REWAM: A UNIFIED FRAMEWORK FOR REPRESENTATION-CENTRIC ANALYSIS OF WORLD-ACTION MODELING

Anonymous authors
Paper under double-blind review

ABSTRACT

World-action models (WAMs) learn future scene dynamics alongside actions for autonomous driving, yet how world representations affect planning remains unclear. We propose ReWAM, a unified framework for studying how world representations shape planning. Under a common architecture, ReWAM instantiates inverse dynamics model, world model, and world-action model across nine representations spanning VAE latents, visual foundation features, and driving-specific BEV features, while making even implicit foundation latents visually inspectable as videos. We evaluate these models progressively through action recovery, appearance-agnostic world-action consistency, and planning, revealing aligned representation trends and an advantage for visual foundation representations. Within the same framework, we further compare world-action interaction paradigms and introduce a new method that directs world-prediction supervision to action generation during training while retaining action-only inference. ReWAM achieves strong planning performance, reaching 91.6 PDMS on `navtest` and 39.3 EPDMS on `navhard`. Its efficient variant retains 91.6 PDMS on `navtest` at 59 ms, 3.6× faster than inference with future-world generation.

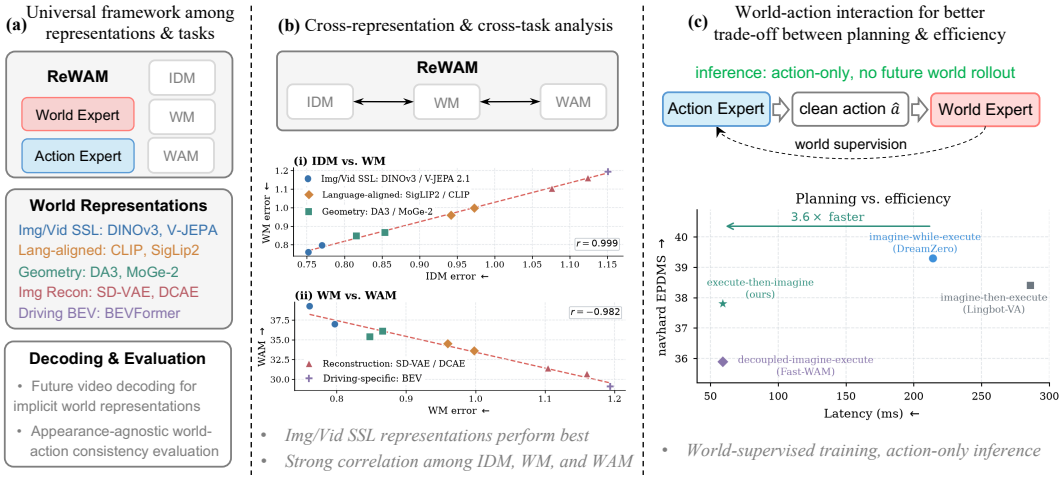

Figure 1: **Overview of ReWAM.** ReWAM provides a unified framework for analyzing how world representations and world-action interaction shape planning. (a) A shared architecture instantiates inverse dynamics model (IDM), world model (WM), and world-action model (WAM) across diverse representations, with video decoding for even implicit latents and appearance-agnostic WM evaluation. (b) Cross-representation analysis reveals aligned trends in action recovery (IDM), world-action consistency (WM), and planning (WAM). (c) Controlled interaction comparisons motivate a new method to direct world supervision to action generation and retain fast action-only inference.

1 INTRODUCTION

World-action models (WAMs) learn future scene evolution together with actions to strengthen autonomous driving (Zhang et al., 2025a; Zhao et al., 2025; Xia et al., 2026; Liu et al., 2026b; Li et al.,

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.

2026b; Jia et al., 2026). Their success raises a central question: *which world representations best support planning, and how should world modeling guide effective and efficient action generation?*

First, representation choice remains tied to individual WAM designs. Video-generative approaches typically inherit VAE latents optimized for reconstruction rather than semantic structure (Zhang et al., 2025a; Xia et al., 2026; Liu et al., 2026b), while implicit feature-predictive approaches use learned visual or BEV states whose prediction is difficult to inspect directly (Li et al., 2025b;c; Wang et al., 2026a). Recent methods enrich world prediction with semantics and geometry (Xu et al., 2026; Fu et al., 2026), but comparisons across architectures leave the effect of representation choice unclear. A common framework is needed to identify representations that best support planning.

Second, world modeling evaluation remains disconnected from planning. Existing studies emphasize downstream planning scores or decoded-video quality (Zhang et al., 2025a; Xia et al., 2026; Liu et al., 2026b). Metrics such as FID and FVD (Heusel et al., 2017; Unterthiner et al., 2018) do not directly measure whether a predicted future follows its action condition, and comparisons across representations also depend on decoder quality. This leaves a fundamental relationship poorly understood: whether action-consistent world prediction translates into stronger planning.

Third, world-action interaction creates a tension between planning performance and inference cost. Generating future worlds before or alongside actions requires expensive iterative prediction (Xia et al., 2026; Ye et al., 2026; Liu et al., 2026b). Decoupled approaches such as Fast-WAM (Yuan et al., 2026) and UNIVERSE (Liu et al., 2026a) retain world supervision during training while enabling action-only inference. However, our controlled comparison shows that fully-decoupled world and action supervision sacrifices planning performance, especially in challenging scenarios. Preserving the benefits of world modeling at low inference cost remains critical for deployment.

Therefore, we present ***Re**presentation-Centric **W**orld-**A**ction **M**odeling* (**ReWAM**) in this paper, a unified framework for studying world representations and action planning. Our contributions are summarized in Figure 1. **(1)** Our central contribution is a common architecture that instantiates inverse dynamics model (IDM), world model (WM), and world-action model (WAM) across low-dimensional reconstruction latents, high-dimensional visual foundation features, and driving-specific BEV features, enabling controlled representation-centric analysis across tasks. It also enables visual inspection for even implicit world representations (Figure 1a). **(2)** We evaluate three capabilities across nine representations: IDM action recovery, WM world-action consistency under an appearance-agnostic protocol, and WAM planning (Figure 1b). The results reveal strongly aligned representation trends: image/video self-supervised representations lead, while foundation representations outperform reconstruction and BEV alternatives, providing empirical guidance for world-representation design. **(3)** A systematic analysis of world-action interaction motivates *execute-then-imagine* (ETI), which directs world-dynamics supervision to action generation during training while retaining action-only inference (Figure 1c). With effective representation selection, ETI reaches 91.6 PDMS on NAVSIM v1, 91.7 EPDMS on NAVSIM v2, and 37.8 EPDMS on `navhard` (Dauner et al., 2024; Cao et al., 2025), while running 3.6× faster than inference with future-world generation.

2 ReWAM: Representation-Centric World-Action Modeling

ReWAM treats visual-encoder latents as world representations in a common architecture for inverse dynamics modeling (IDM), action-conditioned world modeling (WM), and joint world-action modeling (WAM). For representation r, its encoder maps video frames to world tokens f_t, while a shallow MLP maps ego waypoints to action tokens a_t, where $t \leq 0$ denotes history and $t > 0$ the future. Figure 2a–c instantiates the three tasks through their clean conditions, prediction targets, and attention masks. We compare nine representations: image/video self-supervision (DINOv3 and V-JEPA 2.1 (Siméoni et al., 2025; Mur-Labadia et al., 2026)), language alignment (CLIP and SigLIP 2 (Radford et al., 2021; Tschannen et al., 2025)), geometry (DA3 and MoGe-2 (Lin et al., 2026; Wang et al., 2025a)), reconstruction (SD-VAE and DCAE (Rombach et al., 2022; Chen et al., 2025)), and driving-specific BEV from BEVFormer (Li et al., 2022). For a world representation r and $\text{task} \in \{\text{IDM}, \text{WM}, \text{WAM}\}$ (Secs. 2.1–2.3), we denote the corresponding model by $\mathcal{M}_r^{\text{task}}$. All variants of $\mathcal{M}_r^{\text{task}}$ share a Mixture-of-Transformers (MoT) diffusion architecture (Liang et al., 2025b) but are trained independently. Modality-specific world and action experts exchange information through masked joint attention. Condition tokens remain clean, while target tokens are generated through flow matching (Lipman et al., 2023) at denoising times τ_w and τ_a.

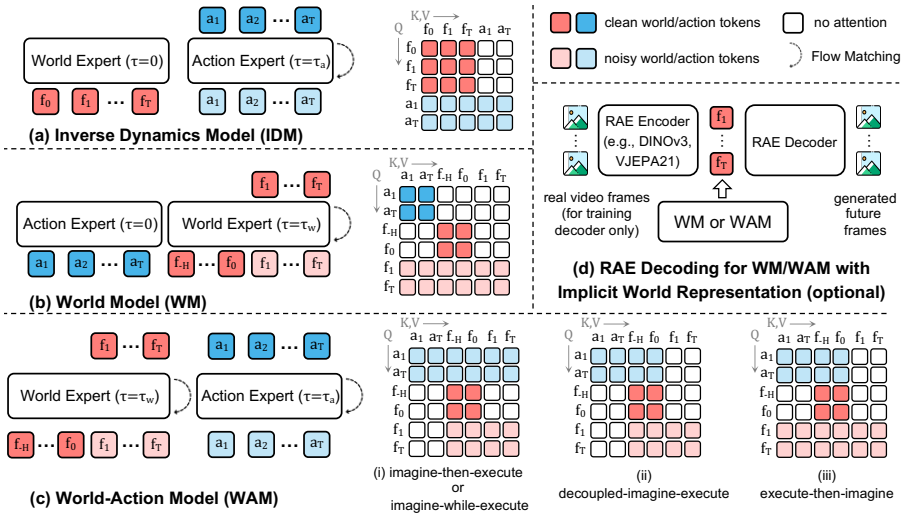

Figure 2: **The ReWAM framework.** Different clean conditions, prediction targets, and attention masks instantiate (a) IDM, (b) action-conditioned WM, and (c) WAM within a common architecture. Historical actions, ego states, and driving commands supplied to the WM and WAM are omitted for clarity. (d) RAE decoder maps implicit world representations to video frames for visual inspection.

To accommodate high-dimensional world latents and low-dimensional actions, we equip MoT with the DDT-inspired wide head from RAE (Wang et al., 2025b; Zheng et al., 2026) (Appendix A.4). For implicit latents such as DINOv3 and V-JEPA 2.1, an optional RAE-style ViT decoder maps predicted world tokens to video frames (Figure 2d), *making implicit world predictions visually inspectable.*

2.1 Inverse Dynamics Model

As shown in Figure 2a, the IDM maps the world tokens of an observed video sequence to the action trajectory underlying that sequence:

$$\hat{a}_{1:T} = \mathcal{M}_r^{\mathrm{IDM}}(f_{0:T}; \epsilon_a), \tag{1}$$

where $f_{0:T} = (f_0, \ldots, f_T)$ are world tokens encoded in representation r, $\hat{a}_{1:T} = (\hat{a}_1, \ldots, \hat{a}_T)$ is the recovered future action trajectory, and ϵ_a initializes action generation with Gaussian noise. Under the mask in Figure 2a, the world stream remains clean ($\tau_w = 0$) and cannot attend to the target actions, whereas the action expert denoises future action tokens at τ_a while attending to the complete world-token sequence. *Recovering actions from video therefore measures how faithfully representation r maps world dynamics into action space, which is essential for planning.*

We train each IDM on the NAVSIM (Dauner et al., 2024) `navtrain` split and evaluate action recovery from video on `navtest`. Table 1 reports average displacement error (ADE), final displacement error (FDE), and their mean (Overall), all measured in meters. Figure 3a visualizes Overall error by representation family, where lower values indicate more accurate action recovery. We can see that reconstruction-oriented SD-VAE and DCAE trail the 2D foundation representations, highlighting the importance of high-level visual knowledge beyond image reconstruction and compression for action alignment. Among foundation representations, self-supervised DINOv3 and V-JEPA 2.1 outperform language-aligned CLIP and SigLIP 2 and geometry-oriented DA3 and MoGe-2. The driving-specific BEV also trails the general-purpose foundation representations. *Overall, image/video self-supervision offers the best action alignment and the strongest recovery capability.*

2.2 World Model

As shown in Figure 2b, the action-conditioned WM maps historical world tokens and a future action trajectory to future world tokens:

$$\hat{f}_{1:T} = \mathcal{M}_r^{\mathrm{WM}}(f_{-H:0}, a_{1:T}; \epsilon_w), \tag{2}$$

where $f_{-H:0}$ provides the observed history, $a_{1:T}$ specifies the future trajectory, and ϵ_w initializes world generation with Gaussian noise. Under the mask in Figure 2b, future world tokens attend

Metric	DINOv3	V-JEPA 2.1	SigLIP2	CLIP	DA3	MoGe-2	SD-VAE	DCAE	BEV
ADE (m) ↓	**0.567**	0.599	0.703	0.750	0.723	0.609	0.942	0.977	0.919
FDE (m) ↓	**0.938**	0.942	1.182	1.196	0.984	1.024	1.211	1.270	1.383
Overall (m) ↓	**0.752**	0.770	0.942	0.973	0.854	0.816	1.076	1.124	1.151

Table 1: **IDM action recovery on NAVSIM `navtest`.**

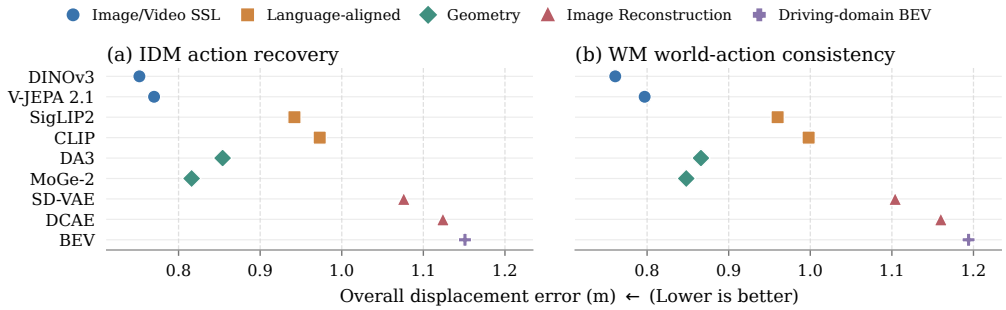

Figure 3: **IDM and WM across representations.** (a) IDM action recovery. (b) appearance-agnostic WM evaluation via world-action consistency. Colors and markers show representation families.

to the historical world stream, clean actions, and themselves, whereas action tokens and historical world tokens cannot attend to future world tokens, avoiding leakage. Thus, action tokens remain clean ($\tau_a = 0$), while the world expert denoises $\hat{f}_{1:T}$ across τ_w. *Predicting future world latents given future actions tests the model's ability to reason about action-induced world dynamics.*

Appearance-agnostic WM evaluation via world-action consistency. Existing WAMs are evaluated mainly through downstream planning (Liu et al., 2026b; Xu et al., 2026) or FID and FVD on decoded videos (Heusel et al., 2017; Unterthiner et al., 2018; Zhang et al., 2025a; Xia et al., 2026). FID and FVD measure visual fidelity and temporal coherence, but do not test whether a predicted future follows the world dynamics induced by actions. They are also confounded by differences in decoder quality across world representations. We therefore evaluate the WM through world-action consistency in predicted latent space. For each representation r, the WM generates $\hat{f}_{1:T}$ conditioned on $a_{1:T}$, and the frozen IDM of the same representation recovers the implied actions:

$$\tilde{a}_{1:T} = \mathcal{M}_r^{\mathrm{IDM}}(f_0, \hat{f}_{1:T}; \epsilon_a). \tag{3}$$

Passing $\hat{f}_{1:T}$ directly to the IDM bypasses video rendering, yielding an appearance-agnostic measure of world-action consistency that jointly reflects WM prediction and IDM action recovery (Figure 4).

We pretrain each WM on the nuPlan (Caesar et al., 2021) and fine-tune it on NAVSIM. We then report world-action consistency as ADE, FDE, and their mean (Overall) between $\tilde{a}_{1:T}$ and $a_{1:T}$ on `navtest` split in Table 2, with Overall visualized in Figure 3b. The results mirror the IDM ordering: self-supervised representations lead language-aligned and geometry-oriented alternatives, while all foundation representations outperform reconstruction representations and driving-specific BEV. *Overall, stronger action recovery across representations is associated with greater world-action consistency in world model predictions.*

2.3 World-Action Model

As shown in Figure 2c, the WAM maps historical world tokens to jointly generate future world and action tokens:

$$(\hat{f}_{1:T}, \hat{a}_{1:T}) = \mathcal{M}_r^{\mathrm{WAM}}(f_{-H:0}; \epsilon_w, \epsilon_a), \tag{4}$$

where $f_{-H:0}$ provides the observed world history, and ϵ_w and ϵ_a denote the initial Gaussian noise for future-world and action generation, respectively. This design follows the premise underlying WAMs that *learning to predict future world evolution improves reasoning about world dynamics and, in turn, provides stronger support for action planning.*

Accordingly, WAMs commonly generate future world latents at inference to guide action generation. In Figure 2c-i, an asymmetric attention mask allows future action tokens to attend to future world tokens while blocking the reverse direction. Although recent methods omit explicit future-world prediction for efficiency (Yuan et al., 2026; Liu et al., 2026a), we use *imagine-while-execute* (IWE),

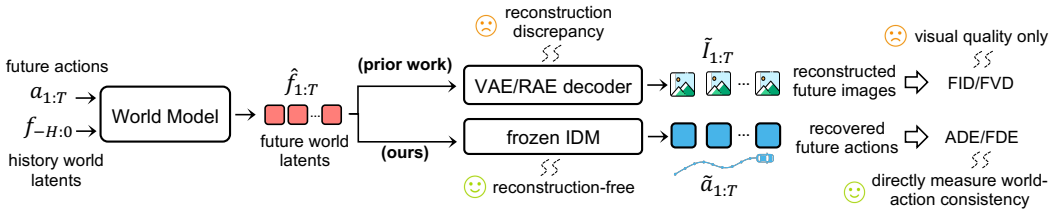

Figure 4: **Appearance-agnostic WM evaluation via world-action consistency.**

Metric	DINOv3	V-JEPA 2.1	SigLIP2	CLIP	DA3	MoGe-2	SD-VAE	DCAE	BEV
ADE (m) ↓	**0.575**	0.669	0.719	0.769	0.734	0.639	0.963	1.001	0.942
FDE (m) ↓	0.947	**0.925**	1.201	1.227	0.998	1.057	1.245	1.319	1.446
Overall (m) ↓	**0.761**	0.797	0.960	0.998	0.866	0.848	1.104	1.160	1.194

Table 2: **Appearance-agnostic WM evaluation on NAVSIM `navtest` via world-action consistency.** Errors are measured between the conditioning and IDM-recovered future action trajectories.

a common design in existing WAMs (Ye et al., 2026; Liu et al., 2026b), as the reference interaction in this section. IWE jointly denoises future world and action tokens during training and inference. At each sampling step, the two experts predict velocity fields at τ_w and τ_a, respectively, and advance both token states toward their clean endpoints. This provides a consistent basis for comparing world representations. Alternative interaction paradigms are studied in Sec. 2.4.

We train the WAM variants on NAVSIM `navtrain`, typically initializing each from the pretrained WM of the same representation. We evaluate them on `navtest` under both v1 and v2 protocols, and on `navhard` (Cao et al., 2025) under v2. Compared with `navtest`, `navhard` adds challenging driving cases and a second stage with perturbed observations. It therefore requires stronger world dynamics reasoning. Table 3 reports the results. Across these evaluations, foundation representations outperform reconstruction-oriented SD-VAE and DCAE as well as driving-specific BEV. Image/video self-supervised representations further lead language-aligned and geometry-oriented representations, especially on `navhard`. *This ordering closely matches the trends in IDM action recovery and WM world-action consistency.* **Unless otherwise specified, we use DINOv3 as the default world representation for ReWAM throughout the remainder of this paper.**

Metric	DINOv3	V-JEPA 2.1	SigLIP2	CLIP	DA3	MoGe-2	SD-VAE	DCAE	BEV
PDMS (v1) ↑	**91.6**	91.2	90.2	89.9	90.3	90.3	88.9	88.4	89.0
EPDMS (v2) ↑	**91.7**	91.5	90.5	90.2	90.9	90.8	89.4	88.9	88.8
EPDMS (navhard) ↑	**39.3**	37.0	34.5	33.6	36.1	35.4	31.4	30.7	29.1

Table 3: **Planning across world representations.** We compare *imagine-while-execute* WAMs on NAVSIM v1 and v2 `navtest`. `navhard` uses the v2 protocol.

To quantify this relationship, Figure 5 reports pairwise Pearson correlations among IDM, WM, and WAM `navhard` metrics across nine representations. All three pairs show strong associations: IDM and WM errors rise together, while both decline as WAM EPDMS improves. *These trends reveal a tight empirical link among IDM action-space alignment, WM world-action consistency, and WAM planning across representations.*

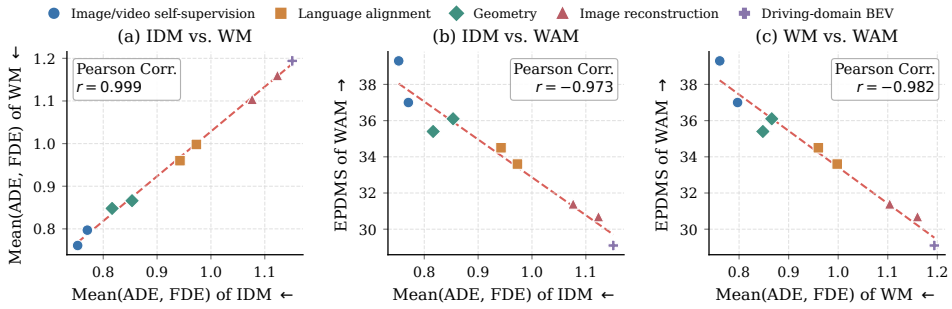

Figure 5: **Cross-representation relationships among IDM, WM, and WAM.** Colors and markers denote representation families, dashed lines show least-squares fits, and inset values report Pearson correlation r. Lower is better for IDM/WM errors, while higher is better for WAM EPDMS.

2.4 Design Choices of World-Action Interaction

As discussed by Yuan et al. (2026), existing WAMs commonly use *imagine-while-execute* (IWE) or *imagine-then-execute* (ITE) interaction methods, as illustrated in Figure 2c-i. Both adopt an asymmetric attention mask: future action tokens attend to future world tokens but not vice versa, allowing predicted dynamics to guide action generation. IWE jointly denoises both streams, so action tokens attend to evolving noisy world states. ITE denoises the world first, then actions against a clean-world key-value (KV) cache. Both require future-world generation at inference, making IWE slow and ITE even slower due to its two-stage world-then-action denoising process.

Several WAM designs retain future-world supervision during training while enabling action-only inference (Yuan et al., 2026; Zhang et al., 2025a; Liu et al., 2026a). In ReWAM, this design corresponds to *decoupled-imagine-execute* (DIE), which mutually masks future world and action tokens such that they interact only through the fixed clean context (Figure 2c-ii). We compare DIE with IWE, ITE, and an *execute-only* (EO) baseline under a shared setup. Table 4 shows that DIE matches the low latency of EO but trails IWE and ITE in planning, especially on `navhard`. We attribute this gap to the bidirectional mask, which removes a direct gradient path between the two streams and limits the effect of world modeling on action planning.

To mitigate this issue, we introduce *execute-then-imagine* (ETI) as shown in Figure 2c-iii. ETI reverses the asymmetric dependency of IWE and ITE: actions are generated without future-world rollout, while future world tokens attend to future actions. *This enables world dynamics supervision to directly shape the action expert while preserving action-only inference.* We use two training variants (Appendix A.6). ETI-sync synchronously denoises both streams, so world tokens attend to noisy action tokens. ETI-cache first rolls out clean actions differentiably, then passes their tokens through the action expert at $\tau_a = 0$ to cache layerwise K/V for world denoising. At inference, both variants plan using only the action rollout. Table 4 reports the results. By removing future-world generation at inference, ETI-sync and ETI-cache run at $\sim$60 ms, matching the latency of DIE and EO while making ETI approximately $3.6\times$ faster than IWE and $4.8\times$ faster than ITE. ETI-cache matches the best result on `navtest` and approaches IWE and ITE on `navhard`, providing a better balance between planning performance and inference latency than DIE. ETI-sync offers only a modest gain over DIE, likely because sparse, noisy actions provide weak conditions for world denoising during synchronous training. ETI-cache instead uses clean self-rolled-out actions, providing a stable information path from world supervision to planning.

Metric	IWE	ITE	EO	DIE	ETI-sync	ETI-cache
Future world generation at inference	Yes	Yes	No	No	No	No
Latency (ms) ↓	214ms	286ms	**59ms**	**59ms**	60ms	**59ms**
PDMS (v1) ↑	**91.6**	91.5	88.4	90.9	91.2	**91.6**
EPDMS (v2) ↑	**91.7**	91.5	88.7	91.2	91.3	**91.7**
EPDMS (navhard) ↑	**39.3**	38.4	30.6	35.9	36.2	37.8

Table 4: **World-action interaction comparison.** We report planning performance and inference latency. NAVSIM v1 and v2 use `navtest`. `navhard` uses the v2 protocol.

3 Comparison with State-of-the-Art Methods

3.1 Planning Performance

Tables 5 and 6 compare ReWAM with representative methods on NAVSIM v1 `navtest` and NAVSIM v2 `navhard`, respectively. Table 5 covers traditional end-to-end and world-action methods, while Table 6 reports WAM model size, latency (see Appendix A.1), and two-stage EPDMS. ReWAM achieves the best overall planning performance on both benchmarks, outperforming recent WAMs including DriveVA (Liu et al., 2026b), DriveWorld-VLA (Jia et al., 2026), UNIVERSE (Liu et al., 2026a), DriveVLA-W0 (Li et al., 2026b), DriveLaW (Xia et al., 2026), and EponaV2 (Xu et al., 2026). It also surpasses methods using multi-camera and LiDAR inputs (Liao et al., 2025; Li et al., 2025c) or simulator-based reinforcement fine-tuning (Li et al., 2026c;a). On `navhard`, this advantage persists under both real observations and perturbed synthetic follow-up observations. ETI-cache preserves strong planning with action-only inference and the lowest measured latency, providing the best balance between planning performance and efficiency.

Method	Sensors	NC	DAC	TTC	C	EP	PDMS
Traditional end-to-end methods							
TransFuser (Chitta et al., 2023)	3C+L	97.7	92.8	92.8	**100.0**	79.2	84.0
DiffusionDrive (Liao et al., 2025)	3C+L	98.2	96.2	94.7	**100.0**	82.2	88.1
ReCogDrive (RFT) (Li et al., 2026c)	3C	97.9	97.3	94.9	**100.0**	**87.3**	90.8
SGDrive (RFT) (Li et al., 2026a)	Front C	98.6	97.8	96.2	**100.0**	85.8	91.1
World-action model methods							
LAW (Li et al., 2025b)	Front C	96.4	95.4	88.7	99.9	81.7	84.6
Epona (Zhang et al., 2025a)	Front C	97.9	95.1	93.8	99.9	80.4	86.2
WoTE (Li et al., 2025c)	3C+L	98.5	96.8	94.9	99.9	81.9	88.3
DriveVLA-W0 (Li et al., 2026b)	Front C	98.4	95.3	95.2	**100.0**	80.9	87.2
PWM (Zhao et al., 2025)	Front C	98.6	95.9	95.4	**100.0**	81.8	88.1
DriveLaW (Xia et al., 2026)	Front C	99.0	97.1	96.7	**100.0**	81.3	89.1
CoWorld-VLA (Huang et al., 2026)	Front C	99.1	97.0	96.5	**100.0**	84.0	90.0
DriveVA (Liu et al., 2026b)	Front C	**99.2**	97.5	**98.7**	**100.0**	83.5	90.9
DriveWorld-VLA (Jia et al., 2026)	3C	99.1	98.2	96.1	**100.0**	85.9	91.3
UNIVERSE (Liu et al., 2026a)	Front C	99.1	97.6	98.5	**100.0**	83.6	91.0
ReWAM (IWE)	Front C	**99.2**	**98.6**	97.4	**100.0**	84.7	**91.6**
ReWAM (ETI-cache)	Front C	**99.2**	**98.6**	96.8	**100.0**	85.2	**91.6**

Table 5: **Comparison on the NAVSIM v1 `navtest` split.** C and L denote camera and LiDAR inputs. RFT denotes reinforcement fine-tuning.

Method	Params	Latency	Stage	NC	DAC	DDC	TLC	EP	TTC	LK	HC	EC	EPDMS
DriveVLA-W0 (Li et al., 2026b)	9.3B	1418ms	S1	96.8	83.3	99.0	99.6	**84.6**	95.3	96.4	97.6	78.2	24.4
			S2	76.8	64.3	79.9	98.3	**89.2**	75.0	46.8	95.8	53.1	
Epona (Zhang et al., 2025a)	2.5B	1055ms	S1	94.7	79.3	96.9	99.3	83.5	93.8	93.3	97.6	57.3	25.0
			S2	80.5	69.2	82.8	98.8	85.5	77.4	46.5	**96.7**	47.4	
DriveLaW (Xia et al., 2026)	8.1B	1547ms	S1	97.3	89.1	99.2	99.6	84.3	97.1	96.2	**97.8**	67.6	30.6
			S2	82.5	67.6	83.5	98.1	84.8	78.5	45.8	96.4	57.3	
DriveVA (Liu et al., 2026b)	11.4B	476ms	S1	97.6	92.4	99.4	99.6	82.7	96.9	97.8	97.6	36.0	30.5
			S2	81.6	74.7	85.5	98.1	87.8	78.4	49.4	94.8	16.3	
EponaV2 (Xu et al., 2026)	6.7B	–	S1	97.3	90.7	99.4	**100.0**	83.3	**97.3**	97.3	97.6	60.9	36.1
			S2	83.6	78.0	**88.0**	**98.9**	86.0	80.3	50.1	96.1	52.0	
ReWAM (IWE)	**698M**	214ms	S1	**98.3**	94.7	99.6	99.8	83.4	97.1	98.2	**97.8**	**80.0**	**39.3**
			S2	**84.2**	**78.1**	86.5	98.6	84.6	**81.0**	**52.5**	96.1	**70.8**	
ReWAM (ETI-cache)	**698M**	**59ms**	S1	97.2	**94.9**	**99.9**	99.6	84.0	**97.3**	**98.7**	**97.8**	78.7	37.8
			S2	81.9	77.8	87.0	98.0	86.4	78.3	51.5	96.3	66.8	

Table 6: **Comparison on NAVSIM v2 `navhard`.** S1 and S2 denote real-observation and synthetic follow-up stages, respectively. EPDMS is the final two-stage score. Latency is unavailable for EponaV2 due to the absence of open-source code implementation.

Our method is further evaluated on NAVSIM v2 and achieves 91.7 EPDMS (Appendix C.1).

3.2 Qualitative World-Action Consistency

Figure 6 compares Epona (Zhang et al., 2025a), DriveLaW (Xia et al., 2026), DriveVA (Liu et al., 2026b), and ReWAM on the same driving scene. Projecting each model's predicted action trajectory onto the current frame allows direct comparison between its planned path and the ego motion implied by its generated future frames. ReWAM follows the ground-truth path closely, and its visual rollout precisely depicts motion along that path. The other WAMs show larger trajectory deviations or weaker correspondence between their planned paths and generated futures. This combination of accurate planning and consistent world prediction illustrates how ReWAM integrates the two tasks. *Additional visualizations can be found in our Appendix C.2 and supplementary videos.*

4 Discussions

How well does ReWAM transfer to unseen domains? Table 7 evaluates zero-shot transfer from NAVSIM to nuScenes (Caesar et al., 2020) and Bench2Drive (CARLA) (Jia et al., 2024), without



Figure 6: **World-action consistency across WAMs.** Predicted (solid) and ground-truth (dashed) trajectories are projected onto the current frame and shown in BEV alongside generated future frames.

target-domain fine-tuning (see Appendix A.1). ReWAM outperforms prior WAM methods in both average L2 error and collision rate on each domain, with particularly large margins on nuScenes. Notably, our ReWAM model is trained from scratch only on driving data, while DriveVA (Liu et al., 2026b) builds on the web-scale pretrained Wan 2.2 (Wan Team et al., 2025). Together with our representation comparisons, these results suggest that informative world representations support planning that generalizes across driving domains. *Qualitative results can be found in our Appendix C.2.*

	nuScenes								Bench2Drive (CARLA)							
	L2 (m) ↓				Collision (%) ↓				L2 (m) ↓				Collision (%) ↓			
Method	1s	2s	3s	Avg.	1s	2s	3s	Avg.	1s	2s	3s	Avg.	1s	2s	3s	Avg.
DriveVLA-W0	0.43	1.26	2.60	1.43	0.22	0.66	1.42	0.77	1.01	2.77	5.22	3.00	1.49	2.53	3.53	2.52
PWM	2.06	3.91	6.00	3.99	0.12	0.15	0.86	0.36	1.70	2.74	3.97	2.80	4.01	3.73	3.53	3.76
DriveVA	0.33	0.76	1.43	0.84	0.00	0.07	0.12	0.06	0.69	1.29	**2.03**	1.33	1.38	1.97	2.65	2.00
ReWAM	**0.12**	**0.32**	**0.64**	**0.36**	**0.00**	**0.00**	**0.04**	**0.01**	**0.55**	**1.23**	2.12	**1.30**	**0.95**	**1.56**	**2.42**	**1.64**

Table 7: **Zero-shot planning on nuScenes and Bench2Drive (CARLA).** All methods are world action models trained only on NAVSIM without target-domain fine-tuning while DriveVA (Liu et al., 2026b) is initialized from web-scale pretrained Wan 2.2. Baseline values are adopted from DriveVA.



Figure 7: **Long-horizon world model generation.** Autoregressive video rollouts from 1 s to 1 min given the same observed history. Our ReWAM decodes predicted DINOv3 latents into video frames.

Can implicit world representations support long-horizon simulation? Figure 7 compares one-minute action-conditioned autoregressive rollouts, far beyond our model's 4-second training horizon. Our ReWAM world model renders implicit world latents as visually inspectable video and

maintains scene consistency and visual stability, with its current world latent predictions fed to itself as the subsequent context. In contrast, the reconstruction-oriented autoencoder representation baselines Epona (Zhang et al., 2025a), DriveLaW (Xia et al., 2026), and DriveVA (Liu et al., 2026b) show drift or lose scene structure over time. This result points to foundation representations such as DINOv3 Siméoni et al. (2025) as a promising basis for long-horizon driving simulators.

Note: *We strongly encourage readers to refer to our Appendix and supplementary videos for additional experiments and qualitative results.*

5 Related Work

Here, we focus on driving world models and joint world-action modeling, with a broader discussion of representations, generative architectures, and evaluation in Appendix D.

World Models for Autonomous Driving. Driving world models learn how traffic scenes evolve, ranging from controllable video generation for simulation (Hu et al., 2023; Wang et al., 2024a) to visual and structured forecasting for planning (Wang et al., 2024b; Zheng et al., 2024; Yang et al., 2025b; Li et al., 2025c). World Action Models (WAMs) take a further step by learning future scenes alongside trajectories, allowing dense temporal supervision to shape action prediction (Zhang et al., 2025a; Liu et al., 2026b). These approaches usually inherit the representation chosen for their particular architecture. Video WAMs use reconstruction-oriented autoencoder latents (Zhang et al., 2025a; Liu et al., 2026b), whereas other planners predict implicit features or driving-specific occupancy and BEV states (Li et al., 2025b; Zheng et al., 2024; Yang et al., 2025b; Li et al., 2025c). Reconstruction latents prioritize appearance, implicit features are difficult to inspect, and structured states encode task-specific spatial abstractions. Consequently, cross-system comparisons conflate representation with model design, leaving its effect on world modeling and planning unresolved.

Joint World-Action Modeling. In robotics, GR-1 and GR-2 jointly predict future images alongside actions (Wu et al., 2024a; Cheang et al., 2024), UWM couples video and action diffusion (Zhu et al., 2025), and DreamZero and Cosmos Policy adapt pretrained video generators for control (Ye et al., 2026; Kim et al., 2026). Across embodied and driving models, a key difference is whether an imagined future enters action generation at inference. In our terminology, imagine-then-execute (ITE) generates a future before predicting actions (Du et al., 2023; Xia et al., 2026). Imagine-while-execute (IWE) instead jointly denoises both streams, keeping imagined dynamics on the inference path (Ye et al., 2026; Liu et al., 2026b). Decoupled-imagine-execute (DIE) retains future-world supervision in training but predicts actions without an imagined future at inference (Yuan et al., 2026; Liu et al., 2026a). ReWAM compares these interactions under a shared architecture, and propose a new method to further improve the balance between planning performance and inference efficiency.

6 Conclusion

We introduce ReWAM, a unified framework that brings low-dimensional reconstruction latents, high-dimensional visual foundation features, and driving-specific BEV features into a common architecture for inverse dynamics model (IDM), world model (WM), and world-action model (WAM). This shared design enables controlled representation-centric comparisons of IDM action recovery, WM world-action consistency, and WAM planning. Across nine representations, the three tasks exhibit aligned trends and close empirical connections: image/video self-supervised representations perform best, while foundation representations consistently outperform reconstruction-oriented and BEV alternatives. ReWAM further makes implicit foundation-space predictions visually inspectable, providing a direct view of the dynamics modeled in these spaces. Beyond representation choice, we systematically compare world-action interaction paradigms and examine how world prediction affects both action generation and inference cost. This analysis motivates *execute-then-imagine* (ETI), which directs world-dynamics supervision to action generation during training while retaining action-only inference. Overall, the DINOv3 configuration performs best across the three tasks. Its WAM reaches 91.6 PDMS on NAVSIM v1, 91.7 EPDMS on NAVSIM v2, and 39.3 EPDMS on `navhard`. The ETI variant reduces inference latency to 59 ms while retaining strong planning performance across these benchmarks. These results establish ReWAM as a general foundation and a strong candidate for future research on efficient, high-performing world-action models.

AI USE STATEMENT

We used an AI assistant primarily to polish the manuscript and, for selected sections, to draft English prose from section and subsection outlines written by the authors. It also assisted with literature lookup, occasional feedback on experimental analysis, and a small number of auxiliary testing and experimental scripts. The authors reviewed and revised the AI-assisted text, executed the experiments, and checked the code and results before including them in the paper. The authors take full responsibility for the final content.

ETHICS STATEMENT

This work studies autonomous-driving planning using recorded driving data and simulation-based benchmarks. Improved planning may benefit road safety, while errors in unfamiliar or challenging scenarios could affect driving decisions. Further safety evaluation would be needed before real-world use.

REPRODUCIBILITY STATEMENT

Section 2 describes the models and world-action interaction designs, and Section 3 presents the evaluation benchmarks and protocols. Appendix A details the data processing, representation extraction, architecture, training configurations, and sampling procedures. Appendices B and C provide additional experiments and results. We will release the code implementation upon acceptance.

REFERENCES

Hidehisa Arai, Keishi Ishihara, Tsubasa Takahashi, and Yu Yamaguchi. ACT-Bench: Towards action controllable world models for autonomous driving. In *International Conference on Learning Representations*, 2025.

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Mojtaba Komeili, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, Sergio Arnaud, Abha Gejji, Ada Martin, Francois Robert Hogan, Daniel Dugas, Piotr Bojanowski, Vasil Khalidov, Patrick Labatut, Francisco Massa, Marc Szafraniec, Kapil Krishnakumar, Yong Li, Xiaodong Ma, Sarath Chandar, Franziska Meier, Yann LeCun, Michael Rabbat, and Nicolas Ballas. V-JEPA 2: Self-supervised video models enable understanding, prediction and planning. arXiv preprint arXiv:2506.09985, 2025.

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 et al. 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 *Proceedings of The 9th Conference on Robot Learning*, pp. 4709–4722, 2025.

Chi-Lam Cheang, Guangzeng Chen, Ya Jing, Tao Kong, Hang Li, Yifeng Li, Yuxiao Liu, Hongtao Wu, Jiafeng Xu, Yichu Yang, Hanbo Zhang, and Minzhao Zhu. GR-2: A generative video-language-action model with web-scale knowledge for robot manipulation. arXiv preprint arXiv:2410.06158, 2024.

Junyu Chen, Han Cai, Junsong Chen, Enze Xie, Shang Yang, Haotian Tang, Muyang Li, and Song Han. Deep compression autoencoder for efficient high-resolution diffusion models. In *International Conference on Learning Representations*, 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, 2023. doi: 10.1109/TPAMI.2022.3200245.

Daniel Dauner, Marcel Hallgarten, Tianyu Li, Xinshuo Weng, Zhiyu Huang, Zetong Yang, Hongyang Li, Igor Gilitschenski, Boris Ivanovic, Marco Pavone, Andreas Geiger, and Kashyap Chitta. NAVSIM: Data-driven non-reactive autonomous vehicle simulation and benchmarking. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2024.

Yilun Du, Mengjiao Yang, Bo Dai, Hanjun Dai, Ofir Nachum, Joshua B. Tenenbaum, Dale Schuurmans, and Pieter Abbeel. Learning universal policies via text-guided video generation. In *Advances in Neural Information Processing Systems*, 2023.

Jiacheng Fu, Yibo Yuan, Meng Tian, Yue Li, Jiangtong Zhu, Jianhua Han, Yueyi Zhang, Jianwu Fang, Jianru Xue, Hang Xu, and Zhiwei Xiong. 4D-WAM: 4d consistent world modeling for autonomous driving. arXiv preprint arXiv:2608.10107, 2026.

Shenyuan Gao, Jiazhi Yang, Li Chen, Kashyap Chitta, Yihang Qiu, Andreas Geiger, Jun Zhang, and Hongyang Li. Vista: A generalizable driving world model with high fidelity and versatile controllability. In *Advances in Neural Information Processing Systems*, 2024.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, 2020.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models. In *Advances in Neural Information Processing Systems*, 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.

Anthony Hu, Lloyd Russell, Hudson Yeo, Zak Murez, George Fedoseev, Alex Kendall, Jamie Shotton, and Gianluca Corrado. GAIA-1: A generative world model for autonomous driving. arXiv preprint arXiv:2309.17080, 2023.

Xiaotao Hu, Mingkai Jia, Xiaoyang Guo, Qian Zhang, Xiao xiao Long, and Wei Yin. DrivingWorld: Constructing world model for autonomous driving via video GPT. In *International Conference on Pattern Recognition*, pp. 276–291, 2026. doi: 10.1007/978-3-032-31583-0_19.

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.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21807–21818, 2024.

Feiyang Jia, Lin Liu, Ziying Song, Caiyan Jia, Hangjun Ye, Xiaoshuai Hao, and Long Chen. DriveWorld-VLA: Unified latent-space world modeling with vision-language-action for autonomous driving. arXiv preprint arXiv:2602.06521, 2026.

Xiaosong Jia, Zhenjie Yang, Qifeng Li, Zhiyuan Zhang, and Junchi Yan. Bench2Drive: Towards multi-ability benchmarking of closed-loop end-to-end autonomous driving. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Moo Jin Kim, Yihuai Gao, Tsung-Yi Lin, Yen-Chen Lin, Yunhao Ge, Grace Lam, Percy Liang, Shuran Song, Ming-Yu Liu, Chelsea Finn, and Jinwei Gu. Cosmos Policy: Fine-tuning video models for visuomotor control and planning. In *International Conference on Learning Representations*, 2026.

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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4032–4042, 2026a.

Shuang Li, Yihuai Gao, Dorsa Sadigh, and Shuran Song. Unified video action model. arXiv preprint arXiv:2503.00200, 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*, 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 27137–27146, 2025c. doi: 10.1109/ICCV51701.2025.02519.

Yingyan Li, Shuyao Shang, Weisong Liu, Bing Zhan, Haochen Wang, Yuqi Wang, Yuntao Chen, Xiaoman Wang, Yasong An, Chufeng Tang, Lu Hou, Lue Fan, and Zhaoxiang Zhang. DriveVLA-W0: World models amplify data scaling law in autonomous driving. In *International Conference on Learning Representations*, 2026b.

Yongkang Li, Kaixin Xiong, Xiangyu Guo, Fang Li, Sixu Yan, Gangwei Xu, Lijun Zhou, Long Chen, Haiyang Sun, Bing Wang, Kun Ma, Guang Chen, Hangjun Ye, Wenyu Liu, and Xinggang Wang. ReCogDrive: A reinforced cognitive framework for end-to-end autonomous driving. In *International Conference on Learning Representations*, 2026c.

Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Yu Qiao, and Jifeng Dai. BEVFormer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In *European Conference on Computer Vision*, 2022. doi: 10.1007/978-3-031-20077-9_1.

Junbang Liang, Pavel Tokmakov, Ruoshi Liu, Sruthi Sudhakar, Paarth Shah, Rares Ambrus, and Carl Vondrick. Video generators are robot policies. arXiv preprint arXiv:2508.00795, 2025a.

Weixin Liang, Lili Yu, Liang Luo, Srini Iyer, Ning Dong, Chunting Zhou, Gargi Ghosh, Mike Lewis, Wen tau Yih, Luke Zettlemoyer, and Xi Victoria Lin. Mixture-of-transformers: A sparse and scalable architecture for multi-modal foundation models. *Transactions on Machine Learning Research*, 2025b.

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

Haotong Lin, Sili Chen, Jun Hao Liew, Donny Y. Chen, Zhenyu Li, Yang Zhao, Sida Peng, Hengkai Guo, Xiaowei Zhou, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. In *International Conference on Learning Representations*, 2026.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023.

Mengmeng Liu, Diankun Zhang, Jiuming Liu, Jianfeng Cui, Hongwei Xie, Guang Chen, Hangjun Ye, Francesco Nex, Hao Cheng, and Michael Ying Yang. UNIVERSE: Unified video action models for autonomous driving with flexible mask-modulated modality generation. arXiv preprint arXiv:2607.05133, 2026a.

Mengmeng Liu, Diankun Zhang, Jiuming Liu, Jianfeng Cui, Hongwei Xie, Guang Chen, Hangjun Ye, Michael Ying Yang, Francesco Nex, and Hao Cheng. DriveVA: Video action models are zero-shot drivers. In *European Conference on Computer Vision*, 2026b.

Lorenzo Mur-Labadia, Matthew Muckley, Amir Bar, Mido Assran, Koustuv Sinha, Mike Rabbat, Yann LeCun, Nicolas Ballas, and Adrien Bardes. V-JEPA 2.1: Unlocking dense features in video self-supervised learning. arXiv preprint arXiv:2603.14482, 2026.

Jingcheng Ni, Yuxin Guo, Yichen Liu, Rui Chen, Lewei Lu, and Zehuan Wu. MaskGWM: A generalizable driving world model with video mask reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22381–22391, 2025.

Maxime Oquab et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

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.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 8748–8763, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.

Oriane Siméoni et al. DINOv3. arXiv preprint arXiv:2508.10104, 2025.

Jaskirat Singh, Boyang Zheng, Zongze Wu, Richard Zhang, Eli Shechtman, and Saining Xie. Improved baselines with representation autoencoders. arXiv preprint arXiv:2605.18324, 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. arXiv preprint arXiv:2603.29163, 2026.

Shengbang Tong, Boyang Zheng, Ziteng Wang, Bingda Tang, Nanye Ma, Ellis Brown, Jihan Yang, Rob Fergus, Yann LeCun, and Saining Xie. Scaling text-to-image diffusion transformers with representation autoencoders. arXiv preprint arXiv:2601.16208, 2026.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier H'enaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. arXiv preprint arXiv:2502.14786, 2025.

Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Rapha"el Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. arXiv preprint arXiv:1812.01717, 2018.

Wan Team et al. Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314, 2025.

Linbo Wang, Yupeng Zheng, Qiang Chen, Shiwei Li, Yichen Zhang, Zebin Xing, Qichao Zhang, Xiang Li, Deheng Qian, Pengxuan Yang, Yihang Dong, Ce Hao, Xiaoqing Ye, Junyu Han, Yifeng Pan, and Dongbin Zhao. 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.

Ruicheng Wang, Sicheng Xu, Yue Dong, Yu Deng, Jianfeng Xiang, Zelong Lv, Guangzhong Sun, Xin Tong, and Jiaolong Yang. MoGe-2: Accurate monocular geometry with metric scale and sharp details. In *Advances in Neural Information Processing Systems*, volume 38, 2025a.

Shuai Wang, Zhi Tian, Weilin Huang, and Limin Wang. DDT: Decoupled diffusion transformer. arXiv preprint arXiv:2504.05741, 2025b.

Xiaofeng Wang, Zheng Zhu, Guan Huang, Xinze Chen, Jiagang Zhu, and Jiwen Lu. DriveDreamer: Towards real-world-driven world models for autonomous driving. In *European Conference on Computer Vision*, 2024a.

Yuqi Wang, Jiawei He, Lue Fan, Hongxin Li, Yuntao Chen, and Zhaoxiang Zhang. Driving into the future: Multiview visual forecasting and planning with world model for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024b.

Ge Wu, Shen Zhang, Ruijing Shi, Shanghua Gao, Zhenyuan Chen, Lei Wang, Zhaowei Chen, Hongcheng Gao, Yao Tang, Jian Yang, Ming-Ming Cheng, and Xiang Li. Representation entanglement for generation: Training diffusion transformers is much easier than you think. arXiv preprint arXiv:2507.01467, 2025.

Hongtao Wu, Ya Jing, Chilam Cheang, Guangzeng Chen, Jiafeng Xu, Xinghang Li, Minghuan Liu, Hang Li, and Tao Kong. Unleashing large-scale video generative pre-training for visual robot manipulation. In *International Conference on Learning Representations*, 2024a.

Jialong Wu, Shaofeng Yin, Ningya Feng, Xu He, Dong Li, Jianye Hao, and Mingsheng Long. iVideoGPT: Interactive VideoGPTs are scalable world models. In *Advances in Neural Information Processing Systems*, 2024b.

Tianze Xia, Yongkang Li, Lijun Zhou, Jingfeng Yao, Kaixin Xiong, Haiyang Sun, Bing Wang, Kun Ma, Guang Chen, Hangjun Ye, Wenyu Liu, and Xinggang Wang. 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, 2026.

Jiawei Xu, Zhizhou Zhong, Zhijian Shu, Mingkai Jia, Mingxiao Li, Jia-Wang Bian, Qian Zhang, Kaicheng Zhang, Jin Xie, Jian Yang, and Wei Yin. EponaV2: Driving world model with comprehensive future reasoning. arXiv preprint arXiv:2605.14696, 2026.

Mingwang Xu, Jiahao Cui, Feipeng Cai, Hanlin Shang, Zhihao Zhu, Shan Luan, Yifang Xu, Neng Zhang, Yaoyi Li, Jia Cai, and Siyu Zhu. WAM-Diff: A masked diffusion VLA framework with MoE and online reinforcement learning for autonomous driving. arXiv preprint arXiv:2512.11872, 2025.

Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas. VideoGPT: Video generation using VQ-VAE and transformers. arXiv preprint arXiv:2104.10157, 2021.

Pengxuan Yang, Yupeng Zheng, Deheng Qian, Zebin Xing, Qichao Zhang, Linbo Wang, Yichen Zhang, Shaoyu Guo, Zhongpu Xia, Qiang Chen, Junyu Han, Lingyun Xu, Yifeng Pan, and Dongbin Zhao. DreamerAD: Efficient reinforcement learning via latent world model for autonomous driving. arXiv preprint arXiv:2603.24587, 2026.

Xuemeng Yang, Licheng Wen, Tiantian Wei, Yukai Ma, Jianbiao Mei, Xin Li, Wenjie Lei, Daocheng Fu, Pinlong Cai, Min Dou, Liang He, Yong Liu, Botian Shi, and Yu Qiao. DriveArena: A closed-loop generative simulation platform for autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 26933–26943, 2025a.

Yu Yang, Jianbiao Mei, Yukai Ma, Siliang Du, Wenqing Chen, Yijie Qian, Yuxiang Feng, and Yong Liu. Driving in the occupancy world: Vision-centric 4d occupancy forecasting and planning via world models for autonomous driving. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025b.

Jingfeng Yao, Bin Yang, and Xinggang Wang. Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15703–15712, 2025.

Ziyang Yao, Haochen Liu, Yuncheng Jiang, Zeyu Zhu, Zibin Guo, Jingru Wang, Tianle Liu, Jianwei Cui, Kuiyuan Yang, Hongwei Xie, Jingwei Zhao, Guang Chen, and Hangjun Ye. DiscreteWAM: Unified discrete vision-action token editing for world-policy learning. arXiv preprint arXiv:2606.05645, 2026.

Seonghyeon Ye et al. World action models are zero-shot policies. arXiv preprint arXiv:2602.15922, 2026.

Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. In *International Conference on Learning Representations*, 2025.

Tianyuan Yuan, Zibin Dong, Yicheng Liu, and Hang Zhao. Fast-WAM: Do world action models need test-time future imagination? arXiv preprint arXiv:2603.16666, 2026.

Kaiwen Zhang, Zhenyu Tang, Xiaotao Hu, Xingang Pan, Xiaoyang Guo, Yuan Liu, Jingwei Huang, Li Yuan, Qian Zhang, Xiao-Xiao Long, Xun Cao, and Wei Yin. Epona: Autoregressive diffusion world model for autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 27220–27230, 2025a.

Xiangdong Zhang, Jiaqi Liao, Shaofeng Zhang, Fanqing Meng, Xiangpeng Wan, Junchi Yan, and Yu Cheng. VideoREPA: Learning physics for video generation through relational alignment with foundation models. In *Advances in Neural Information Processing Systems*, 2025b.

Zhida Zhao, Talas Fu, Yifan Wang, Lijun Wang, and Huchuan Lu. From forecasting to planning: Policy world model for collaborative state-action prediction. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Boyang Zheng, Nanye Ma, Shengbang Tong, and Saining Xie. Diffusion transformers with representation autoencoders. In *International Conference on Learning Representations*, 2026.

Wenzhao Zheng, Weiliang Chen, Yuanhui Huang, Borui Zhang, Yueqi Duan, and Jiwen Lu. OccWorld: Learning a 3d occupancy world model for autonomous driving. In *European Conference on Computer Vision*, 2024.

Yupeng Zheng, Pengxuan Yang, Zebin Xing, Qichao Zhang, Yuhang Zheng, Yinfeng Gao, Pengfei Li, Teng Zhang, Zhongpu Xia, Peng Jia, XianPeng Lang, and Dongbin Zhao. World4Drive: End-to-end autonomous driving via intention-aware physical latent world model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 28632–28642, 2025.

Yang Zhou, Xiaofeng Wang, Hao Shao, Letian Wang, Guosheng Zhao, Jiangnan Shao, Jiagang Zhu, Tingdong Yu, Zheng Zhu, Guan Huang, and Steven L. Waslander. DriveDreamer-Policy: A geometry-grounded world-action model for unified generation and planning. arXiv preprint arXiv:2604.01765, 2026.

Chuning Zhu, Raymond Yu, Siyuan Feng, Benjamin Burchfiel, Paarth Shah, and Abhishek Gupta. Unified world models: Coupling video and action diffusion for pretraining on large robotic datasets. In *Robotics: Science and Systems*, 2025.

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 Implementation Details

A.1 Evaluation Details

IDM action recovery and WM world-action consistency are measured on `navtest`. WAM planning is scored on `navtest` under NAVSIM v1 and v2, and on `navhard` under the v2 two-stage protocol. For appearance-agnostic WM evaluation, the frozen IDM for the same representation reads predicted latents directly, without rendering and re-encoding images.

Latency is measured for the complete planning path, including the visual encoder, with batch size 1 in bfloat16 on a single NVIDIA A100 GPU. It excludes the RAE decoder because video decoding is not used for planning. Latencies for other methods are obtained by running their official code under the same hardware, batch size, and precision. Parameter counts include the visual encoder and all modules used by the planning path, but exclude the RAE decoder. For zero-shot transfer to nuScenes and Bench2Drive, we follow the official DriveVA implementation and evaluation settings without modification (Liu et al., 2026b), with no target-domain fine-tuning.

A.2 Data and Task Protocols

All three tasks sample video at 2 Hz, using the front camera (`CAM_F0`) for image-based representations and all eight cameras for BEV. Images are undistorted and resized before encoding, with sequence-consistent color jitter and occasional Gaussian blur applied during training. We represent each future action as (x, y, ψ) relative to the last observed ego pose, rather than as a displacement from the preceding waypoint. Predicted trajectories are converted to the NAVSIM scoring format for evaluation. Table 8 summarizes how training data, model sequence lengths, and conditioning differ across tasks. IDM recovers actions from video alone, using the last historical frame as visual context. WM and WAM additionally receive the historical action trajectory, current ego state (v_x, v_y, a_x, a_y), and driving command as clean context.

Task / stage	Training data	History / Future	Additional clean context
IDM	NAVSIM `navtrain`	1 / 8 frames	None
WM pretrain	nuPlan train	4 / 8 frames	actions, ego state, command
WM fine-tune	NAVSIM `navtrain`	4 / 8 frames	actions, ego state, command
WAM	NAVSIM `navtrain`	4 / 8 frames	historical actions, ego state, command

Table 8: **Data protocols for the three tasks.**

A.3 World Representations

To compare world representations under a common modeling setup, we use frozen pretrained visual encoders and keep the MoT configuration fixed across the IDM, WM, and WAM variants, changing only the encoder and representation-specific tokenization (Table 9). For foundation representations, we follow RAEv2 (Singh et al., 2026): each ViT encoder uses the same seven transformer layers and decoder design. Their latents are normalized using encoder statistics computed on nuPlan. The corresponding image RAE decoders each have 28 transformer blocks of width 1152 with 16 heads and are used only for visualizing implicit world latents. SD-VAE and DCAE use final encoder latents, as in common latent video generators. The BEV variant retains BEVFormer's native eight-camera input and final BEV grid. Because the grid is not aligned with pixels, the image-space RAE decoder cannot render BEV predictions as RGB video. A dedicated BEV-to-video decoder remains future work.

Encoder-specific processing. Image encoders use 272×480-pixel inputs, except patch-14 models, which use dimensions divisible by 14 (MoGe-2 and DA3 use 238×420 pixels). V-JEPA 2.1 encodes four-frame clips with temporal tubelets of size two, so each latent token spans two consecutive frames. We design an untubelet decoder to linearly expand each tubelet token from C to $2C$, reshape the two channel groups into temporally ordered per-frame tokens, and add learned frame embeddings before a shared 2D ViT decoder for per-frame reconstruction. SD-VAE groups 2×2 latent cells into each world token, while DCAE pads its input to 288×480 pixels so both dimensions are divisible by 32 as in its official implementation. DA3 receives no camera poses. History and future are encoded

Representation	Model	Used layers	Camera/Frame	$C \times H \times W$	Normalization
DINOv3	ViT-L/16	11,13,15,17,19,21,23	1C/1F	$1024 \times 17 \times 30$	encoder stats
V-JEPA 2.1	ViT-L/16	11,13,15,17,19,21,23	1C/4F	$1024 \times 17 \times 30$	encoder stats
CLIP	ViT-L/14	11,13,15,17,19,21,23	1C/1F	$1024 \times 17 \times 30$	encoder stats
SigLIP 2	ViT-L/16	11,13,15,17,19,21,23	1C/1F	$1024 \times 17 \times 30$	encoder stats
DA3	ViT-L/14	11,13,15,17,19,21,23	1C/4F or 1C/9F	$1024 \times 17 \times 30$	encoder stats
MoGe-2	ViT-L/14	11,13,15,17,19,21,23	1C/1F	$1024 \times 17 \times 30$	encoder stats
SD-VAE	KL-f8	final encoder latents	1C/1F	$16 \times 17 \times 30$	fixed scale
DCAE	f32-c32	final encoder latents	1C/1F	$32 \times 9 \times 15$	fixed scale
BEVFormer	R50-base	final BEV features	8C/1F	$256 \times 32 \times 40$	LN w/o affine

Table 9: **Processing configurations for different world representations.**

separately, with the first historical frame also included in the future clip as a common reconstruction reference. Reference tokens are excluded from future-world modeling. The BEV encoder retains its pretrained camera resolution. We crop its features to the near-ego region, spanning 0–40 m ahead and 16 m to either side. Output shapes and normalization are summarized in Table 9.

A.4 MoT–DDT Architecture

High-dimensional latents from visual foundation models make it costly to widen every layer of a Diffusion Transformer. For image generation, RAE addresses this with a standard-width DiT backbone and a shallow, wider DDT head (Wang et al., 2025b; Zheng et al., 2026). We adopt this depth-width split design for all representations. Notably, both the original DDT and RAE's wide-head variant model a single image-latent stream. To model world and action together, we extend the deep encoder and shallow decoder into parallel MoT experts (Liang et al., 2025b). Each expert retains its own embeddings, attention projections, feed-forward layers, and output head. Within each block, world and action queries, keys, and values are projected to a common width and concatenated for masked attention, then returned to their respective streams. The encoder processes clean context alongside noisy future tokens. Its future-token states condition the wider decoders, where noisy world and action tokens interact again before separate heads predict future latents and trajectories. Table 10 gives the shared dimensions. The same backbone yields IDM, WM, and WAM through task-specific tokens, masks, and losses (Secs. 2.1–2.3).

Stage	World Expert	Action Expert	World-Action Exchange
Input	world latents $\to$ 1024-D	actions $\to$ 512-D	separate embeddings
Encoder	10×1024	10×512	joint masked attention at 1024-D
Bridge	world latents $\to$ 1536	actions $\to$ 1024	separate projections
Decoder	2×1536	2×1024	joint masked attention at 1536-D
Output	representation-latent head	3-D action head	separate predictions

Table 10: **Shared MoT–DDT architecture.** World input and output projections adapt to each representation's latent dimension, while the experts exchange information at the listed attention widths.

Action tokens use one-dimensional rotary position embeddings (1D RoPE) to encode temporal order. Trajectory coordinates are normalized as $\bar{x} = 2x/(50\,\mathrm{m}) - 1$, $\bar{y} = y/(10\,\mathrm{m})$, and $\bar{\psi} = \psi/(1.2\,\mathrm{rad})$, and predictions are mapped back to physical units to obtain executable driving waypoints. Ego state is standardized with training-set statistics, clipped to $[-5, 5]$, and projected into a token. Driving commands and ego state serve as clean action-stream context rather than prediction targets.

A.5 Training and Sampling

Flow matching and sampling. The world and action experts predict clean future latents and normalized trajectories, respectively. During WAM training, we sample $u_w, u_a \sim \mathcal{U}(0, 1)$ independently and apply the same time shift to obtain $\tau_j = 5u_j/(1 + 4u_j)$ for $j \in \{w, a\}$. Let x_w denote future world latents and x_a normalized future actions. With independent Gaussian noise ϵ_j, the noisy input is $x_{j,\tau_j} = (1 - \tau_j)x_j + \tau_j\epsilon_j$. Following Singh et al. (2026), we reparameterize each x-prediction $\hat{x}_j$ to a velocity and compute its flow-matching loss as

$$\hat{v}_j = \frac{x_{j,\tau_j} - \hat{x}_j}{\max(\tau_j, 0.05)}, \qquad v_j^* = \frac{x_{j,\tau_j} - x_j}{\max(\tau_j, 0.05)}, \qquad \mathcal{L}_j = \mathrm{MSE}(\hat{v}_j, v_j^*). \tag{5}$$

The denominator is floored at 0.05 during training. The three task objectives are $\mathcal{L}_{\text{IDM}} = \mathcal{L}_a$, $\mathcal{L}_{\text{WM}} = \mathcal{L}_w + \mathcal{L}_w^{\text{base}}$, and $\mathcal{L}_{\text{WAM}} = \mathcal{L}_w + 5\mathcal{L}_a$. Here $\mathcal{L}_w^{\text{base}}$ applies the same velocity MSE to an auxiliary WM prediction from the encoder following Singh et al. (2026). IDM keeps the world context clean ($\tau_w = 0$) and samples τ_a with the shifted uniform rule. On the contrary, WM keeps actions clean ($\tau_a = 0$) and draws τ_w from a shifted uniform distribution. At WAM inference, world and action generation each use Euler sampling from Gaussian noise by 3 denoising steps.

For the RAE decoders, we follow the reconstruction objective of RAE (Zheng et al., 2026): $\mathcal{L}_{\text{RAE}} = \ell_1 + \text{LPIPS} + 0.75\lambda_{\text{adv}}\mathcal{L}_{\text{GAN}}$, where λ_{adv} adaptively balances the adversarial term.

Training schedules and optimization. Table 11 summarizes the training recipes. We train RAE decoders on nuPlan before training the diffusion models. IDM is trained on NAVSIM `navtrain`. WM is pretrained on nuPlan and fine-tuned on `navtrain`, and WAM is initialized from the corresponding pretrained WM before its `navtrain` training. Within each task, the MoT training schedule is shared across representations.

Stage	Training data	Steps	Batch/GPU	Peak LR	Initialization
RAE decoder	nuPlan train	120k	32	1×10^{-4}	from scratch
IDM	NAVSIM `navtrain`	30k	16	2×10^{-4}	from scratch
WM pretrain	nuPlan train	60k	16	2×10^{-4}	from scratch
WM fine-tune	NAVSIM `navtrain`	8k	16	5×10^{-5}	nuPlan WM
WAM	NAVSIM `navtrain`	30k	16	2×10^{-4}	NAVSIM WM

Table 11: **Training schedules across representations.** Initializations from WM use the matching representation. Batch sizes are per GPU, and all stages use 8 NVIDIA A100 GPUs.

RAE decoders and their discriminators use AdamW with $(\beta_1, \beta_2) = (0.9, 0.95)$ and zero weight decay. Decoder learning rates follow cosine decay after an 8k-step warmup, with EMA decay 0.9978. The discriminator peak learning rate is 2×10^{-4}. IDM, WM, and WAM use Muon with momentum 0.95 and Nesterov updates, while remaining parameters use AdamW with the same betas. These models use zero weight decay, linear learning-rate decay, and EMA decay 0.9995. Their warmup lasts 10k steps, except for WM fine-tuning, which uses 1k steps. All stages use bfloat16 and gradient-norm clipping at 1.0.

A.6 Execute-then-Imagine Variants

Figure 8 details the two ETI training variants. Both use the action-first asymmetric mask in Figure 2c-iii: future world tokens attend to future action tokens, while future action tokens cannot attend to future world tokens. Action prediction is therefore independent of future-world rollout, yet world-prediction supervision can propagate through the action tokens. Let c denote the shared clean context, and let $\mathcal{A}$ and $\mathcal{W}$ denote the action and world experts, respectively.

ETI-sync. As illustrated in Figure 8a, ETI-sync processes independently noisy world tokens w_{τ_w} and action tokens a_{τ_a} in the same forward pass:

$$\mathbf{a}^{\text{sync}} = \mathcal{A}(a_{\tau_a}, \tau_a; c), \qquad \widehat{w}_0^{\text{sync}} = \mathcal{W}(w_{\tau_w}, \tau_w; c, \mathbf{a}^{\text{sync}}). \tag{6}$$

Here $\mathbf{a}^{\text{sync}}$ denotes the layerwise action tokens exposed to masked attention. Because τ_a and τ_w are sampled independently, world tokens interact with the current noisy action tokens throughout the shared MoT forward pass.

ETI-cache. Figure 8b shows the two-stage forward path of ETI-cache. A differentiable K-step Euler sampler first produces a self-predicted clean trajectory $\widetilde{a}_0$, with ($K = 3$) matching the inference-time sampling schedule. This trajectory is embedded into clean action tokens and passed through the action expert again at $\tau_a = 0$. The resulting layerwise keys and values form a fixed cache $\mathbf{a}^{\text{cache}}$ for world denoising:

$$\begin{aligned} \widetilde{a}_0 &= \mathcal{S}_{\mathcal{A}}(\epsilon_a; c), \\ \mathbf{a}^{\text{cache}} &= \text{KV}[\mathcal{A}(\widetilde{a}_0, 0; c)], \\ \widehat{w}_0^{\text{cache}} &= \mathcal{W}(w_{\tau_w}, \tau_w; c, \mathbf{a}^{\text{cache}}). \end{aligned} \tag{7}$$

Here $\mathcal{S}_{\mathcal{A}}$ denotes action sampling driven by $\mathcal{A}$, and $\text{KV}[\mathcal{A}(\cdot)]$ denotes the per-layer key/value tensors produced while the action expert processes the clean action tokens. ETI-cache retains the standard

flow-matching loss for action prediction. Its additional rollout and cache path are not detached, allowing world-prediction supervision to reach the action expert through the clean action tokens. During planning inference, both ETI variants omit future-world prediction and run only the action path. The clean-action cache is constructed only when ETI-cache performs future-world prediction. We isolate this gradient path in Appendix B.3.

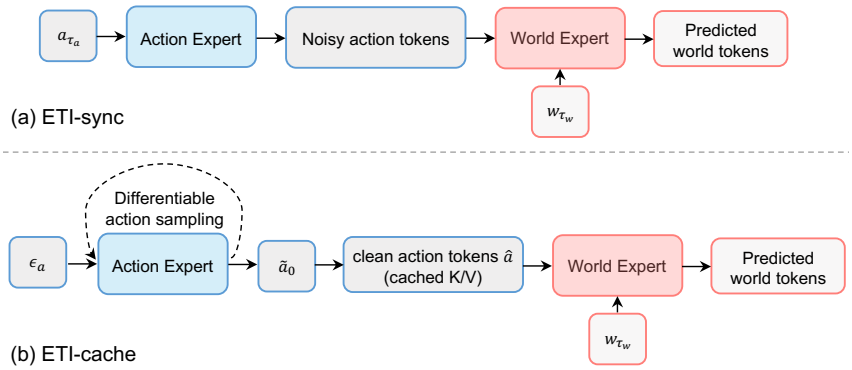

Figure 8: **Training paths of ETI-sync and ETI-cache.** (a) ETI-sync passes noisy action tokens directly to the world expert. (b) ETI-cache rolls out clean action tokens, and caches their layerwise K/V for world denoising. Both variants use action-only planning inference.

B Ablation Studies

We first compare conventional reconstruction and video-generation metrics across world representations. The remaining studies isolate two design choices under the DINOv3 WAM setting. Unless noted, data, encoder, initialization, budget, losses, and evaluation follow Table 11.

B.1 Reconstruction and Video Generation Quality across Representations

Reconstruction quality. We evaluate the encoder–decoder reconstruction of each RGB-decodable representation on 8-second video clips from the nuPlan `val` split using PSNR and LPIPS (Table 12). These metrics quantify information preservation and decoder fidelity, but do not measure whether a latent space aligns with actions or supports planning.

Video generation quality. We further evaluate action-conditioned WM predictions on the same 8-second clips using FID and FVD between decoded predictions and ground-truth videos. FID measures frame-level distributional similarity, while FVD captures spatiotemporal quality. Both metrics depend on representation-specific reconstruction as well as predicted dynamics, and can therefore vary for reasons unrelated to planning. We report them as visual metrics complementary to the appearance-agnostic WM evaluation via world-action consistency in Sec. 2.2.

Evaluation	Metric	DINOv3	V-JEPA 2.1	CLIP	SigLIP 2	DA3	MoGe-2	SD-VAE	DCAE
Reconstruction	PSNR (dB) ↑	28.67	24.84	26.14	26.50	29.74	32.05	30.29	28.17
	LPIPS ↓	0.103	0.198	0.170	0.165	0.128	0.094	0.123	0.161
Generation	FID ↓	2.91	3.79	4.54	4.15	3.94	3.37	6.82	8.00
	FVD ↓	29.31	47.59	40.38	37.71	31.92	23.01	60.37	71.58

Table 12: **Reconstruction and generation quality across world representations on nuPlan `val`.** Higher PSNR is better, while lower LPIPS, FID, and FVD are better. BEV is excluded because it has no RGB decoder.

Table 12 shows no clear generation-quality ordering among foundation representation families, whose reconstruction quality also varies. Yet world models using visual foundation representations achieve substantially lower FID and FVD than the SD-VAE and DCAE variants, despite weaker reconstruction in several cases. Together with the WM world-action consistency results in Sec. 2.2, this highlights the limitations of reconstruction-oriented latents for world-dynamics modeling.

B.2 Standard MoT versus MoT–DDT

Table 13 compares 12-block standard MoT (world/action widths 1152/576, attention width 1152) with the 10+2-block MoT–DDT in Table 10. DINOv3 tokens, objectives, time sampling, optimizer, and training schedule are matched. The MoT–DDT row is the main IWE result. MoT–DDT improves PDMS by 1.8 points and `navhard` EPDMS by 13.3 points with only 25M more parameters and 6 ms more inference latency. This gain is consistent with RAE's finding that high-dimensional representation latents benefit from a shallow, wide denoising head, which expands denoising capacity without widening the entire deep backbone (Zheng et al., 2026).

Backbone	MoT Params (M)	Visual Encoder Params (M)	Latency (ms)	PDMS (v1) ↑	EPDMS (navhard) ↑
Standard MoT	370M	303M	208ms	89.8	26.0
MoT–DDT	395M	303M	214ms	91.6	39.3

Table 13: **MoT backbone ablation.** All non-architectural settings are matched. PDMS is measured on NAVSIM v1 `navtest`. EPDMS is measured on `navhard` under NAVSIM v2.

B.3 World-to-Action Gradient in ETI-cache

ETI-cache is designed to propagate world-prediction supervision into the action expert through its self-rolled clean-action cache. To isolate this path, we detach the fully constructed cache in Eq. 7 before world denoising:

$$\bar{\mathbf{a}}^{\text{cache}} = \text{detach}\left(\mathbf{a}^{\text{cache}}\right), \qquad \hat{w}_0^{\text{SG}} = \mathcal{W}\left(w_{\tau_w}, \tau_w; c, \bar{\mathbf{a}}^{\text{cache}}\right). \tag{8}$$

The two variants therefore have identical action rollouts, cache values, world conditioning, objectives, parameterization, and forward computation. Detaching blocks only the world-prediction gradient into the action expert, while direct action-prediction supervision remains unchanged.

Variant	World-to-Action Gradient	PDMS (v1) ↑	EPDMS (v2) ↑	EPDMS (`navhard`) ↑
ETI-cache	Yes	91.6	91.7	37.8
ETI-cache (stop-grad)	No	91.0	91.2	36.3

Table 14: **Effect of the world-to-action gradient in ETI-cache.** The stop-gradient variant preserves the complete forward computation but blocks gradients through the clean-action cache.

Table 14 shows that detaching the cache consistently reduces planning performance on NAVSIM v1, NAVSIM v2, and `navhard`. Since the two variants differ only in this backward path, the result shows that ETI-cache benefits from directing world-prediction supervision into action generation rather than from its additional rollout and cache computation alone.

B.4 Timestep Coupling and Time Shift

Independent sampling draws separate raw flow times from uniform distribution $u_w, u_a \sim \mathcal{U}(0,1)$ for the world and action experts, while shared sampling uses the same draw for both. Following the time-shift form in RAE (Zheng et al., 2026), we map each draw to $\tau_j = su_j/[1 + (s-1)u_j]$. Thus, $s = 1$ leaves the uniform draw unchanged, while $s = 5$ favors noisier states. Table 15 shows modest performance differences, with independent sampling and $s = 5$ performing best overall across `navtest` and `navhard`.

Time sampling	Shift s	PDMS on `navtest` (%) ↑	EPDMS on `navhard` (%) ↑
Independent	1 (no shift)	91.0	36.9
Independent	5	91.6	39.3
Shared	5	91.6	37.5

Table 15: **Effect of timestep sampling and time shift.** Independent sampling with $s = 5$ is the default IWE configuration.

C Additional Results and Visualizations

C.1 NAVSIM v2 Results

Table 16 compares ReWAM with prior methods on NAVSIM v2 `navtest`. Both ReWAM variants reach 91.7 EPDMS, exceeding the scores of the listed methods. ETI-cache preserves this planning score while running 3.6× faster than IWE at inference (Table 4).

Method	NC	DAC	DDC	TLC	EP	TTC	LK	HC	EC	EPDMS
Traditional end-to-end methods										
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
SparseDriveV2 (Sun et al., 2026)	98.1	98.1	99.6	99.8	**91.1**	97.3	96.9	98.2	78.4	90.1
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
World-action model methods										
DriveWorld-VLA (Jia et al., 2026)	98.6	**99.1**	99.6	99.8	87.4	97.9	97.0	97.8	78.6	86.8
DreamerAD (Yang et al., 2026)	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
Discrete-WAM (Yao et al., 2026)	98.5	98.2	99.7	99.8	90.5	97.9	97.2	**98.3**	78.1	90.4
ReWAM (IWE)	99.1	98.6	**99.8**	**99.9**	87.6	**98.8**	**98.5**	**98.3**	87.1	**91.7**
ReWAM (ETI-cache)	**99.2**	98.6	99.7	**99.9**	87.1	**98.8**	98.3	**98.3**	88.2	**91.7**

Table 16: **Planning comparison on NAVSIM v2 `navtest`.**

C.2 Additional Visualizations

More world-action consistency cases. Figure 9 presents additional challenging cases in the format of Figure 6, with DriveVLA-W0 (Li et al., 2026b) as an additional WAM baseline. Pairing predicted future frames with planned trajectories makes agreement between the generated ego motion, surrounding traffic, and actions directly inspectable. *We strongly recommend watching `consistent_world_action.mp4` in the supplementary materials to better visualize world-action consistency over time.*

Counterfactual reasoning. Figure 10 varies the driving command while holding the current visual context fixed. By alternating among left, right, and straight commands across rollouts, we observe the action trajectories and generated future frames evolve together consistently: the ego motion reflected across successive frames follows the alternative command as well as the corresponding planned path. This command-world-action agreement shows potential for simulating command-conditioned world-action rollouts. *We recommend watching `counterfactual_imagination.mp4` in the supplementary materials for better visualization.*

Zero-shot transfer. Figure 11 complements Table 7 with zero-shot world-action rollouts on Bench2Drive and nuScenes. These results further illustrate ReWAM's strong generalization to unseen driving domains.

Additional video results. The supplementary materials include three videos. `consistent_world_action.mp4` shows additional joint predictions with strong world-action consistency. `counterfactual_imagination.mp4` shows alternative command-conditioned rollouts. `long_horizon.mp4` extends Figure 7 with up to one minute of autoregressive world-video generation compared against prior methods. In this comparison, ReWAM maintains the most stable visual content and the most plausible world dynamics throughout the rollout. *We recommend watching these videos to inspect temporal behavior beyond the still figures.*

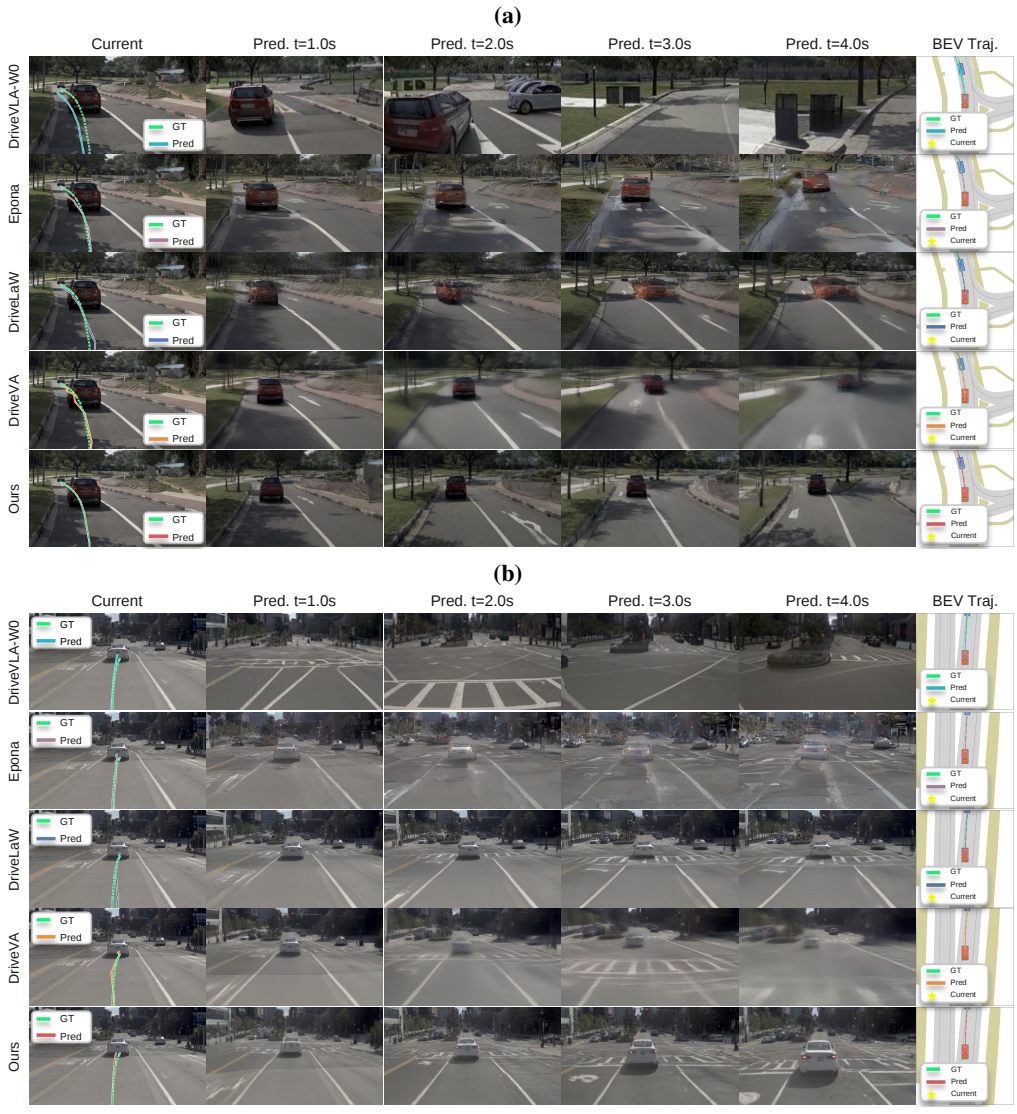

Figure 9: **More world-action consistency cases.** Predicted (solid) and ground-truth (dashed) trajectories are projected onto the current frame and shown in BEV alongside generated future frames.



Figure 10: **Counterfactual world-action rollouts.** Alternative driving commands generate paired action trajectories and future frames from the same visual context.

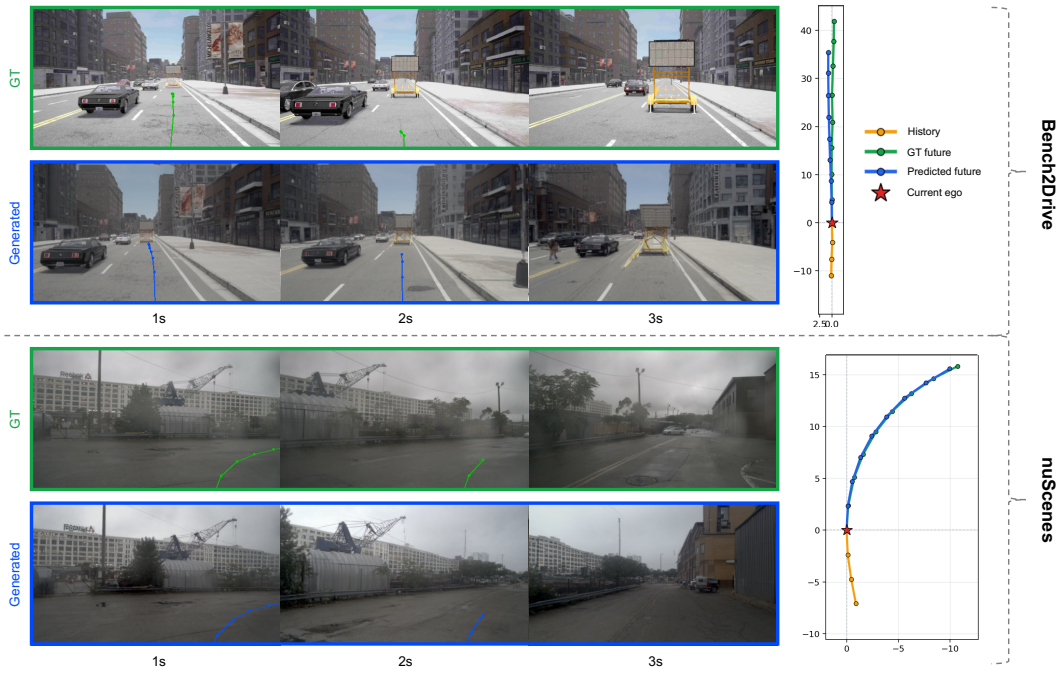

Figure 11: **Zero-shot world-action rollouts.** Ground-truth and generated futures with action trajectories on Bench2Drive (top) and nuScenes (bottom).

D Extended Related Work

Visual Representations for World States. A world state inherits the priorities of its visual encoder. Image self-supervision produces transferable features in DINOv2 and DINOv3 (Oquab et al., 2024; Siméoni et al., 2025), while V-JEPA 2 and 2.1 learn predictive representations from video (Assran et al., 2025; Mur-Labadia et al., 2026). Language-aligned encoders such as CLIP and SigLIP 2 emphasize cross-modal semantics (Radford et al., 2021; Tschannen et al., 2025). Depth Anything 3 and MoGe-2 instead target scene geometry (Lin et al., 2026; Wang et al., 2025a). Reconstruction-oriented autoencoders such as SD-VAE and DCAE prioritize faithful visual compression (Rombach et al., 2022; Chen et al., 2025), whereas BEVFormer constructs a driving-specific bird's-eye-view state from camera observations (Li et al., 2022). These objectives provide distinct candidates for world prediction, motivating a controlled comparison of IDM action recovery, WM world-action consistency, and WAM planning.

High-Level Representations in Generation. Pretrained visual features can enter generative modeling through different paths. VA-VAE aligns its autoencoder latent with a vision foundation model (Yao et al., 2025), while REPA aligns intermediate denoiser features with frozen image representations (Yu et al., 2025). VideoREPA distills relations from a video understanding model into a text-to-video generator (Zhang et al., 2025b), and REG denoises a pretrained class token alongside image latents (Wu et al., 2025). RAEs instead make frozen encoder features the diffusion space and learn a decoder back to images (Zheng et al., 2026; Singh et al., 2026), an approach subsequently scaled to text-to-image generation (Tong et al., 2026). Together, these works establish multiple ways to use high-level features for generation. Our focus is how the choice of world representation affects action-relevant prediction and driving planning.

Generative Architectures. Diffusion models progressed from pixel-space denoising to latent generation and transformer backbones (Ho et al., 2020; Rombach et al., 2022; Peebles & Xie, 2023), and were extended to video (Ho et al., 2022). Flow matching provides a continuous-time training formulation (Lipman et al., 2023). For multimodal modeling, MoT assigns modality-specific transformer parameters while retaining attention across the full sequence (Liang et al., 2025b). DDT separates condition encoding from velocity decoding (Wang et al., 2025b). Drawing on it, RAE attaches a shallow, wide DDT head to a base DiT for high-dimensional representation latents (Zheng et al., 2026). Autoregressive models offer another route: VideoGPT predicts discrete video tokens (Yan et al., 2021), iVideoGPT interleaves visual, action, and reward tokens for interactive prediction (Wu et al., 2024b), and DrivingWorld combines next-state and next-token prediction for driving video (Hu et al., 2026). ReWAM builds on MoT and the RAE-style head to model world and action tokens with different dimensions.

Driving World Models and Planning. Driving world models span simulation and planning. GAIA-1, DriveDreamer, and Vista generate controllable future video (Hu et al., 2023; Wang et al., 2024a; Gao et al., 2024), while DriveArena closes the simulation loop by feeding generated observations to a driving agent (Yang et al., 2025a). For planning, OccWorld and Drive-OccWorld predict occupancy (Zheng et al., 2024; Yang et al., 2025b), whereas Drive-WM evaluates candidate trajectories through generated multiview futures (Wang et al., 2024b). MaskGWM augments video diffusion with masked reconstruction (Ni et al., 2025). Recent latent approaches introduce other priors. World4Drive uses vision-foundation features for trajectory generation and selection (Zheng et al., 2025). Drive-JEPA adapts video-predictive features for a trajectory planner (Wang et al., 2026b), while Latent-WAM distills geometric knowledge into compact autoregressive world states (Wang et al., 2026a). DriveDreamer-Policy jointly generates depth, video, and actions (Zhou et al., 2026). These systems establish several world-state choices, yet differences in architecture and training obscure what the representation itself contributes to prediction and planning.

Embodied World-Action Models. In robotics, UniPi generates a future video plan before extracting actions (Du et al., 2023), and Video Policy trains video and action generation together to recover robot policies from imagined behavior (Liang et al., 2025a). GR-1 predicts future images alongside actions, and GR-2 extends video pretraining to subsequent video-and-action finetuning (Wu et al., 2024a; Cheang et al., 2024). UWM uses separate video and action diffusion times to support forward and inverse dynamics within one model (Zhu et al., 2025). DreamZero jointly models future video

and actions with a pretrained video backbone (Ye et al., 2026). UVA learns a joint video-action latent but decodes the modalities separately, allowing action inference without video generation (Li et al., 2025a), whereas Cosmos Policy encodes actions and future states as latent frames of a video model (Kim et al., 2026). Their different inference paths motivate studying how world prediction communicates with action generation.

Evaluating Action-Relevant World Models. FID and FVD evaluate distributional image and video quality (Heusel et al., 2017; Unterthiner et al., 2018), while VBench probes perceptual and temporal properties of generated videos (Huang et al., 2024). These metrics do not test whether a predicted driving future follows a specified action trajectory. ACT-Bench adds that test by estimating ego motion from generated video (Arai et al., 2025). Its video-based evaluation, however, does not isolate the contribution of the world representation from the rendering pipeline or relate action fidelity to planning. ReWAM compares representations within one architecture through complementary tests: IDM recovers actions from observed world states, WM is evaluated through world-action consistency in predicted latents without decoding, and WAM measures downstream planning.