Drift-Resistant Navigation World Model with Anchored Epipolar Guidance

Po-Chien Luan **Zimin Xia**[†] **Wuyang Li** **Yang Gao** **Alexandre Alahi**[†]
VITA@EPFL
firstname.lastname@epfl.ch [†] Corresponding Author

Project Page: https://vita-epfl.github.io/DR-NWM.io/

Abstract

We propose Drift-Resistant Navigation World Model, a generative model that mitigates both *perceptual drift* and *geometric drift* in conventional rollout-based navigation world models. Existing methods recursively feed generated content into subsequent steps, causing noise accumulation and degraded predictions, *i.e.*, *perceptual drift*. Meanwhile, their predictions often deviate from the agent's motion, resulting in *geometry drift*. We address both types of drift by redesigning world-model prediction as an *anchor-guided rollout*. Instead of rolling out every frame sequentially, we first predict sparse future anchors that serve as stable long-range targets, and then generate intermediate frames within each chunk conditioned on both past context and future anchors. Importantly, these sparse anchors also provide geometric constraints, supported by bidirectional epipolar geometry, to localize where corresponding content should appear in the intermediate frames. Experiments on four benchmarks demonstrate consistent improvements over strong baselines in long-horizon visual quality, geometric consistency, and multi-view coherence. These gains further translate into improved downstream planning performance under the same planners, highlighting the importance of drift-resistant, geometry-aware prediction for reliable navigation world models.

1 Introduction

World models [9] learn the dynamics of the environment, enabling an agent to anticipate future observations before acting [19, 10, 1]. Navigation is a fitting setting for such prediction [29, 4]. Because each action changes the robot's viewpoint, navigation depends on predicting what it will observe after moving. Therefore, a world model provides a natural basis for navigation by forecasting future observations under candidate actions, which can then be used to compare possible trajectories.

Despite their promise, current world models for navigation suffer from two drifts (fig. 1a). The first is ***perceptual drift*** [37, 25, 21, 20]. Standard practice [4] involves forecasting future autoregressively, where previous predictions are recursively fed back as inputs for subsequent steps. Consequently, minor per-step errors propagate and compound over time, leading to perceptual drift over extended horizons [5]. While some methods attempt to mitigate this through inference-like rollout training [37], exposing the model to its own errors in this manner introduces prohibitive computational overhead.

The second is ***geometric drift*** [37, 35, 3]. There are two main reasons for this drift. First, predicted futures may fail to stay aligned with the changes of the agent's actions, even when they remain perceptually plausible. Second, perceptual drift further corrupts geometric consistency, since degraded predictions provide unreliable visual cues for subsequent viewpoint changes. To enforce geometric fidelity, prior approaches typically rely on costly 3D supervision, such as depth maps or point clouds [17, 36, 7, 38]. Such supervision is difficult to acquire at scale and often limits generalizability. Both

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

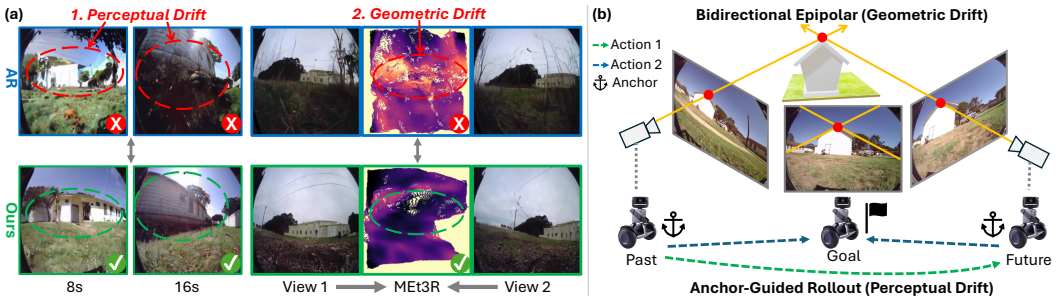


Figure 1: (a) Our method improves perceptual and geometric drifts (left: visual quality in long horizons; right: MEt3R [2] (darker is better)). We use NWM [4] as an example of an autoregressive (AR) process. Red circles denote the drifts. (b) Anchor-guided rollout improves perceptual drift. Through two anchors, we can form bidirectional epipolar constraints to improve geometric drift.

failures are critical bottlenecks for navigation: effective planning requires future simulations that remain stable over long horizons and faithful to the geometry induced by the agent's motion.

To address these challenges, we adopt a strategy mirrored in both classical robotics [18] and human behavior [23]: first planning a distant goal and then determining the steps to reach it. Following this intuition, we decompose future prediction into a two-level hierarchy. Our model first establishes a series of sparse temporal *anchors* that act as long-term goals for the trajectory. Once these stable anchors are fixed, the model tracks the path between them by generating intermediate frames for each chunk. This effectively partitions the full rollout into shorter, manageable chunks, preventing recursive error accumulation and fundamentally mitigating perceptual drift (fig. 1a left).

Furthermore, the anchors provide not only temporal guidance but also a geometric benefit. Once both the past and future anchors are available, they define two views of the same rollout, while the action sequence between them provides a motion prior for the intermediate frames. Based on these two views, we use bidirectional epipolar geometry to localize where corresponding content should appear in the intermediate frames (proof in app. A), as illustrated in fig. 1(b). This gives the model stronger spatial grounding without requiring explicit 3D supervision (fig. 1a right).

We propose **Drift-Resistant Navigation World Model (DR-NWM)**, a generative model that combines *anchor-guided rollout* and *bidirectional epipolar geometry* with a *Anchor-conditioned Diffusion Transformer* (AC-DiT). DR-NWM first predicts sparse anchors at fixed intervals, then generates intermediate frames within each chunk by jointly conditioning on past context and the future anchor. By leveraging bidirectional actions and views, we inject geometry awareness into AC-DiT through bidirectional epipolar masks. As a result, the generated sequences not only remain stable over time but also stay faithful to the geometry induced by the agent's motion.

This paper makes two main contributions. 1) We propose DR-NWM. It jointly mitigates perceptual drift and geometric drift, leading to consistent improvements in long-horizon future prediction and multi-view consistency. These improvements further translate into stronger downstream planning performance. 2) At the core of DR-NWM is anchor-guided rollout, our ingenious design that redefines the role of future anchors: they serve not only as stable long-horizon targets for mitigating recursive drift, but also as key geometric anchors for bidirectional epipolar grounding, enabling geometry-aware prediction without explicit 3D supervision.

2 Related Work

Recent work has explored world models as predictive simulators for navigation, improving the paradigm in several directions. NWM establishes controllable future-view generation for navigation with a Conditional Diffusion Transformer backbone, following a conventional autoregressive rollout pipeline that predicts future observations sequentially from past observations and actions [4]. While effective, this design remains vulnerable to long-horizon error accumulation. MWM improves rollout consistency and deployment efficiency [37], while RAE-NWM studies navigation modeling in dense visual representation space to improve long-horizon behavior and action accuracy [38]. UniWM moves toward a more integrated formulation by unifying visual foresight and planning within

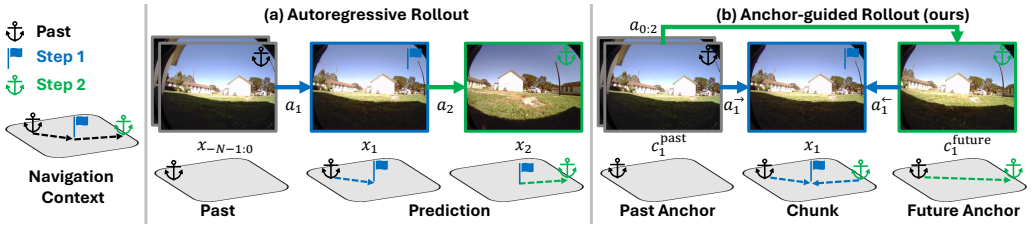


Figure 2: Comparison of inference strategies. (a) Conventional autoregressive rollout predicts future frames step by step, recursively feeding previous predictions into later inputs. (b) Our anchor-based inference first predicts sparse future anchors and then fills in the intermediate frames within each chunk under bidirectional conditions.

a memory-augmented autoregressive architecture [16]. In parallel, EgoWM leverages pretrained internet-scale video priors for controllable egocentric prediction [3], CompACT reduces planning cost through compact latent tokenization [16], and one-step world models improve efficiency by replacing iterative diffusion with one-step generation [30]. Despite these advances, prior work does not directly address the joint problem we consider: reducing long-horizon perceptual drift and improving geometric drift in a monocular and unified setting. Our method instead reformulates future prediction as anchor-guided rollout with bidirectional geometric grounding, which jointly mitigates recursive rollout error and strengthens geometry-aware conditioning without explicit 3D supervision.

3 Method

Conventional world models for navigation typically rely on one-sided autoregressive rollout [4], where previously generated outputs are recursively used as conditions for subsequent steps, as illustrated in fig. 2 (a). Formally, a world model p_θ observes N past frames $x_{s-N:s-1}$ and a possible future action a_s at step $s-1$ to predict the next frame x_s:

$$\hat{x}_s \sim p_\theta(x_s \mid \tilde{x}_{s-N:s-1}, a_s), \qquad s = 0, \ldots, S, \tag{1}$$

where S is the rollout horizon. In eq. (1), $\tilde{x}_s = x_s$ for $s < 0$, and $\tilde{x}_s = \hat{x}_s$ for $s \geq 0$. As rollout proceeds, predicted frames gradually replace the observed history in the conditioning window. This recursive dependency makes long-horizon prediction prone to prediction drift and geometric drift under ego-motion with accumulated noisy predictive observation $\hat{x}_s$.

3.1 Overview

The spirit of our method is anchor-based guidance to address perceptual and geometric drifts jointly. We introduce sparse temporal anchors with two complementary roles. First, they serve as stable long-horizon targets that partition the rollout into shorter chunks, reducing recursive error accumulation (sec. 3.2). Second, they act as spatial reference points for bidirectional geometric masking (sec. 3.3). This unified design is realized by AC-DiT, which generates the intermediate frames by jointly conditioning on past and future anchors, together with the geometry-aware correspondences derived from them (sec. 3.4).

3.2 Anchor-Guided Rollout

To reduce the dependence on drifted intermediate predictions, we redesign the rollout order. Instead of predicting every future frame step by step, we first generate sparse future key frames, referred to as *anchors*, and then generate the intermediate frames within the chunk between two anchors.

Anchor generation. Let $\mathcal{K} = \{k_1, \ldots, k_M\}, 0 \leq k_1 < \cdots < k_M \leq S$, denote the set of anchor steps, where adjacent anchors are separated by a fixed interval. Each anchor is predicted by directly jumping from the observation history to a farther future step using the corresponding action subsequence:

$$\hat{x}_{k_m} \sim p_\theta(x_{k_m} \mid x_{-N+1:-1}, a_{0:k_m}), \qquad m = 1, \ldots, M. \tag{2}$$

These anchors provide coarse future anchors that capture the large-scale scene evolution under the action sequence while avoiding a long recursive rollout that repeatedly conditions on noisy intermediate predictions.

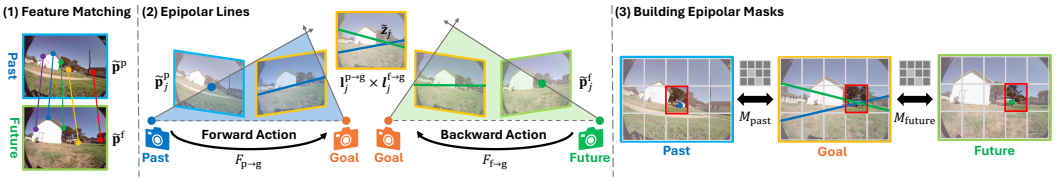


Figure 3: Bidirectional epipolar masking. (1) Match features of past and future anchors. (2) Matched points from the past context and future anchor induce two epipolar lines in the goal frame and form an intersection. (3) This intersection is converted into attention masks that guide cross-attention toward geometry-consistent regions.

Chunk generation. We then decompose the full prediction horizon into chunks between consecutive anchors. For the first chunk, the past anchor consists entirely of observed frames, while for later chunks, it is progressively updated with previously predicted anchors. Let $k_0 = 0$, for each chunk (k_{m-1}, k_m), we generate the intermediate frames under both past and future anchors. Specifically, we define the past and future conditions as $c_m^{\text{past}} = \hat{x}_{k_{m-1}-N-1:k_{m-1}}, c_m^{\text{future}} = \hat{x}_{k_m}$. The intermediate frames inside the chunk are then generated as:

$$\hat{x}_m^{\text{chunk}} \sim p_\theta\left(x_m^{\text{chunk}} \mid c_m^{\text{past}}, c_m^{\text{future}}, a_m^{\rightarrow}, a_m^{\leftarrow}\right), \qquad m = 1, \ldots, M, \tag{3}$$

where $\hat{x}_m^{\text{chunk}}$ contains the internal frames $\hat{x}_s$ satisfying $k_{m-1} < s < k_m$. Here, $a_m^{\rightarrow} = a_{k_{m-1}:k_m}$ denotes the forward action sequence from the start of the chunk to the end, and $a_m^{\leftarrow} = a_{k_m:k_{m-1}}$ denotes the corresponding inverse action sequence defined from the future anchor backward. For a target frame x_s within chunk m, the forward action $a_s^{\rightarrow} = a_{k_{m-1}:s}$ encodes motion from the start of the chunk to the target, while the inverse action $a_s^{\leftarrow} = a_{k_m:s}$ encodes motion from the end of the chunk back to the target.

3.3 Bidirectional Epipolar Masking

Our masking strategy is motivated by the need to inject geometric consistency without explicit 3D priors. While the Diffusion Transformer (DiT) [24] is effective at generating high-quality visual content, it does not explicitly reason about scene geometry. The anchors provide an additional view-based constraint that defines regions of interest for attention, allowing the model to focus on spatially valid correspondences between the past and future views.

To implement this, we first extract sparse feature matching [32] and estimate the fundamental matrices [11] $F_{\text{p}\to\text{g}}$ and $F_{\text{f}\to\text{g}}$ (fig. 3(1)). Given matched points in the past and future frames, we project them to epipolar lines in the goal frame:

$$\mathbf{l}_j^{\text{p}\to\text{g}} = F_{\text{p}\to\text{g}}\tilde{\mathbf{p}}_j^{\text{p}}, \qquad \mathbf{l}_j^{\text{f}\to\text{g}} = F_{\text{f}\to\text{g}}\tilde{\mathbf{p}}_j^{\text{f}}, \qquad \tilde{\mathbf{z}}_j = \mathbf{l}_j^{\text{p}\to\text{g}} \times \mathbf{l}_j^{\text{f}\to\text{g}}, \tag{4}$$

where $\mathbf{l}_j^{\text{p}\to\text{g}}$ and $\mathbf{l}_j^{\text{f}\to\text{g}}$ are the epipolar lines induced by the j-th matched points in the observation and future, respectively, and $\tilde{\mathbf{z}}_j$ is their intersection in the goal (fig. 3 (2)). We then discretize these correspondences onto the transformer token grid and construct binary masks $M_{\text{past}}, M_{\text{fut}} \in \{0,1\}^{L\times L}$, which allow attention only between geometry-consistent source and target (fig. 3 (3)). If no valid correspondence is available for a source token, its attention row remains unmasked.

3.4 Anchor-conditioned Diffusion Transformer

AC-DiT is the generative backbone that integrates our temporal anchors and geometric constraints into a single framework. By conditioning on pairs of bidirectional anchors, the model generates the constituent frames of each chunk. As illustrated in fig. 4, the architecture extends the DiT into a multi-conditional framework that treats future prediction as a task of constrained interpolation.

Encode bidirectional anchors. Scalar conditions, such as actions and time, are encoded using sinusoidal features, followed by a two-layer MLP. For the forward direction, the action $a \in \mathbb{R}^3$, diffusion timestep t, and relative time offset k are embedded as $\psi_a, \psi_t, \psi_k \in \mathbb{R}^d$. For the backward direction, we encode the inverse action a^{-1} and future relative time k_f as $\psi_{a^{-1}}, \psi_{k_f} \in \mathbb{R}^d$. The resulting scalar conditioning is:

$$\xi_{\text{base}} = \psi_t + \psi_k + \psi_a, \qquad \xi_{\text{fut}} = \psi_{a^{-1}} + \psi_{k_f}, \qquad \xi = \xi_{\text{base}} + \gamma_{\text{cond}}\,\xi_{\text{fut}}, \tag{5}$$

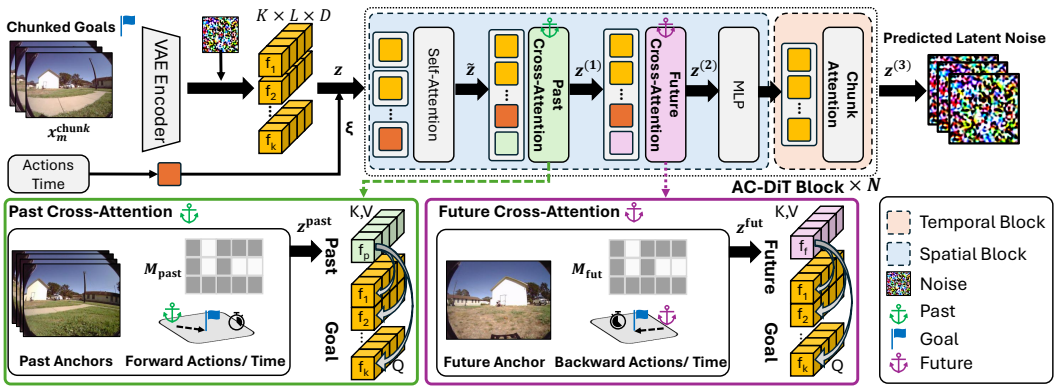


Figure 4: **Overview of AC-DiT.** Given past and future anchors, AC-DiT generates the intermediate frames within each chunk.

where $\gamma_{\text{cond}} = 0$ at initialization. This preserves pretrained behavior at the beginning of finetuning while allowing the model to progressively incorporate future-side information. The term ξ represents the embeddings derived from the diffusion timestep and bidirectional anchors.

Chunk generation with geometric guidance. To generate the frames within a chunk, AC-DiT processes target latent tokens z by conditioning on the past anchors z^{past} and the future anchor z^{fut}. Formally, for target tokens at a given diffusion step, the block computation is defined as follows. First, we use self-attention (SA) to interact with the embeddings ξ denoted as $\tilde{z} = \text{SA}(z; \xi)$. Next, we interact with past and future anchors through past (PA) and future cross-attention (FA):

$$z^{(1)} = \tilde{z} + \gamma_{\text{past}} \, \text{PA}(\tilde{z}, z^{\text{past}}; M_{\text{past}}, \xi), \quad z^{(2)} = z^{(1)} + \gamma_{\text{fut}} \, \text{FA}(z^{(1)}, z^{\text{fut}}; M_{\text{fut}}, \xi). \tag{6}$$

To interact the frames within a chunk, we use chunk attention (CA) along the temporal axis, denoted as $z^{(3)} = z^{(2)} + \gamma_\tau \, \text{CA}(z^{(2)}; K)$, where K is the chunk length. The bidirectional geometric guidance is enforced within the cross-attention layers PA and FA. Specifically, the masks M_{past} and M_{fut} are applied to the attention score matrices to restrict the model's focus to geometrically consistent regions. To ensure a stable training start, we employ zero-initialized gates γ_{p}, γ_{f}, and γ_τ, which allow the model to gradually incorporate conditional and temporal information during the optimization.

3.5 Training

We train AC-DiT using the standard denoising diffusion loss [14]. For a target chunk x, we minimize the mean-squared error between the added Gaussian noise ϵ and the noise predicted by our model ϵ_θ:

$$\mathcal{L} = \mathbb{E}_{x,\epsilon,t} \left[\left| \epsilon - \epsilon_\theta(x^{(t)}, t, c) \right|_2^2 \right], \tag{7}$$

where $x^{(t)}$ is the noisy target at step t and c represents the conditioning context, including the past observation, future anchor, and action sequences. This objective encourages the model to recover clean, geometry-consistent predictions from the bidirectional anchor constraints.

4 Experiments

4.1 Experimental Setups

Datasets. We follow the benchmark setup of [4] and evaluate on four navigation datasets: RECON [28], HuRoN [13] [1], SCAND [15], and TartanDrive [33].

Generation metrics. We evaluate generation quality from two complementary perspectives aligned with our main goals: *perceptual drift* and *geometric drift*. To quantify perceptual drift, we measure visual quality across increasing prediction horizons using FID [12], DreamSim [8], and LPIPS [39],

[1] The higher-resolution version used in the original work is not publicly available. We therefore use the publicly available lower-resolution version (160×120).

Table 1: Perceptual drift across horizons. Gain (%) denotes the relative improvement over NWM.

	Metric	RECON				SCAND				HuRoN*				TartanDrive			
		EgoWM	NWM	Ours	Gain	EgoWM	NWM	Ours	Gain	EgoWM	NWM	Ours	Gain	EgoWM	NWM	Ours	Gain
1s	LPIPS ↓	0.504	0.330	**0.281**	14.8	0.497	0.353	**0.297**	15.9	0.653	0.445	**0.334**	24.9	0.481	0.381	**0.303**	20.5
	DreamSim ↓	0.239	0.125	**0.103**	17.6	0.322	0.211	**0.183**	13.3	0.453	0.236	**0.178**	24.6	0.264	0.177	**0.137**	22.6
	FID ↓	97.095	52.367	**47.443**	9.4	119.958	69.264	**65.007**	6.2	152.512	97.726	**78.541**	19.6	89.769	50.642	**42.894**	15.3
2s	LPIPS ↓	0.591	0.395	**0.325**	17.7	0.524	0.413	**0.345**	16.5	0.676	0.508	**0.389**	23.4	0.527	0.446	**0.373**	16.4
	DreamSim ↓	0.366	0.159	**0.123**	22.6	0.369	0.256	**0.213**	16.8	0.498	0.291	**0.203**	30.2	0.318	0.241	**0.174**	27.8
	FID ↓	119.380	61.500	**54.929**	10.7	119.526	73.670	**68.514**	7.0	152.669	101.240	**81.341**	19.7	85.874	57.526	**47.899**	16.7
4s	LPIPS ↓	0.651	0.448	**0.385**	14.1	0.601	0.473	**0.401**	15.2	0.733	0.559	**0.439**	21.5	0.642	0.519	**0.439**	15.4
	DreamSim ↓	0.497	0.212	**0.154**	27.4	0.519	0.316	**0.257**	18.7	0.607	0.359	**0.260**	27.6	0.463	0.331	**0.234**	29.3
	FID ↓	165.934	70.929	**58.350**	17.7	187.576	85.046	**76.430**	10.1	195.754	114.395	**87.457**	23.5	158.649	68.391	**54.247**	20.7
8s	LPIPS ↓	0.727	0.504	**0.444**	11.9	0.680	0.547	**0.464**	15.2	0.774	0.603	**0.506**	16.1	0.748	0.572	**0.504**	11.9
	DreamSim ↓	0.686	0.280	**0.200**	28.6	0.681	0.422	**0.321**	23.9	0.689	0.425	**0.320**	24.7	0.659	0.428	**0.305**	28.7
	FID ↓	265.686	90.968	**67.298**	26.0	272.919	109.985	**88.976**	19.1	263.516	131.694	**104.536**	20.6	258.002	88.272	**62.614**	29.1
16s	LPIPS ↓	0.765	0.557	**0.508**	8.8	0.719	0.595	**0.510**	14.3	0.787	0.627	**0.544**	13.2	0.796	0.624	**0.553**	11.4
	DreamSim ↓	0.775	0.360	**0.263**	26.9	0.761	0.486	**0.374**	23.0	0.729	0.492	**0.366**	25.6	0.772	0.495	**0.393**	20.6
	FID ↓	365.810	112.679	**76.898**	31.8	330.710	131.263	**100.424**	23.5	313.405	141.702	**112.330**	20.7	325.248	107.086	**78.562**	26.6

following [4]. These metrics assess how prediction quality degrades as rollout extends farther in time. To quantify geometric, we evaluate whether generated frames remain aligned with the conditioned actions and the underlying scene geometry. Specifically, we use epipolar distance (ED) [11] and Sampson error (SE) [27, 17] to measure geometric consistency with the induced viewpoint change, and MEt3R [2] to assess multi-view consistency after geometry-aware alignment.

Planning metrics. To evaluate downstream planning benefit, we follow the planning setup of [4]. In this setting, the planner itself is kept unchanged, so the comparison isolates the effect of improved future prediction quality. We report Absolute Trajectory Error (ATE), Final Displacement Error (FDE), and Relative Pose Error (RPE).

Implementation details. We use the pretrained CDiT/XL model from [4] as the backbone of AC-DiT. The anchor-conditioned, bidirectional, and temporal modules are initialized from scratch. Following [4], we adopt the VAE tokenizer from [6]. We fine-tune the model using the AdamW optimizer [22] with a batch size of 4 and a learning rate of 8×10^{-6}. All experiments are conducted on NVIDIA A100 GPUs. Detailed experimental setups are provided in the appendix.

4.2 Generation

4.2.1 Perceptual Drift

We evaluate perceptual drift on four datasets across horizons from 1s to 16s in table 1, using NWM [4] and EgoWM [3] as baselines under the same 224×224 image resolution. Our method consistently outperforms both baselines across all datasets, horizons, and visual-quality metrics, showing that it degrades much more gracefully as rollout extends.

The advantage becomes increasingly clear at longer horizons, where autoregressive prediction is more vulnerable to recursive error accumulation. NWM progressively deteriorates as prediction errors are fed back into later steps, while EgoWM becomes blurred and abstract after only a few seconds. In contrast, our anchor-based inference and anchor-conditioned generation reduce recursive corruption and maintain more stable long-horizon prediction. This trend is also visible in fig. 5: NWM deteriorates noticeably beyond 4s, EgoWM degrades even earlier, while our method preserves clearer structure and more interpretable content over time.

4.2.2 Geometric Drift

We evaluate geometric drift from two complementary perspectives: geometric alignment with the conditioned motion and multi-view consistency across viewpoint changes. In table 2, our method achieves lower epipolar distance and Sampson error than both NWM and EgoWM in nearly all settings across datasets and horizons, indicating stronger geometric consistency with the action-induced viewpoint change. The advantage becomes more pronounced at longer horizons, where autoregressive baselines accumulate more rollout error. In contrast, our method remains better aligned with the conditioned motion because future anchors provide explicit endpoint constraints and bidirectional epipolar masks further restrict attention to geometry-consistent regions. This trend is also visible in fig. 6 left, where our method preserves scene structure more faithfully and produces more geometry-consistent correspondences.



Figure 5: Qualitative results of perceptual drift over time. Red frames highlight regions where the baselines become visibly corrupted or difficult to interpret.

Table 2: Geometric drift comparison across datasets and prediction horizons.

		RECON			SCAND			HuRoN			TartanDrive		
		EgoWM	NWM	Ours	EgoWM	NWM	Ours	EgoWM	NWM	Ours	EgoWM	NWM	Ours
1s	ED ↓	15.35	9.40	**8.69**	10.45	4.22	**3.58**	14.11	8.99	**7.48**	11.84	6.41	**5.52**
	SE ↓	178.41	80.46	**63.41**	84.56	14.98	**11.10**	140.75	60.37	**50.52**	97.47	35.63	**30.83**
2s	ED ↓	16.06	9.46	**9.15**	9.17	4.58	**3.61**	10.71	11.28	**9.57**	13.09	6.84	**6.54**
	SE ↓	202.20	71.25	**65.99**	64.29	20.39	**11.38**	83.61	88.18	**81.41**	115.02	39.18	**36.82**
4s	ED ↓	19.34	9.80	**8.93**	8.64	4.23	**3.77**	12.21	10.82	**9.43**	11.50	7.13	**6.86**
	SE ↓	277.52	75.28	**65.09**	60.90	18.89	**13.65**	101.38	92.67	**82.05**	95.14	41.71	**38.75**
8s	ED ↓	18.11	9.81	**9.61**	7.16	5.00	**3.83**	11.71	12.53	**9.91**	9.35	8.83	**7.35**
	SE ↓	255.13	80.54	**75.27**	51.52	21.87	**14.88**	105.80	132.04	**89.75**	75.71	72.48	**49.25**
16s	ED ↓	16.51	11.37	**9.92**	6.65	5.40	**4.02**	**10.70**	14.83	11.15	8.31	10.09	**7.85**
	SE ↓	204.85	107.14	**80.83**	49.69	33.09	**17.66**	**91.75**	191.26	108.71	63.89	94.32	**55.96**

We further evaluate geometric drift under viewpoint change using MEt3R in table 3. Our method consistently achieves lower MEt3R scores than NWM across all datasets and skip lengths, showing stronger multi-view coherence after geometry-aware alignment across different range of actions. As shown in fig. 6 right, our method better preserves scene structure and viewpoint consistency. Together, these results show that reducing geometric drift also improves cross-view coherence.

4.3 Planning

The previous results show that our method produces future predictions with lower perceptual drift and geometric drift. We next examine whether these gains translate into downstream planning benefit. Since the planner itself is kept unchanged, any performance improvement can be attributed to better predicted futures rather than changes to the planning algorithm.

Following the real-world planning setup of [4], we sample 32 candidate trajectories over a 4-second horizon and report results for both Cross-Entropy Method (CEM) [26] and NoMaD (NMD) [31] in table 4. Our method consistently improves planning performance over NWM on both datasets,

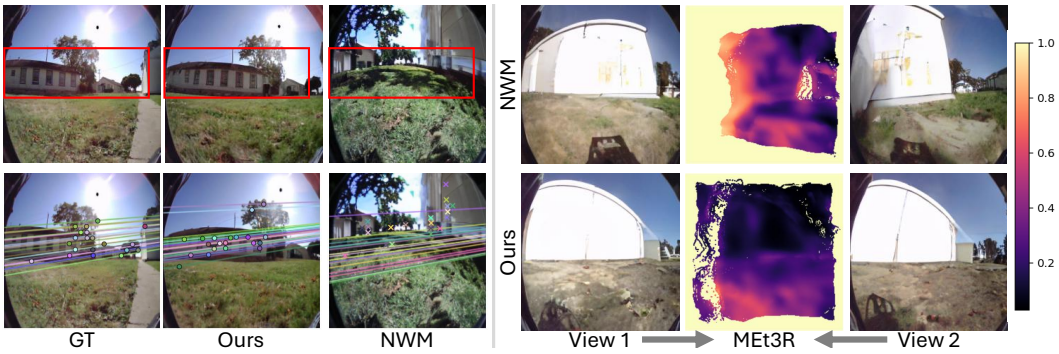


Figure 6: Qualitative results of geometric drift. **Left:** Epipolar geometry visualization. Red boxes mark regions sensitive to viewpoint change under the commanded motion. In the bottom row, dots denote matches within an 8-pixel epipolar threshold, while crosses indicate larger violations. Our method better preserves scene structure and object layout, yielding more valid matches and stronger geometric alignment. **Right:** MEt3R visualization. Compared with NWM, our method preserves more coherent scene structure and viewpoint-consistent content with better MEt3R scores.

Table 3: MEt3R comparison. Lower is better.

	Skip	1s	2s	4s
RECON	NWM	0.331	0.400	0.463
	Ours	**0.306**	**0.361**	**0.416**
SCAND	NWM	0.348	0.433	0.510
	Ours	**0.342**	**0.388**	**0.439**
HuRoN	NWM	0.466	0.558	0.628
	Ours	**0.429**	**0.489**	**0.540**
Tartan	NWM	0.385	0.464	0.532
	Ours	**0.355**	**0.411**	**0.466**

Table 4: Planning performance.

		ATE↓	FDE↓	RPE↓	LPIPS↓
RECON	CEM+NWM	3.18	5.86	0.55	0.438
	CEM+Ours	**2.78**	**5.15**	**0.51**	**0.398**
	NMD+NWM	2.97	5.38	0.52	0.448
	NMD+Ours	**2.70**	**4.75**	**0.49**	**0.366**
SCAND	CEM+NWM	2.49	4.34	0.37	0.506
	CEM+Ours	**2.30**	**3.99**	**0.35**	**0.483**
	NMD+NWM	2.72	4.82	**0.40**	0.512
	NMD+Ours	**2.69**	**4.76**	**0.40**	**0.454**



Figure 7: Qualitative planning results. Using the same planners, our method yields more coherent predictions and selects trajectories that better follow the intended motion and end closer to the goals.

indicating that more stable and action-consistent prediction leads to more accurate downstream trajectory selection.

This effect is also visible in fig. 7. Because the planners operate under the same setup, the performance gap mainly comes from the quality of the predicted futures. In particular, reduced perceptual drift and geometric drift help the selected trajectory remain better aligned with the intended motion and closer to the target observation. These results show that improving future prediction quality also improves the practical usefulness of world models for downstream planning.

Table 5: Ablation study on inference strategy. AR: autoregressive rollout. Ours: Anchored-guided rollout.

	1s			2s			4s			8s			16s		
	LPIPS ↓	FID ↓	SE ↓	LPIPS ↓	FID ↓	SE ↓	LPIPS ↓	FID ↓	SE ↓	LPIPS ↓	FID ↓	SE ↓	LPIPS ↓	FID ↓	SE ↓
AR	0.333	53.233	65.76	0.390	64.975	66.59	0.451	78.687	67.75	0.518	105.839	79.19	0.577	140.292	84.74
Ours	**0.281**	**47.443**	**63.41**	**0.325**	**54.929**	**65.99**	**0.385**	**58.350**	**65.09**	**0.444**	**67.298**	**75.27**	**0.508**	**76.898**	**80.83**

Table 6: Effect of anchor quality on intermediate-frame FID on RECON. Ground-truth anchors provide an approximate upper bound for intermediate-frame generation.

	0–4 s	4–8 s	8–12 s	12–16 s	Avg.
Generated anchors	52.78	63.11	70.87	76.28	65.76
Ground-truth anchors	**50.76**	**49.99**	**51.79**	**54.04**	**51.65**

4.4 Ablation Study

We next analyze how each component contributes to the final improvements. Since our method is built on two tightly connected ideas: anchor-guided rollout and bidirectional epipolar masking. Hence, we organize the ablations according to the role of each design. We analyze four aspects of our method: the inference strategy, anchor quality, anchor spacing, and the components of AC-DiT.

Effect of anchor-based inference. We first compare conventional autoregressive rollout (AR) with the proposed anchor-based inference while keeping the same backbone model. The results in table 5 show that simply replacing AR with anchor-based inference already improves visual quality, confirming that the new rollout order effectively reduces recursive error accumulation and mitigates perceptual drift, but it does not directly contribute to geometric drift.

Effect of anchor quality. We replace predicted anchors with ground-truth frames while keeping the same intermediate-frame generator. As shown in table 6, ground-truth anchors improve all prediction windows, confirming that intermediate-frame quality depends on the quality of the endpoint anchors. This setting provides an approximate upper bound for the interpolation stage.

Effect of anchor spacing. We vary the inference anchor spacing from $0.5\,\mathrm{s}$ to $4\,\mathrm{s}$. As shown in table 7, increasing the spacing consistently improves intermediate-frame FID, with the $4\,\mathrm{s}$ setting performing best across all four temporal windows. However, the $1\,\mathrm{s}$ setting achieves the lowest Sampson error at four of the five evaluation horizons, indicating stronger geometric consistency. We therefore use $1\,\mathrm{s}$ as the default balance between perceptual quality and geometric consistency.

Effect of model components. We next examine the contribution of each component in AC-DiT by ablating the future condition, bidirectional epipolar masking, and the chunked generation structure. The results are summarized in table 8. Across variants, FID and LPIPS remain relatively similar, indicating that the main gain in long-horizon visual quality comes from the anchor-guided rollout. By contrast, the differences in Sampson error are much larger, showing that geometric drift depends strongly on the internal design of AC-DiT.

In particular, the full model achieves the best average Sampson error, indicating that all three components are needed for the strongest geometric consistency. Removing the future condition weakens the bidirectional constraint, while removing the epipolar mask removes the geometry-aware guidance. The degradation is especially clear when epipolar masking is used without future conditioning: In this case, the geometric cue degenerates from an intersection-based localization signal to a weaker line-based constraint, leading to worse epipolar consistency. These results support our main claim that bidirectional anchors and epipolar masking are complementary: The future anchor provides the second view needed for bidirectional reasoning, and the epipolar mask turns this structure into effective geometry-aware attention.

5 Conclusion

We propose DR-NWM, a generative model that addresses two coupled failures of current approaches: perceptual drift over long autoregressive rollouts and geometric drift under ego-motion. Our method

Table 7: Effect of inference anchor spacing on RECON. We report FID over intermediate frames, excluding anchors, across four temporal windows, and Sampson Error, across five prediction horizons.

Anchor spacing	Intermediate-frame FID ↓				Sampson Error ↓				
	0–4 s	4–8 s	8–12 s	12–16 s	1 s	2 s	4 s	8 s	16 s
0.5 s	55.43	74.12	87.66	100.04	72.58	70.26	70.60	79.32	83.89
1 s	52.78	63.11	70.87	76.28	**63.41**	**65.99**	**65.09**	75.27	**80.83**
2 s	50.93	59.38	63.64	67.56	66.20	78.75	69.70	**73.77**	93.69
4 s	**50.70**	**57.33**	**63.14**	**65.83**	69.53	67.85	79.25	79.52	94.37

Table 8: Ablation of the key components of AC-DiT. Pair separation denotes the temporal separation between generated frame pairs used to compute the Sampson error; it is an evaluation setting and is unrelated to the spacing between inference anchors. We additionally report LPIPS and FID averaged across evaluation horizons.

Future Anchor	Epi. Mask	Chunk	Sampson Error ↓				Avg. LPIPS ↓	Avg. FID ↓
			Pair sep.=3	Pair sep.=2	Pair sep.=1	Avg.		
×	×	×	67.99	61.24	53.06	60.76	0.392	60.040
✓	✓	×	65.51	67.50	51.46	61.49	0.389	60.312
×	✓	✓	76.63	61.76	54.45	64.31	0.389	**59.447**
✓	×	✓	66.49	69.24	52.90	62.88	0.390	59.456
✓	✓	✓	**63.41**	**60.89**	**44.80**	**56.37**	**0.388**	60.983

combines anchor-guided rollout, bidirectional epipolar grounding, and the AC-DiT to generate future observations that remain more stable over extended horizons while staying more faithful to the geometry induced by the agent's actions, all without requiring explicit 3D supervision. By redesigning future prediction rollout around sparse anchors and bidirectional chunk generation, DR-NWM reduces recursive error accumulation and introduces stronger geometric constraints. Experiments on four navigation benchmarks demonstrate consistent improvements in long-horizon visual prediction, geometric consistency, multi-view coherence, and downstream planning performance under the same planners. Taken together, these results show that reducing drift and enforcing action-consistent geometric grounding are both essential for building reliable and practically useful world models for navigation. We hope this work motivates further research on world models that are not only visually plausible but also stable, geometrically grounded, and directly useful for embodied decision making.

Acknowledgment

This work was funded by Honda R&D Co., Ltd, SwissAI, and Sportradar. We would like to express our gratitude to Bastien Van Delft, Mohamed Abdelfattah, and Yasamin Borhani for their insightful feedback.

References

[1] Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, et al. Cosmos world foundation model platform for physical ai. *arXiv preprint arXiv:2501.03575*, 2025.

[2] Mohammad Asim, Christopher Wewer, Thomas Wimmer, Bernt Schiele, and Jan Eric Lenssen. Met3r: Measuring multi-view consistency in generated images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6034–6044, 2025.

[3] Anurag Bagchi, Zhipeng Bao, Homanga Bharadhwaj, Yu-Xiong Wang, Pavel Tokmakov, and Martial Hebert. Walk through paintings: Egocentric world models from internet priors. *arXiv preprint arXiv:2601.15284*, 2026.

[4] Amir Bar, Gaoyue Zhou, Danny Tran, Trevor Darrell, and Yann LeCun. Navigation world models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 15791–15801, 2025.

[5] Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. *Advances in neural information processing systems*, 28, 2015.

[6] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

[7] Yixiang Dai, Fan Jiang, Chiyu Wang, Mu Xu, and Yonggang Qi. Fantasyworld: Geometry-consistent world modeling via unified video and 3d prediction. *arXiv preprint arXiv:2509.21657*, 2025.

[8] Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. Dreamsim: Learning new dimensions of human visual similarity using synthetic data. *arXiv preprint arXiv:2306.09344*, 2023.

[9] David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. *Advances in neural information processing systems*, 31, 2018.

[10] Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*, 2023.

[11] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003.

[12] 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. *Advances in neural information processing systems*, 30, 2017.

[13] Noriaki Hirose, Dhruv Shah, Ajay Sridhar, and Sergey Levine. Sacson: Scalable autonomous control for social navigation. *IEEE Robotics and Automation Letters*, 9(1):49–56, 2023.

[14] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[15] Haresh Karnan, Anirudh Nair, Xuesu Xiao, Garrett Warnell, Sören Pirk, Alexander Toshev, Justin Hart, Joydeep Biswas, and Peter Stone. Socially compliant navigation dataset (scand): A large-scale dataset of demonstrations for social navigation. *IEEE Robotics and Automation Letters*, 7(4):11807–11814, 2022.

[16] Dongwon Kim, Gawon Seo, Jinsung Lee, Minsu Cho, and Suha Kwak. Planning in 8 tokens: A compact discrete tokenizer for latent world model. *arXiv preprint arXiv:2603.05438*, 2026.

[17] Orest Kupyn, Fabian Manhardt, Federico Tombari, and Christian Rupprecht. Epipolar geometry improves video generation models. *arXiv preprint arXiv:2510.21615*, 2025.

[18] Jean-Claude Latombe. *Robot motion planning*, volume 124. Springer Science & Business Media, 2012.

[19] Yann LeCun et al. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review*, 62(1):1–62, 2022.

[20] Wuyang Li, Yang Gao, Mariam Hassan, Lan Feng, Wentao Pan, Po-Chien Luan, and Alexandre Alahi. Everanimate: Minute-scale human animation via latent flow restoration. *arXiv preprint arXiv:2605.15042*, 2026.

[21] Wuyang Li, Wentao Pan, Po-Chien Luan, Yang Gao, and Alexandre Alahi. Stable video infinity: Infinite-length video generation with error recycling. *arXiv preprint arXiv:2510.09212*, 2025.

[22] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[23] George A Miller, Galanter Eugene, and Karl H Pribram. Plans and the structure of behaviour. In *Systems research for behavioral science*, pages 369–382. Routledge, 2017.

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

[25] Xuanchi Ren, Tianchang Shen, Jiahui Huang, Huan Ling, Yifan Lu, Merlin Nimier-David, Thomas Müller, Alexander Keller, Sanja Fidler, and Jun Gao. Gen3c: 3d-informed world-consistent video generation with precise camera control. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6121–6132, 2025.

[26] Reuven Y Rubinstein. Optimization of computer simulation models with rare events. *European Journal of Operational Research*, 99(1):89–112, 1997.

[27] Paul D Sampson. Fitting conic sections to "very scattered" data: An iterative refinement of the bookstein algorithm. *Computer graphics and image processing*, 18(1):97–108, 1982.

[28] Dhruv Shah, Benjamin Eysenbach, Gregory Kahn, Nicholas Rhinehart, and Sergey Levine. Rapid exploration for open-world navigation with latent goal models. *arXiv preprint arXiv:2104.05859*, 2021.

[29] Dhruv Shah, Ajay Sridhar, Nitish Dashora, Kyle Stachowicz, Kevin Black, Noriaki Hirose, and Sergey Levine. Vint: A foundation model for visual navigation. *arXiv preprint arXiv:2306.14846*, 2023.

[30] Wangtian Shen, Ziyang Meng, Jinming Ma, Mingliang Zhou, and Diyun Xiang. An efficient and multi-modal navigation system with one-step world model. *arXiv preprint arXiv:2601.12277*,

2026.

[31] Ajay Sridhar, Dhruv Shah, Catherine Glossop, and Sergey Levine. Nomad: Goal masked diffusion policies for navigation and exploration. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 63–70. IEEE, 2024.

[32] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8922–8931, 2021.

[33] Samuel Triest, Matthew Sivaprakasam, Sean J Wang, Wenshan Wang, Aaron M Johnson, and Sebastian Scherer. Tartandrive: A large-scale dataset for learning off-road dynamics models. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 2546–2552. IEEE, 2022.

[34] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20697–20709, 2024.

[35] Zhouxia Wang, Ziyang Yuan, Xintao Wang, Yaowei Li, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. Motionctrl: A unified and flexible motion controller for video generation. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024.

[36] Haoyu Wu, Diankun Wu, Tianyu He, Junliang Guo, Yang Ye, Yueqi Duan, and Jiang Bian. Geometry forcing: Marrying video diffusion and 3d representation for consistent world modeling. *arXiv preprint arXiv:2507.07982*, 2025.

[37] Han Yan, Zishang Xiang, Zeyu Zhang, and Hao Tang. Mwm: Mobile world models for action-conditioned consistent prediction. *arXiv preprint arXiv:2603.07799*, 2026.

[38] Mingkun Zhang, Wangtian Shen, Fan Zhang, Haijian Qin, Zihao Pei, and Ziyang Meng. Rae-nwm: Navigation world model in dense visual representation space. *arXiv preprint arXiv:2603.09241*, 2026.

[39] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

A Geometric Justification of Bidirectional Epipolar Intersection

We briefly justify why the intersection of two bidirectional epipolar lines provides a valid localization cue for the target view. Consider three views: a past view I^{past}, a future view I^{fut}, and a target (goal) view I^{goal}. Let X denote a 3D scene point, and let $C_{\mathrm{past}}, C_{\mathrm{fut}}, C_{\mathrm{goal}}$ be the corresponding camera centers. Denote by $\mathbf{x}^{\mathrm{past}}, \mathbf{x}^{\mathrm{fut}}, \mathbf{x}^{\mathrm{goal}}$ the image projections of X in the three views.

An image point does not correspond to a unique 3D point; rather, it defines a back-projected ray starting at the camera center and passing through the image plane. Thus, the point $\mathbf{x}^{\mathrm{past}}$ defines a 3D ray R_{past} through C_{past}, and $\mathbf{x}^{\mathrm{fut}}$ defines a 3D ray R_{fut} through C_{fut}. Because both image points are projections of the same scene point X, the point X must lie on both rays.

Now consider the goal view. The projection of the ray R_{past} onto the goal image plane defines an epipolar line $\mathbf{l}^{\mathrm{past}\to\mathrm{goal}}$. Since $X \in R_{\mathrm{past}}$, its projection $\mathbf{x}^{\mathrm{goal}}$ must lie on this line. Similarly, the projection of the ray R_{fut} onto the goal image plane defines another epipolar line $\mathbf{l}^{\mathrm{fut}\to\mathrm{goal}}$, and $\mathbf{x}^{\mathrm{goal}}$ must also lie on this line. Therefore, in the noise-free case, the target point is exactly the intersection of the two epipolar lines:

$$\mathbf{x}^{\mathrm{goal}} = \mathbf{l}^{\mathrm{past}\to\mathrm{goal}} \cap \mathbf{l}^{\mathrm{fut}\to\mathrm{goal}}. \tag{8}$$

Using the fundamental matrix, the two epipolar lines in the goal view are given by

$$\mathbf{l}^{\mathrm{past}\to\mathrm{goal}} = F_{\mathrm{goal,past}}\,\tilde{\mathbf{x}}^{\mathrm{past}}, \qquad \mathbf{l}^{\mathrm{fut}\to\mathrm{goal}} = F_{\mathrm{goal,fut}}\,\tilde{\mathbf{x}}^{\mathrm{fut}}, \tag{9}$$

where $\tilde{\mathbf{x}}$ denotes homogeneous image coordinates. The target point $\tilde{\mathbf{x}}^{\mathrm{goal}}$ satisfies both epipolar constraints,

$$\tilde{\mathbf{x}}^{\mathrm{goal}\top}\mathbf{l}^{\mathrm{past}\to\mathrm{goal}} = 0, \qquad \tilde{\mathbf{x}}^{\mathrm{goal}\top}\mathbf{l}^{\mathrm{fut}\to\mathrm{goal}} = 0. \tag{10}$$

Since $\tilde{\mathbf{x}}^{\mathrm{goal}}$ lies on both lines, it can be obtained as their cross product:

$$\tilde{\mathbf{x}}^{\mathrm{goal}} = \mathbf{l}^{\mathrm{past}\to\mathrm{goal}} \times \mathbf{l}^{\mathrm{fut}\to\mathrm{goal}}. \tag{11}$$

In practice, estimated correspondences and fundamental matrices are noisy, so the two lines may not intersect exactly at the true target correspondence. Nevertheless, their intersection still provides a substantially tighter localization cue than the one-sided point-to-line constraint available in standard autoregressive rollout. This is the geometric basis for our bidirectional epipolar masks.

B Details of Bidirectional Epipolar Masking

B.1 Detailed Construction of Epipolar Attention Masks

This section provides the implementation details omitted from section 3.3. Our goal is to use multi-view geometry to constrain cross-view attention without introducing explicit 3D reconstruction or depth supervision. The mask is constructed for each triplet consisting of a past observation, a target goal frame, and a later future anchor.

Triplet formation. During training, we use the last frame of the past context as the reference observation for geometry. For each target goal frame, we pair this observation with a future anchor sampled later in the same trajectory. In the chunked training setup used in our experiments, several consecutive goal frames share the same future anchor, which encourages temporally consistent geometric constraints across nearby targets.

Sparse matching and geometric estimation. For each triplet, we compute sparse correspondences for three image pairs: observation-goal, observation-future, and future-goal. We extract correspondences using LoFTR [32] and discard low-confidence matches. To improve robustness, we further reject matches that lie on unreliable regions, including sky, dynamic objects, and invalid image borders. In our implementation, this filtering combines a semantic segmentation model with lightweight color-based heuristics for sky/cloud detection and a simple border-validity test for dark fisheye boundaries.

After filtering, we estimate the fundamental matrices $F_{\mathrm{p}\to\mathrm{g}}$ and $F_{\mathrm{f}\to\mathrm{g}}$ with RANSAC [11]. If too few filtered matches remain or RANSAC fails to produce a valid model, the corresponding pair is marked unreliable and the geometric mask is disabled for that training example.

Epipolar line construction. Let $\tilde{\mathbf{p}}_j^{\mathrm{p}}$ and $\tilde{\mathbf{p}}_j^{\mathrm{f}}$ denote the homogeneous coordinates of the j-th matched points in the observation and future frames, respectively. We use these points together with the estimated fundamental matrices to project each source point into the goal frame as an epipolar line:

$$\mathbf{l}_j^{\mathrm{p}\to\mathrm{g}} = F_{\mathrm{p}\to\mathrm{g}}\tilde{\mathbf{p}}_j^{\mathrm{p}}, \qquad \mathbf{l}_j^{\mathrm{f}\to\mathrm{g}} = F_{\mathrm{f}\to\mathrm{g}}\tilde{\mathbf{p}}_j^{\mathrm{f}}. \tag{12}$$

The two lines provide two independent geometric constraints on the location of the same scene element in the goal frame. We therefore compute their intersection as

$$\tilde{\mathbf{z}}_j = \mathbf{l}_j^{\mathrm{p}\to\mathrm{g}} \times \mathbf{l}_j^{\mathrm{f}\to\mathrm{g}}. \tag{13}$$

If the intersection is numerically unstable or lies outside the image domain, we discard that correspondence. Otherwise, it is treated as the geometry-consistent target location in the goal frame.

Discretization onto the token grid. The geometric constraints are applied in the latent token space used by the transformer. Let the latent spatial grid contain L tokens. In our implementation, 224×224 input images are encoded into a 28×28 latent map and then patchified with patch size 2, yielding a 14×14 token grid and therefore $L = 196$ spatial tokens.

Each valid observation point, future point, and goal intersection is mapped from pixel coordinates to its corresponding token index on this grid. We then build two binary masks, $M_{\mathrm{past}}, M_{\mathrm{fut}} \in \{0,1\}^{L\times L}$, where rows correspond to source tokens and columns correspond to goal tokens. For a valid correspondence j, the observation-source token is connected only to the goal token containing $\tilde{\mathbf{z}}_j$ in M_{past}, and the future-source token is connected only to the same goal token in M_{fut}. If multiple valid correspondences fall into the same source token, that row may enable multiple goal tokens. If no valid correspondence exists for a source token, its row remains fully unmasked, allowing the model to revert to unconstrained attention rather than enforcing an incorrect geometric hypothesis.

Reliability-aware gating. To avoid using noisy geometry, we compute a reliability score for each pairwise matching result from the number of RANSAC inliers. For a pair $a \to b$, we define

$$r_{a\to b} = \mathbb{1}[F_{a\to b} \text{ exists}] \cdot \mathrm{clip}\left(\frac{n_{a\to b} - n_{\min}}{n_{\mathrm{sat}} - n_{\min}}, 0, 1\right), \tag{14}$$

where $n_{a\to b}$ is the number of inlier correspondences, $n_{\min}$ is the minimum inlier count required before the score becomes non-zero, and n_{sat} is the count at which the score saturates to 1. The final triplet reliability is the minimum of the observation-goal, observation-future, and future-goal scores:

$$r_{\mathrm{triplet}} = \min\big(r_{\mathrm{p}\to\mathrm{g}}, r_{\mathrm{p}\to\mathrm{f}}, r_{\mathrm{f}\to\mathrm{g}}\big). \tag{15}$$

If r_{triplet} falls below a threshold, we disable the geometric mask and revert to full attention for that example.

Temporal smoothing in chunked training. When several consecutive goal frames are trained jointly as a temporal chunk, the mask can vary abruptly from frame to frame due to noisy sparse matches. To reduce this jitter, we smooth the binary masks over time with an exponential moving average. If M_t denotes the binary mask for the t-th frame in a chunk, we compute

$$\bar{M}_t = \alpha \bar{M}_{t-1} + (1-\alpha) M_t, \tag{16}$$

followed by thresholding to recover a binary mask. This encourages neighboring goal frames in the same chunk to use consistent geometric constraints.

Use inside the transformer. The resulting masks are applied as cross-attention masks, rather than as an auxiliary loss. Specifically, they constrain which goal tokens may attend to observation and future tokens during cross-view conditioning. In this way, the model retains the expressive power of DiT while being biased toward spatially valid correspondences supported jointly by the observation and future anchor.

B.2 Hyperparameters

In the chunked training setup, we use a chunk length of $K = 3$, so that each training sample contains short temporal windows over which the geometric and temporal consistency constraints are enforced.

For sparse matching, we retain LoFTR [32] correspondences with confidence above $\tau_{\text{match}} = 0.8$. We then estimate the fundamental matrices with RANSAC using a reprojection threshold of $\tau_{\text{ransac}} = 3$ pixels, and only attempt geometric estimation when at least 8 filtered matches are available. For the reliability-aware mask gating described in section B.1, we set the minimum inlier count to $n_{\min} = 16$, the saturation count to $n_{\text{sat}} = 64$, and the reliability threshold to $\tau_{\text{rel}} = 0.1$. For chunked mask stabilization, we smooth the binary masks temporally with an exponential moving average using $\alpha = 0.6$, and then binarize the smoothed masks with threshold $\tau_{\text{temp}} = 0.5$. Unless otherwise stated, we use the point-intersection version of the epipolar mask.

C Detailed Experimental Setups

C.1 Datasets

We follow the benchmark protocol of [4] and evaluate on four navigation datasets: RECON [28], HuRoN [13], SCAND [15], and TartanDrive [33]. Together, these datasets cover a diverse range of navigation scenarios, including real-world indoor and outdoor robot navigation as well as large-scale synthetic driving environments. This diversity is important for evaluating both long-horizon prediction quality and action-conditioned viewpoint consistency under different scene layouts, motion patterns, and visual appearances.

RECON is a real-world robotic navigation dataset designed for visual navigation and exploration. It contains egocentric image observations paired with robot motion, making it suitable for evaluating action-conditioned future prediction in realistic embodied settings. Its real-world scene diversity and viewpoint changes make it a useful benchmark for studying long-horizon rollout stability.

HuRoN is a real-world navigation dataset collected in human-centered environments. Compared with standard robot-navigation benchmarks, it contains more frequent scene changes caused by clutter, pedestrians, and dynamic surroundings, which makes future prediction and action-conditioned consistency more challenging. Following [4], we use the publicly available lower-resolution version, since the higher-resolution version used in the original work is not publicly available.

SCAND focuses on socially aware navigation in indoor environments. The dataset includes robot trajectories collected in the presence of people and obstacles, introducing complex viewpoint changes and interaction-rich motion patterns. These properties make SCAND particularly relevant for evaluating whether predicted futures remain aligned with commanded motion in crowded or structured environments.

TartanDrive is a large-scale driving dataset that provides egocentric observations under diverse outdoor conditions. Compared with the indoor robot-navigation datasets above, TartanDrive involves larger scene depth, faster ego-motion, and longer-range viewpoint changes. It therefore provides a complementary testbed for evaluating prediction drift and action-conditioned coherence in more challenging long-horizon settings.

C.2 Baselines

For NWM [4], we use the CDiT/XL checkpoint released by the authors as our baseline. This checkpoint is trained on the additional Ego4D dataset to provide better generalizability. For EgoWM [3], we use the official pretrained Stable Video Diffusion (SVD) weights [6]. To ensure a fair comparison, we directly generate images at the same resolution as NWM.

C.3 Metrics

Epipolar Distance. We measure epipolar distance as the pixel distance between a matched point and its corresponding epipolar line induced by the estimated fundamental matrix. For a correspondence $x \leftrightarrow x'$ and a fundamental matrix F, the epipolar line in the second view is

$$\mathbf{l}' = Fx = (a, b, c)^\top. \tag{17}$$

The epipolar distance from the point $x' = (u', v', 1)^\top$ to this line is

$$d_{\text{epi}}(x', \mathbf{l}') = \frac{|au' + bv' + c|}{\sqrt{a^2 + b^2}}. \tag{18}$$

We compute this quantity for valid correspondences and report the average over pairs. Lower values indicate that generated observations remain more consistent with the conditioned viewpoint change. The point-to-line mapping induced by F and the standard point-to-line distance are classical epipolar-geometry constructions.

Sampson Error. We also report the Sampson epipolar error, a first-order approximation to geometric reprojection error:

$$\mathrm{SE} = \frac{(x'^{\top} F x)^2}{(Fx)_1^2 + (Fx)_2^2 + (F^{\top}x')_1^2 + (F^{\top}x')_2^2}. \tag{19}$$

Lower Sampson error indicates better satisfaction of the epipolar constraint and stronger geometric consistency in the generated observations.

MEt3R. We use MEt3R [2] as a metric of multi-view consistency. Given two images I_i and I_j, MEt3R first applies DUSt3R [34] to infer a dense shared 3D reconstruction from the image pair. This reconstruction is used to warp one image into the viewpoint of the other, after which the aligned views are compared in feature space rather than pixel space. Abstractly, the metric can be written as

$$\mathrm{MEt3R}(I_i, I_j) = D(\phi(W_{i\to j}(I_i))\,,\, \phi(I_j))\,, \tag{20}$$

where $W_{i\to j}(\cdot)$ denotes geometry-aware warping, $\phi(\cdot)$ denotes the feature extractor, and $D(\cdot,\cdot)$ denotes the feature-space discrepancy used by the metric. Because it evaluates consistency after geometry-aware alignment, MEt3R is well-suited for measuring whether generated observations remain coherent under viewpoint change. Lower values indicate stronger multi-view consistency.

Absolute Trajectory Error (ATE). ATE measures the overall deviation between the predicted trajectory and the ground-truth trajectory over the full horizon. Let $\hat{\mathbf{p}}_t$ and $\mathbf{p}_t$ denote the predicted and ground-truth positions at time step t, respectively. We compute ATE as

$$\mathrm{ATE} = \frac{1}{T}\sum_{t=1}^{T} \left\|\hat{\mathbf{p}}_t - \mathbf{p}_t\right\|_2. \tag{21}$$

Lower ATE indicates that the predicted trajectory remains closer to the ground-truth path throughout the horizon.

Final Displacement Error (FDE). FDE measures the endpoint error of the predicted trajectory. It is defined as the Euclidean distance between the predicted final position and the ground-truth final position:

$$\mathrm{FDE} = \left\|\hat{\mathbf{p}}_T - \mathbf{p}_T\right\|_2. \tag{22}$$

Lower FDE indicates better accuracy at the final destination.

Relative Pose Error (RPE). RPE measures the local consistency of the predicted trajectory by comparing the relative motion between consecutive steps. Let $\Delta\hat{\mathbf{p}}_t = \hat{\mathbf{p}}_t - \hat{\mathbf{p}}_{t-1}$ and $\Delta\mathbf{p}_t = \mathbf{p}_t - \mathbf{p}_{t-1}$ denote the predicted and ground-truth relative displacements, respectively. We define RPE as

$$\mathrm{RPE} = \frac{1}{T-1}\sum_{t=2}^{T} \left\|\Delta\hat{\mathbf{p}}_t - \Delta\mathbf{p}_t\right\|_2. \tag{23}$$

Lower RPE indicates better preservation of local motion consistency along the trajectory.

D Additional Ablations

Effect on downstream planning. Finally, we analyze when the improved world model is most beneficial for downstream planning. table 9 studies the effect of the number of sampled trajectories. Our method consistently improves ATE over NWM, with the largest gains appearing when the number of samples is limited. As the number of samples increases, the gap gradually narrows, suggesting that better future prediction is especially helpful in low-sample regimes where the planner cannot rely on exhaustive search. When the number of samples becomes very large, the planner already covers a wide range of possible trajectories, reducing the relative importance of improved ranking from the world model.

Table 9: Effect of the number of samples on ATE. Lower is better.

Samples	10	20	30	60
CEM + NWM	3.93	3.28	3.17	2.97
CEM + Ours	**3.69**	**3.06**	**2.78**	**2.71**

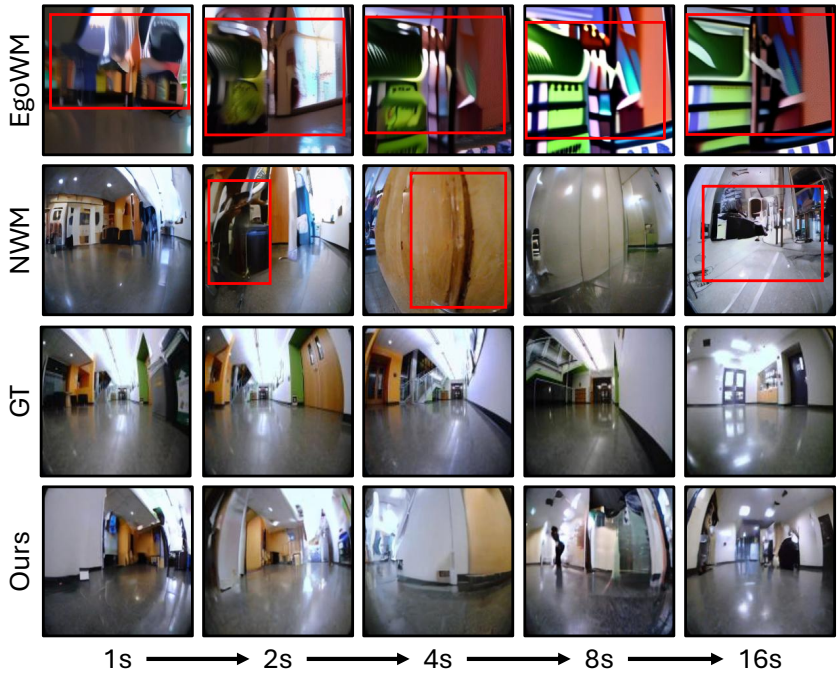


Figure 8: **Qualitative results over time on HuRoN.** As the prediction horizon increases, NWM and EgoWM exhibit stronger blur, drift, and structural degradation, while our method preserves clearer layouts and more stable object appearance over longer rollouts. Red frames highlight regions where the baselines become visibly corrupted or difficult to interpret.

E Additional Qualitative Results

E.1 Perceptual Drift

We provide additional qualitative comparisons on HuRoN (fig. 8), SCAND (fig. 9), and TartanDrive (fig. 10) to illustrate how prediction quality evolves as the rollout horizon increases.

E.2 Geometry Drift

We further visualize geometry drift through additional MEt3R in fig. 11 and fig. 12, which highlight the difference in multi-view consistency between our method and the baseline under viewpoint changes.

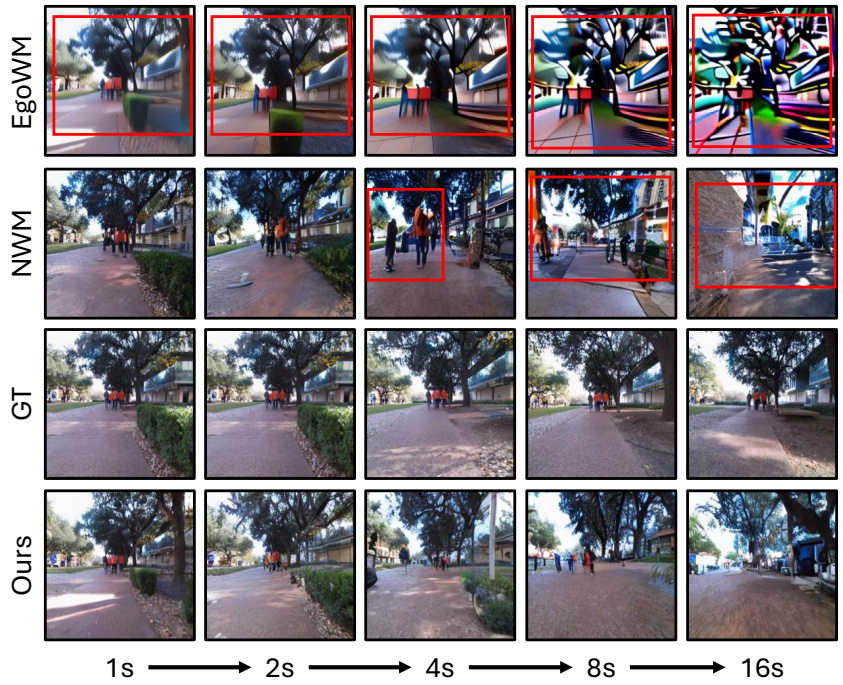


Figure 9: **Qualitative results over time on SCAND.** Compared with the baselines, our method better preserves scene structure and viewpoint consistency as the rollout extends, showing reduced perceptual drift over long horizons. Red frames highlight regions where the baselines become visibly corrupted or difficult to interpret.

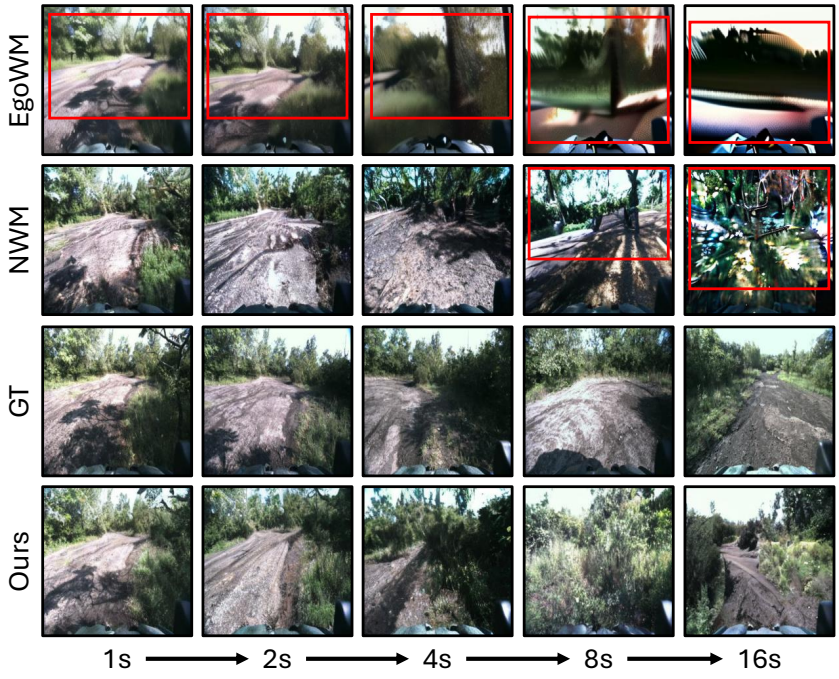


Figure 10: **Qualitative results over time on TartanDrive.** In this more challenging driving setting, our method maintains more stable long-horizon predictions, whereas the baselines deteriorate more noticeably as the horizon increases. Red frames highlight regions where the baselines become visibly corrupted or difficult to interpret.

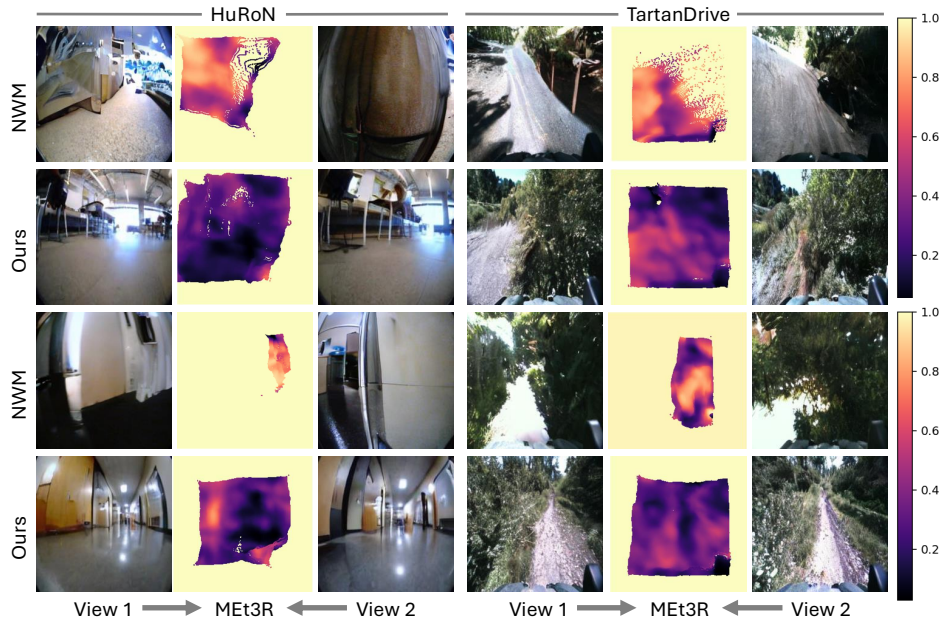


Figure 11: **Qualitative results of MEt3R on HuRoN and TartanDrive.** We visualize multi-view inconsistency measured by MEt3R after geometry-aware alignment. Darker responses indicate lower inconsistency and therefore stronger cross-view coherence. Our method preserves more coherent scene structure and viewpoint-consistent content.



Figure 12: **Qualitative results of MEt3R on RECON and SCAND.** Across different scenes, our method produces more consistent cross-view content after alignment, while the baseline shows stronger geometric inconsistency and structural mismatch.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We accurately reflect the contributions and scope in the abstract and introduction.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We have addressed the limitations in the appendix.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: This paper provides the full set of assumptions and a complete proof for theoretical results.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We have shown the implementation details and setups in the main paper and the appendix. The checkpoints would also be released.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: All the data are acquired from open-sourced platforms. The code is not accessible currently. We will make the codebase open source upon publication.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We have presented the training and test details in the paper and the appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: There is no experimental result related to error bars in this paper.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We report the used GPUs in the paper.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in the paper conforms with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss broader impacts in the Appendix.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: This paper does not have such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All external assets, including data, code, and models, are credited to their original owners, with their licenses and usage terms clearly specified and fully honored.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: All new assets are well documented and provided alongside the assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This study does not include human subjects

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.