sampling. Supplementary Sec. A gives the slot layout and pose conventions.

3.2 Cross-Modal Information Exchange

Trajectory interface. The trajectory interface must preserve metric precision while making action features usable by a pretrained visual backbone. ActionConvAE encodes normalized waypoints into an image-shaped latent and reconstructs their positions and headings. We train it with

$$\mathcal{L}_{\text{codec}} = \mathcal{L}_{\text{rec}}(a, g_\psi(g_\phi(a))) + \lambda_d \widehat{\text{MMD}}^2(U, V), \tag{2}$$

where reconstruction combines position error and circular heading error. U and V are balanced samples of equally sized trajectory and visual latent patches, normalized using fixed training-visual statistics before positional embeddings. Kernel MMD (Gretton et al., 2012) reduces their distribution discrepancy while the visual reference remains fixed. Reconstruction keeps the trajectory recoverable, while distribution matching reduces statistical differences at the shared input interface. Their combination makes latent compatibility serve geometric prediction. Figure 4(a) illustrates the marginal statistics; Supplementary Sec. B gives the patch-level objective.

Shared attention. At each denoising step, trajectory, future ego-state, and future visual tokens attend to one another. The trajectory velocity can depend on the evolving visual prediction, while the visual velocity can depend on the emerging plan. This repeated exchange lets both predictions be revised before either is finalized. Attention paths remain independently controllable; Supplementary Sec. B defines the masks used to isolate information exchange.

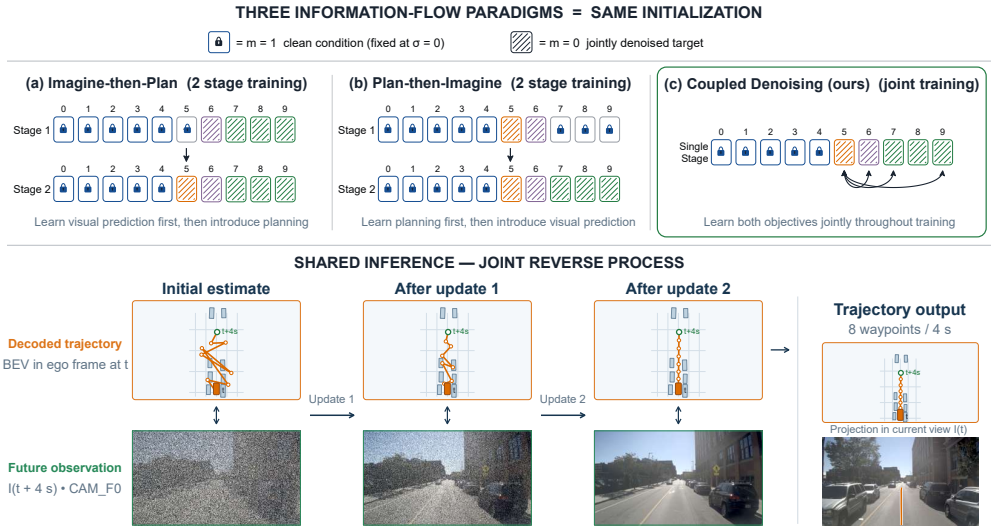


Figure 3: Three training paradigms in a shared architecture. The upper panels show visual-first, planning-first, and joint learning through their active latent slots. In the ordered paradigms, the first objective remains active when the second is introduced. The lower panel illustrates bidirectional exchange between evolving trajectory and visual predictions during joint denoising.

Three training paradigms. We ask whether planning benefits more from an already learned visual objective or from learning both objectives together (Fig. 3). *Imagine-then-Plan* first learns future visual prediction and subsequently introduces trajectory learning while retaining the visual objective. *Plan-then-Imagine* reverses this order. *Coupled Denoising* learns both objectives jointly, allowing their intermediate representations to develop through bidirectional interaction. Each paradigm is trained from the same backbone initialization with matched data and total optimization budgets. All three trained models jointly denoise the active future targets at inference, so the comparison concerns how cross-modal priors develop during learning.

3.3 Pose-Consistent Denoising Dynamics

Feature exchange alone leaves an ambiguity: a turning trajectory and a future view suggesting continued straight motion may each appear plausible in isolation. Similar latent statistics do not resolve this disagreement. We compare the ego motion implied by the predictions at matched timestamps, then use that discrepancy to coordinate their evolution. Physical pose provides the common meaning that their latent coordinates lack.

Base velocity field. We use linear flow matching (Lipman et al., 2023), with data at diffusion time $\tau = 0$ and noise at $\tau = 1$:

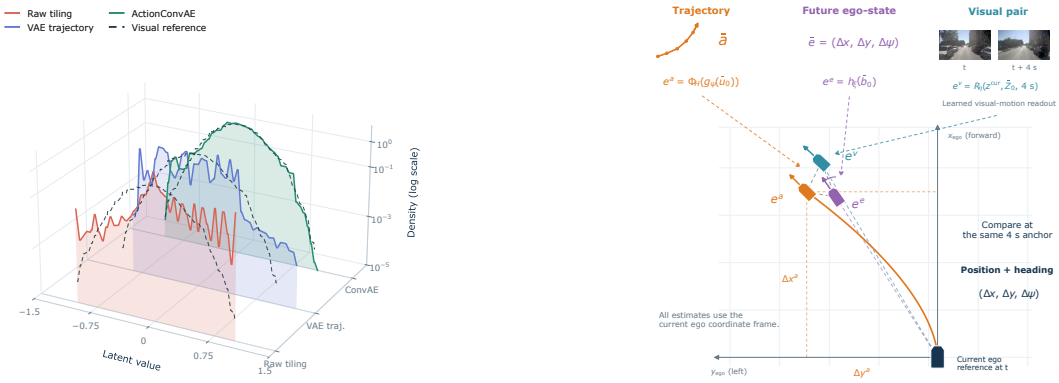
$$Y_\tau = (1-\tau)Y_0 + \tau\epsilon, \quad v_\star = \epsilon - Y_0, \quad \bar{Y}_0 = Y_\tau - \tau v_\theta(Y_\tau, c, \tau). \tag{3}$$

These expressions apply to target blocks, with current conditions fixed. Diffusion time τ controls noise level; physical time t specifies the predicted event. We decode $\bar{a} = g_\psi(\bar{u}_0)$ and $\bar{e} = h_\xi(\bar{b}_0)$, comparing their clean estimates so that the consistency signal concerns the future implied at each step.

Motion implied by visual predictions. Figure 4(b) shows how the three predictions meet in a common pose space. A visual motion readout estimates relative ego pose from current and future visual latents:

$$\bar{e}_j^{\text{vis}} = R_\eta(z^{\text{cur}}, \bar{z}_j, t_{k_j}). \tag{4}$$

The readout receives only visual inputs and the requested time, so its pose estimate must come from the observation pair. We supervise it using recorded relative poses and freeze it during DiT training and sampling. This preserves a common reference for measuring visual motion, while gradients through its inputs connect the pose discrepancy to visual generation.



(a) Latent distribution compatibility

(b) Common relative-pose space

Figure 4: Latent compatibility and physical consistency. (a) Reconstructed marginal density profiles compare trajectory encodings with the visual reference; the depth axis separates representations. (b) Trajectory decoding, future ego-state decoding, and visual motion readout produce relative poses at the same anchor in the current ego frame. Pose disagreement couples their denoising updates.

For poses $e = (p, \psi)$ and $e' = (p', \psi')$, define

$$d(e, e') = \|p - p'\|_2^2 / s_p^2 + \beta_\psi [1 - \cos(\psi - \psi')], \tag{5}$$

with a fixed position scale $s_p = 10\,\mathrm{m}$. Our consistency energy is

$$\mathcal{C}(\bar{Y}_0; c) = \frac{\gamma_{av}}{K} \sum_{j=1}^{K} d\big(\Phi_{k_j}(\bar{a}), \bar{e}_j^{\mathrm{vis}}\big) + \gamma_{ae} d(\Phi_H(\bar{a}), \bar{e}) + \gamma_{ev} d\big(\bar{e}, \bar{e}_K^{\mathrm{vis}}\big)\,. \tag{6}$$

The first term compares visual motion with the trajectory at each anchor. The other two connect both terminal predictions to the auxiliary ego-state. Each prediction is also supervised by its own target. When a modality is absent, we omit its associated consistency terms: visual-first training retains only visual–ego consistency, and planning-first training retains only trajectory–ego consistency. Joint training uses all three terms (Supplementary Sec. B.3).

Learning coordinated predictions. For target dimension d_i and modality weight w_i, the flow loss is

$$\mathcal{L}_{\mathrm{FM}} = \mathbb{E}\left[\frac{\sum_i q_i w_i d_i^{-1} \|v_\theta^i - v_\star^i\|_2^2}{\sum_i q_i w_i}\right]. \tag{7}$$

We apply pose consistency to the clean estimates obtained at the sampled noise level:

$$\mathcal{L}_{\mathrm{train}} = \mathcal{L}_{\mathrm{FM}} + \alpha\, \mathbb{E}_{Y_0, \epsilon, \tau}[w_{\mathrm{CT}}(\tau) \mathcal{C}(Y_\tau - \tau v_\theta; c)]\,. \tag{8}$$

The flow objective teaches each modality to recover its target; the consistency term teaches their clean estimates to agree on motion. Its gradient reaches the DiT through the decoded predictions, shaping the velocity field itself. This gives the model a learned preference for compatible futures even when sampling correction is disabled. We use $w_{\mathrm{CT}}(\tau) = (1 - \tau)^2$ to weight this signal across noise levels.

Coupled reverse-time updates. The same energy also provides a correction specific to the current generated sample. Write $\mathcal{C}_\tau(Y) = \mathcal{C}(Y - \tau v_\theta(Y, c, \tau); c)$ and let P_q be block diagonal with blocks $q_i \rho_i I$, where $\rho_i > 0$ controls each modality's correction scale. We reformulate the reverse dynamics as

$$\frac{dY_\tau}{d\tau} = v_\theta(Y_\tau, c, \tau) + \lambda(\tau) P_q \nabla_Y \mathcal{C}_\tau(Y_\tau). \tag{9}$$

For decreasing times $\tau_k > \tau_{k-1}$ and $h_k = \tau_k - \tau_{k-1}$, the update becomes

$$Y_{\tau_{k-1}} = Y_{\tau_k} - h_k v_\theta - h_k \lambda(\tau_k) P_q \nabla_Y \mathcal{C}_{\tau_k}. \tag{10}$$

Because integration runs toward decreasing τ, the added field produces a negative-gradient correction. A pose discrepancy thus induces changes in the contributing modalities through their own

Table 1: NAVSIM *navtest* results. Higher is better. MV: multi-view cameras; SV: single-view camera; L: LiDAR. Bold: best in the WAM group, including ties.

Method	Ref	Sensors	NC ↑	DAC ↑	TTC ↑	C. ↑	EP ↑	**PDMS**
Human	–	–	100	100	100	99.9	87.5	94.8
Vision-Based End-to-End Methods								
UniAD (Hu et al., 2023)	CVPR'23	MV	97.8	91.9	92.9	100.0	78.8	83.4
TransFuser (Chitta et al., 2023)	TPAMI'23	MV&L	97.7	92.8	92.8	100.0	79.2	84.0
PARA-Drive (Weng et al., 2024)	CVPR'24	MV	97.9	92.4	93.0	99.8	79.3	84.0
LAW (Li et al., 2025)	ICLR'25	SV	96.4	95.4	88.7	99.9	81.7	84.6
DiffusionDrive (Liao et al., 2025)	CVPR'25	MV&L	98.2	96.2	94.7	100.0	82.2	88.1
WoTE (Li et al., 2025)	ICCV'25	MV&L	98.5	96.8	94.4	99.9	81.9	88.3
VLA-based Methods								
ReCogDrive (Li et al., 2026)	ICLR'26	MV	98.1	94.7	94.2	100.0	80.9	86.5
DriveVLA-W0 (Li et al., 2026)	ICLR'26	SV	98.7	96.2	95.5	100.0	82.2	88.4
AutoVLA (Zhou et al., 2025)	NeurIPS'25	MV	98.4	95.6	98.0	99.9	81.9	89.1
DriveDreamer-Policy (Zhou et al., 2026)	arXiv'26	MV	98.4	97.1	95.1	100.0	83.5	89.2
World-Action Models								
Epona (Zhang et al., 2025)	ICCV'25	SV	97.9	95.1	93.8	99.9	80.4	86.2
WorldDrive (Gui et al., 2026)	arXiv'26	SV	98.4	95.8	95.2	99.8	83.3	89.0
DriveWAM (Shi et al., 2026)	arXiv'26	SV	98.3	98.1	95.2	**100.0**	84.3	90.1
DriveWeave (Imagine→Plan)	–	MV	97.7	95.8	95.8	99.8	80.2	86.6
DriveWeave (Plan→Imagine)	–	MV	98.1	95.1	94.8	99.9	81.4	86.8
DriveWeave (Coupled)	–	MV	**99.0**	**98.6**	**96.1**	**100.0**	**85.4**	**91.3**

prediction paths, without comparing latent velocity directions directly. This supplies sample-specific coordination on top of the learned field. We use $\lambda(\tau) = \lambda_0(1-\tau)^2$ to emphasize correction as the predictions become cleaner. On the two-step grid $[1, 0.5, 0]$, $\lambda(1) = 0$, so only the second update receives sampling correction. The first update still uses the velocity field shaped by consistency training. Supplementary Sec. C gives the gradients and two-step expansion.

From pose disagreement to latent correction. Writing the collected poses as $F(\bar{Y}_0)$, the correction gradient is $\nabla_{Y_\tau}\mathcal{C}_\tau = (I - \tau J_{v_\theta, Y_\tau})^\top J^\top_{F,\bar{Y}_0} \nabla_F \mathcal{C}$. The readout and decoders map a common pose error into distinct latent corrections, accounting for the clean prediction's dependence on the velocity field. This gradient translates physical disagreement into coordinated updates.

3.4 Sparse Anchored Future Imagination

A visual anchor gives the plan a future viewpoint to agree with. The terminal observation is particularly useful because its implied ego pose can be matched directly to the trajectory endpoint. We retain all eight supervised waypoints while generating only selected visual anchors, allowing intermediate motion to be learned without reconstructing every intermediate image.

With n_z tokens per camera observation, K anchors require $3Kn_z$ future visual tokens. A single terminal anchor uses one eighth of the visual targets required by half-second sampling over the same four-second horizon. This concentrates visual supervision on a compact future target while preserving the full trajectory. All targets are jointly denoised in two reverse steps in our fast setting. An endpoint observation does not uniquely determine the path. Its constraint reaches intermediate waypoints through the shared trajectory representation and attention, while supervision on all eight waypoints learns the route to that endpoint.

4 Experiments

4.1 Experimental Setup

We evaluate four-second trajectories on the original NAVSIM *navtest* protocol using its non-reactive simulator (Dauner et al., 2024). We report PDM Score (PDMS), no at-fault collision (NC), drivable



Figure 5: Qualitative examples. Current front-camera observations and BEV plots show predicted four-second trajectories (green) and human references (gray). The future left/front/right observations are decoded after one denoising update.

area compliance (DAC), time-to-collision (TTC), comfort, and ego progress (EP), all on a 0–100 scale. Inputs are current left/front/right camera observations and ego measurements.

Training and inference. We fine-tune Cosmos-Predict2-2B (NVIDIA, 2025) for 40K updates using AdamW (Loshchilov & Hutter, 2019), batch size 24, peak LR 10^{-4}, and 1,000 warm-up steps. Inputs are resized to 224×224, and experiments use NVIDIA H20 GPUs. Pretrained codecs and visual readout remain frozen. Imagine-then-Plan uses 30K visual-only and 10K joint updates; Plan-then-Imagine uses 10K planning-only and 30K joint updates. The auxiliary ego-state is predicted throughout. Coupled Denoising uses 40K joint updates. All three share initialization, data, and total training budget, and use two-step joint inference with the same sampling correction. Mechanism ablations retain the same representation, visual targets, and denoising budget.

Cross-modal agreement. At four seconds, E_{pose} is the mean of the three pairwise distances in Eq. equation 5 among the trajectory endpoint, auxiliary ego-state, and visual-readout pose. Fixed normalization makes this a dimensionless agreement metric. PDMS evaluates the resulting planning behavior alongside this cross-modal agreement. A held-out future-image shuffling test examines whether the readout uses scene-specific future content; the results are reported in Appendix D.

4.2 Planning Performance

Table 1 compares DriveWeave with representative planners and world-action models. Coupled Denoising reaches a PDMS of 91.3, exceeding the two ordered training variants by 4.7 and 4.5 points. It improves all five components over both variants, with NC of 99.0, TTC of 96.1, and EP of 85.4, while retaining 100.0 comfort. The two ordered variants finish at similar scores despite their different learning orders and stage allocations. With 40K total updates in each setting, learning both objectives from the outset gives the stronger result. DriveWeave also exceeds the strongest listed learned baseline by 1.2 PDMS points. Sensor configurations are listed alongside the scores.

Qualitative results. Figure 5 pairs full-horizon trajectories with future visual predictions decoded after one denoising update across stationary, straight-driving, and turning scenes. In the second row, the predicted trajectory makes a small lateral adjustment near the parked vehicle on the right. In the third row, the predicted turn follows a path closer to the inside of the bend. The examples show how trajectory shape changes with the scene, while the multi-view predictions retain the surrounding road layout and camera viewpoints.

Table 2: Information exchange (IE) versus dynamic alignment (DA). DA enables both consistency training (CT) and sampling correction (SC). Codec, visual targets, and two-step budget are fixed.

IE	**DA**	**PDMS** ↑	E_{pose} ↓	**ms**
Off	Off	86.3	1.42	40.8
On	Off	87.9	0.97	45.1
Off	On	89.4	0.81	60.4
On	On	**91.3**	**0.52**	66.2

Table 3: Effect of visual sampling density under coupled denoising.

Imagination target	Visual timestamps	**PDMS** ↑
None (plan only)	–	85.1
Sparse anchored (ours)	4 s	**91.3**
Dense frame-by-frame	per 0.5 s	90.1

Table 4: Consistency training (CT) and sampling correction (SC) with information exchange enabled. All configurations use two denoising steps.

Configuration	CT	SC	PDMS ↑	E_{pose} ↓	Latency (ms)
A	Off	Off	87.9	0.97	45.1
B	Off	On	89.1	0.78	62.3
C	On	Off	90.0	0.72	47.6
D	On	On	**91.3**	**0.52**	66.2

4.3 Ablation Studies

Information exchange and dynamic alignment. Table 2 separates information exchange from dynamic alignment, with DA enabling both consistency training and sampling correction. IE alone raises PDMS from 86.3 to 87.9 and reduces E_{pose} from 1.42 to 0.97. DA alone reaches 89.4 PDMS and 0.81 pose disagreement. Combining them gives 91.3 PDMS and 0.52 pose disagreement at 66.2 ms, compared with 40.8 ms for the model without either mechanism. DA improves planning even when the attention path is blocked, showing that the geometric coupling contributes beyond feature access. Adding IE further improves the result: richer cross-modal context and a shared physical constraint serve complementary roles.

Consistency training and sampling correction. Table 4 separates consistency training from sampling correction. Without correction, training with consistency improves PDMS from 87.9 to 90.0 and reduces E_{pose} from 0.97 to 0.72 (A versus C). With correction, it improves PDMS from 89.1 to 91.3 and reduces disagreement from 0.78 to 0.52 at comparable latency (B versus D). Sampling correction adds 1.2 and 1.3 PDMS points to the two models, respectively. The C–A comparison locates part of the benefit in the learned velocity field, while D–B shows that this benefit remains when both models receive sampling correction. The gains therefore support using the same consistency objective during learning and generation. C also outperforms B at lower latency (47.6 versus 62.3 ms), while D improves further, supporting learned coordination with sample-specific correction.

Visual target density. Table 3 fixes the four-second planning horizon and varies the visual targets. A single terminal anchor achieves 91.3 PDMS, compared with 85.1 without future visual generation and 90.1 with half-second visual targets. Both visual-generation settings improve on the plan-only model. Sparse anchoring further gains 1.2 PDMS points over dense prediction while reducing inference latency from 340.2 to 66.2 ms. The gain over plan-only prediction shows the value of visual imagination; the gain over dense prediction shows that its value does not simply increase with the number of generated frames. In this setting, a terminal visual target provides a more effective planning signal than a dense rollout.

5 Conclusion

DriveWeave uses future imagination to inform planning through repeated feature exchange and pose-consistent denoising. A shared geometric relation gives heterogeneous predictions a common objective, shaping both the learned velocity field and its reverse-time updates. Sparse visual anchors retain this connection without requiring dense future-frame prediction. The 91.3 PDMS result and mechanism ablations support a central finding: useful visual imagination depends on how it is coupled to planning, as well as how much future imagery is generated. Richer scene-level constraints offer a natural extension beyond the ego-motion relation studied here.

AI USE STATEMENT

Generative AI tools assisted with developing and refining the methodological formulation, discussing experimental designs and interpreting results, drafting text. The authors conducted the experiments, checked the resulting formulations and manuscript.

ETHICS STATEMENT

This work studies trajectory planning using existing driving data and non-reactive simulation in NAVSIM. Generated futures and planned trajectories may contain errors, and benchmark performance does not establish safety for deployment on public roads. Real-world use requires further validation under diverse traffic conditions and appropriate safety safeguards.

REPRODUCIBILITY STATEMENT

Section 3 defines the joint representation, training objectives, and reverse-time updates. Section 4 reports the evaluation protocol, training budgets, and controlled comparisons. Supplementary Sections A–D provide the pose conventions, attention masks, stage-specific consistency terms, gradient derivations, fixed consistency settings, and metric definitions.

REFERENCES

B. Zhang, N. Song, J. Li, X. Zhu, J. Deng, and L. Zhang, "Future-aware end-to-end driving: Bidirectional modeling of trajectory planning and scene evolution," in *NeurIPS*, vol. 38, 2025.

T. Xia *et al.*, "DriveLaW: Unifying planning and video generation in a latent driving world," in *CVPR*, 2026, pp. 39701–39712.

Y. Chen, Y. Wang, and Z. Zhang, "DrivingGPT: Unifying driving world modeling and planning with multi-modal autoregressive transformers," in *ICCV*, 2025, pp. 26890–26900.

J. Ho, T. Salimans, A. Gritsenko, W. Chan, M. Norouzi, and D. J. Fleet, "Video diffusion models," 2022, arXiv:2204.03458.

J. Yang *et al.*, "Generalized predictive model for autonomous driving," in *CVPR*, 2024, pp. 14662–14672.

A. Hu, L. Russell, H. Yeo, Z. Murez, G. Fedoseev, A. Kendall, J. Shotton, and G. Corrado, "GAIA-1: A generative world model for autonomous driving," 2023, arXiv:2309.17080.

C. Chi, S. Feng, Y. Du, Z. Xu, E. Cousineau, B. C. M. Burchfiel, and S. Song, "Diffusion Policy: Visuomotor policy learning via action diffusion," in *Robotics: Science and Systems (RSS)*, 2023.

M. Janner, Y. Du, J. B. Tenenbaum, and S. Levine, "Planning with diffusion for flexible behavior synthesis," in *ICML*, 2022, pp. 9902–9915.

A. Ajay, Y. Du, A. Gupta, J. B. Tenenbaum, T. S. Jaakkola, and P. Agrawal, "Is conditional generative modeling all you need for decision making?" in *ICLR*, 2023.

Y. Wang, J. He, L. Fan, H. Li, Y. Chen, and Z. Zhang, "Driving into the future: Multiview visual forecasting and planning with world model for autonomous driving," in *CVPR*, 2024, pp. 14749–14759.

S. Gao, J. Yang, L. Chen, K. Chitta, Y. Qiu, A. Geiger, J. Zhang, and H. Li, "Vista: A generalizable driving world model with high fidelity and versatile controllability," in *NeurIPS*, vol. 37, 2024.

A. Rahimi, V. Gerard, E. Zablocki, M. Cord, and A. Alahi, "MAD: Motion appearance decoupling for efficient driving world models," in *CVPR*, 2026, pp. 18364–18374.

W. Peebles and S. Xie, "Scalable diffusion models with transformers," in *ICCV*, 2023, pp. 4195–4205.

Y. Lipman, R. T. Q. Chen, H. Ben-Hamu, M. Nickel, and M. Le, "Flow matching for generative modeling," in *ICLR*, 2023.

K. Zhang *et al.*, "Epona: Autoregressive diffusion world model for autonomous driving," in *ICCV*, 2025, pp. 27220–27230.

Y. Hu *et al.*, "Planning-oriented autonomous driving," in *CVPR*, 2023, pp. 17853–17862.

X. Weng, B. Ivanovic, Y. Wang, Y. Wang, and M. Pavone, "PARA-Drive: Parallelized architecture for real-time autonomous driving," in *CVPR*, 2024.

K. Chitta, A. Prakash, B. Jaeger, Z. Yang, K. Renz, and A. Geiger, "Transfuser: Imitation with transformer-based sensor fusion for autonomous driving," *IEEE TPAMI*, vol. 45, no. 11, pp. 12878–12895, 2023.

S. Hu, L. Chen, P. Wu, H. Li, J. Yan, and D. Tao, "ST-P3: End-to-end vision-based autonomous driving via spatial-temporal feature learning," in *ECCV*, 2022, pp. 533–549.

B. Jiang *et al.*, "VAD: Vectorized scene representation for efficient autonomous driving," in *ICCV*, 2023, pp. 8340–8350.

S. Doll, N. Hanselmann, L. Schneider, R. Schulz, M. Cordts, M. Enzweiler, and H. P. A. Lensch, "DualAD: Disentangling the dynamic and static world for end-to-end driving," in *CVPR*, 2024, pp. 14728–14737.

P. Wu, X. Jia, L. Chen, J. Yan, H. Li, and Y. Qiao, "Trajectory-guided control prediction for end-to-end autonomous driving: A simple yet strong baseline," in *NeurIPS*, vol. 35, 2022.

Z. Li, Z. Yu, S. Lan, J. Li, J. Kautz, T. Lu, and J. M. Alvarez, "Is ego status all you need for open-loop end-to-end autonomous driving?" in *CVPR*, 2024, pp. 14864–14873.

W. Sun, X. Lin, Y. Shi, C. Zhang, H. Wu, and S. Zheng, "SparseDrive: End-to-end autonomous driving via sparse scene representation," in *ICRA*, 2025, pp. 8795–8801.

D. Dauner *et al.*, "NAVSIM: Data-driven non-reactive autonomous vehicle simulation and benchmarking," in *NeurIPS Datasets and Benchmarks*, 2024.

Z. Li *et al.*, "Hydra-MDP: End-to-end multimodal planning with multi-target Hydra-distillation," 2024, arXiv:2406.06978.

S. Shang, Y. Chen, Y. Wang, Y. Li, and Z.-X. Zhang, "DriveDPO: Policy learning via safety DPO for end-to-end autonomous driving," in *NeurIPS*, vol. 38, 2025.

K. Li, Z. Li, S. Lan, Y. Xie, Z. Zhang, J. Liu, Z. Wu, Z. Yu, and J. M. Alvarez, "Hydra-MDP++: Advancing end-to-end driving via expert-guided Hydra-distillation," 2025, arXiv:2503.12820.

B. Jiang, S. Chen, H. Gao, B. Liao, Q. Zhang, W. Liu, and X. Wang, "VADv2: End-to-end vectorized autonomous driving via probabilistic planning," in *ICLR*, 2026.

C. Jiang, A. Cornman, C. Park, B. Sapp, Y. Zhou, and D. Anguelov, "MotionDiffuser: Controllable multi-agent motion prediction using diffusion," in *CVPR*, 2023, pp. 9644–9653.

B. Liao *et al.*, "DiffusionDrive: Truncated diffusion model for end-to-end autonomous driving," in *CVPR*, 2025.

D. Ha and J. Schmidhuber, "Recurrent world models facilitate policy evolution," in *NeurIPS*, 2018, pp. 2451–2463.

D. Hafner, T. Lillicrap, J. Ba, and M. Norouzi, "Dream to control: Learning behaviors by latent imagination," in *ICLR*, 2020.

D. Hafner, J. Pasukonis, J. Ba, and T. Lillicrap, "Mastering diverse control tasks through world models," *Nature*, vol. 640, pp. 647–653, 2025.

T. Brooks *et al.*, "Video generation models as world simulators," OpenAI, Tech. Rep., 2024.

NVIDIA, "Cosmos world foundation model platform for physical AI," 2025, arXiv:2501.03575. [Online]. Available: https://arxiv.org/abs/2501.03575

A. Hu *et al.*, "Model-based imitation learning for urban driving," in *NeurIPS*, vol. 35, 2022.

C. Shi, J. Xu, S. Shi, K. Sheng, B. Zhang, and L. Jiang, "DriveWAM: Video generative priors enable scalable world-action modeling for autonomous driving," 2026, arXiv:2605.28544.

C. Min *et al.*, "DriveWorld: 4D pre-trained scene understanding via world models for autonomous driving," in *CVPR*, 2024, pp. 15522–15533.

Y. Yang, J. Mei, Y. Ma, S. Du, W. Chen, Y. Qian, Y. Feng, and Y. Liu, "Driving in the occupancy world: Vision-centric 4D occupancy forecasting and planning via world models for autonomous driving," 2024, arXiv:2408.14197.

X. Gui, M. Zhang, T. Yan, W. Han, J. Gong, F. Tan, C.-Z. Xu, and J. Shen, "Bridging scene generation and planning: Driving with world model via unifying vision and motion representation," 2026, arXiv:2603.14948.

F. Jia, L. Liu, Z. Song, C. Jia, H. Ye, X. Hao, and L. Chen, "DriveWorld-VLA: Unified latent-space world modeling with vision-language-action for autonomous driving," 2026, arXiv:2602.06521.

Y. Li *et al.*, "DriveVLA-W0: World models amplify data scaling law in autonomous driving," in *ICLR*, 2026.

D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," 2013, arXiv:1312.6114. [Online]. Available: https://arxiv.org/abs/1312.6114

R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *CVPR*, 2022, pp. 10684–10695.

J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *NeurIPS*, 2020.

J. Yang, K. Chitta, S. Gao, L. Chen, Y. Shao, X. Jia, H. Li, A. Geiger, X. Yue, and L. Chen, "ReSim: Reliable world simulation for autonomous driving," in *NeurIPS*, vol. 38, 2025.

A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, "A kernel two-sample test," *JMLR*, vol. 13, pp. 723–773, 2012.

Y. Li, L. Fan, J. He, Y. Wang, Y. Chen, Z. Zhang, and T. Tan, "Enhancing end-to-end autonomous driving with latent world model," in *ICLR*, 2025.

Y. Li, Y. Wang, Y. Liu, J. He, L. Fan, and Z. Zhang, "End-to-end driving with online trajectory evaluation via BEV world model," in *ICCV*, 2025, pp. 27137–27146.

Y. Li *et al.*, "ReCogDrive: A reinforced cognitive framework for end-to-end autonomous driving," in *ICLR*, 2026.

Z. Zhou, T. Cai, S. Z. Zhao, Y. Zhang, Z. Huang, B. Zhou, and J. Ma, "AutoVLA: A vision-language-action model for end-to-end autonomous driving with adaptive reasoning and reinforcement fine-tuning," in *NeurIPS*, 2025.

Y. Zhou *et al.*, "DriveDreamer-Policy: A geometry-grounded world-action model for unified generation and planning," 2026, arXiv:2604.01765.

J. Ho and T. Salimans, "Classifier-free diffusion guidance," 2022, arXiv:2207.12598.

I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *ICLR*, 2019.

R. Feng, C. Yu, W. Deng, P. Hu, and T. Wu, "On the guidance of flow matching," in *ICML*, 2025. arXiv:2502.02150.