CAUSALALIGN: CAUSAL ALIGNMENT OF VIDEO GENERATION MODELS FOR CONSISTENT EGO-MOTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Diffusion-based video generation models (VGMs) can produce visually plausible videos, yet their camera ego-motion often drifts over long horizons. Camera ego-motion determines how the viewpoint changes between frames and is essential for maintaining consistent geometry and trajectories. To address this problem, we propose **CausalAlign**, a causal alignment framework with two components. First, we diagnose how precise ego-motion is represented, used, and retained within VGMs using linear probes and causal interventions, revealing a rise-and-fall pattern during denoising. Second, we construct a delta-token space from a 3D foundation model and align VGM representations at causally identified positions. Delta tokens encode changes between consecutive frames, allowing the alignment to focus on ego-motion rather than static scene content. With Wan2.2 fine-tuned on DL3DV and RealEstate10K, CausalAlign preserves visual quality while improving motion accuracy, geometric consistency, and temporal smoothness. For example, it reduces geometric consistency errors by more than **20%** relative to the base model. Moreover, using the fine-tuned VGM, CausalAlign improves downstream world action models (WAMs) on the RoboCasa365 benchmark. Overall, CausalAlign improves ego-motion consistency in diffusion-based VGMs and yields gains in downstream embodied models on motion-intensive tasks.

1 INTRODUCTION

Recent advances in diffusion-based video generation models (VGMs) have led to remarkable models such as Wan2.2 (Team Wan et al., 2025) and Cosmos3 (NVIDIA Cosmos Team, 2026). Trained on Internet-scale collections of 2D videos, these models can synthesize visually compelling content with high perceptual quality (Huang et al., 2024; 2026). However, trained primarily to model 2D appearance and local motion statistics, standard VGMs struggle to capture global 3D camera trajectories and to control viewpoint changes consistently across frames (Wu et al., 2026; Kupyn et al., 2025). We refer to these consistent cross-frame viewpoint changes as camera ego-motion.

As shown in Figure 1 (left panel), prior work has attempted to address this challenge by introducing explicit camera controls or 3D priors (Wang et al., 2024b; Ren et al., 2025; Zhang et al., 2025; Bahmani et al., 2025), yet generating videos with stable and geometrically consistent camera trajectories remains an open problem. At the same time, recent work shows that VGMs can implicitly estimate camera poses from their internal representations, even when their generated trajectories are imperfect (Bahmani et al., 2025). This raises a more fundamental question: **how is precise ego-motion represented, used, and retained within VGMs?** This question becomes increasingly important when VGMs are integrated into world action models (WAMs) for robotic mobile manipulation: visually plausible yet geometrically inconsistent predictions of the future can induce systematic errors in downstream action prediction.

As the first step, we perform analysis for understanding the ego-motion information within VGMs. Inspired by prior works (Alain & Bengio, 2016; Meng et al., 2022), we adopt linear probing to measure the internal representations of VGMs and causal intervention to reveal the mechanisms of VGMs. Specifically, for linear probing, we define "decodability" as the amount of ego-motion information. Meanwhile, for causal intervention, we define "causal effect" as the extent to which

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

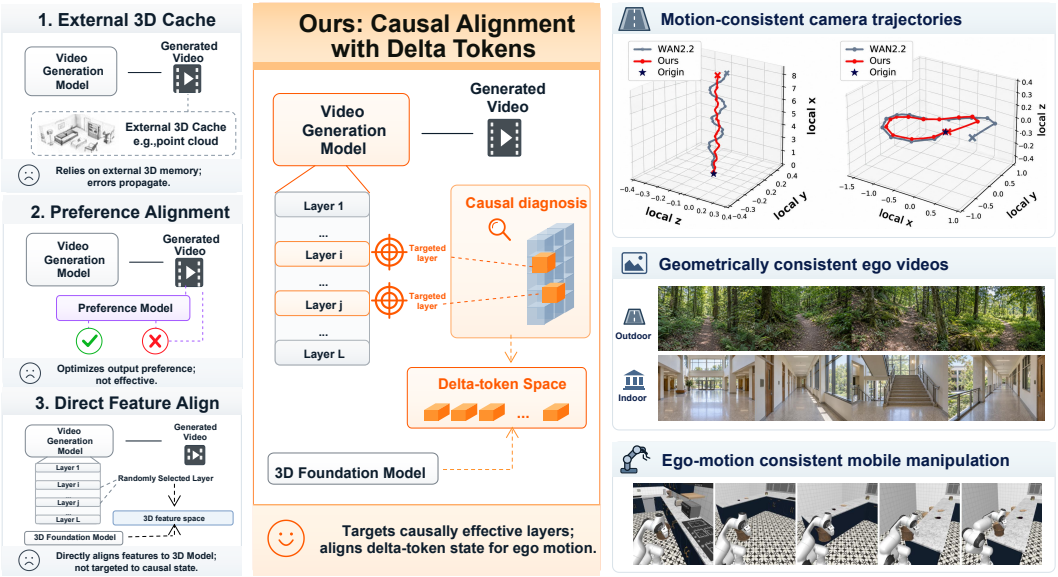

Figure 1: **Overview of CausalAlign.** Left: **CausalAlign** combines causal diagnosis and delta-token alignment. Right: **CausalAlign** largely improves ego-motion consistency.

interventions on VGM change the generation. Our analysis results demonstrate that both measurements exhibit the same trend, that is, increasing first and then declining, across both denoising timesteps and denoising Transformer layers. These results suggest that ego-motion state emerges as the denoising progresses, however, fades away at the late timesteps and the last few Transformer layers at each step. Therefore, ego-motion is already learned and represented in VGMs, increasingly used at intermediate stages, yet not well retained at later stages. We refer to this progressive pattern as **"geometric state forgetting"**.

Motivated by this diagnosis, our goal is to preserve and strengthen the ego-motion state within VGMs without compromising the quality of visual generation. To this end, we propose **CausalAlign**, a framework with two components: causal diagnosis and delta-token alignment, as shown in Figure 1 (middle panel). First, causal diagnosis leverages linear probing and causal intervention to identify where ego-motion is represented, used, and forgotten within a VGM. Second, delta-token alignment supervises the internal ego-motion state at the positions identified as forgetting. There are two key challenges in our alignment: *how to represent ego-motion with the hidden states in VGMs*, and *how to provide meaningful geometry supervision for such representations*. For the first challenge, we adopt the idea of delta tokens, which was initially designed for improving the efficiency of world model (Kerssies et al., 2026). We adapt this idea to measure the ego-motion via the difference between two consecutive frames. Delta tokens serve as a natural carrier for motion information while reducing redundant static scene content. For the second challenge, we construct the delta-token space by leveraging the pretrained 3D foundation model. These 3D models are capable of estimating camera poses, depth, and 3D structure, and thus can provide high-quality geometric supervision. A learnable projector then maps consecutive hidden-state transitions into this teacher delta-token space, explicitly supervising the VGM's latent ego-motion state. By aligning these projected tokens with the corresponding teacher delta tokens at the identified layers, CausalAlign transfers geometric knowledge to the locations where ego-motion information starts to fade.

With our causal diagnosis and delta-token alignment method, we fine-tune the VGM on the DL3DV (Ling et al., 2024) and RealEstate10K (Zhou et al., 2018) datasets. To evaluate what CausalAlign can improve over VGMs, we establish a comprehensive evaluation protocol with three aspects: visual quality, ego-motion consistency, and geometric state retention. Visual quality is measured with VBench-I2V (Huang et al., 2026) and complementary metrics. Ego-motion consistency is evaluated through motion magnitude, geometric consistency, and temporal dynamics, while geometric-state retention is assessed through probe decodability and causal intervention effects.

CausalAlign preserves visual quality while improving camera ego-motion consistency, as shown in Figure 1 (right panel). For example, it reduces reprojection error by 22.6% and improves trajectory smoothness by 23.9% relative to the base model. It also shifts the decodability and causal-effect patterns from *rise-and-fall* to *rise-and-maintain*, supporting improved retention of the internal ego-motion state. When integrated into a WAM, the adapted VGM further increases the success rate on RoboCasa365 (Nasiriany et al., 2026) from 43.5% to 54.2%.

Our contributions are summarized as follows: (1) We identify a pattern called **"geometric state forgetting"** which reveals the ego-motion state in diffusion-based VGMs. Through linear probing and causal intervention, our analysis reveals that ego-motion information is progressively represented and used during denoising process, yet not retained at late stages of denoising. (2) We propose **CausalAlign**, a causal alignment method that aims to preserve the ego-motion state at the VGM layers identified by our causal analysis. Our technical innovation lies in adapting the idea of delta tokens to strengthen the ego-motion via 3D foundation models. (3) Extensive experiments and comprehensive metrics demonstrate the benefits of our method, which largely improves ego-motion consistency, well preserves ego-motion state during denoising, and offers remarkable gains for WAM in motion-intensive robotic tasks.

2 Related Work

Video diffusion and world action models. Video diffusion extends image denoising to joint spatiotemporal generation (Ho et al., 2022), while open backbones such as Wan have broadened its use (Team Wan et al., 2025). Unlike image generation, VGMs must preserve temporal correspondences and scene geometry across multiple frames (Huang et al., 2024; 2026; Ren et al., 2025; Zhang et al., 2025). These priors increasingly serve as action-conditioned simulators or policies: Genie and IRASim learn controllable future dynamics, while VideoVLA, DreamZero, and LingBot-VA couple observations with executable actions; Fast-WAM shows that video co-training can improve action representations without test-time video decoding (Bruce et al., 2024; Zhu et al., 2025; Shen et al., 2025; Ye et al., 2026; Li et al., 2026a; Yuan et al., 2026). Thus, ego-motion consistency affects WAM performance: erroneous camera motion can corrupt predicted states and actions. Rather than proposing another WAM architecture, CausalAlign restores and strengthens the internal ego-motion capability of the VGM to improve WAM performance. This leverages capabilities that emerge naturally during VGM pretraining and avoids the need for expensive action-annotated data.

Feed-forward 3D foundation models. Feed-forward 3D reconstruction has progressed from point-map predictors such as DUSt3R and VGGT (Wang et al., 2024a; 2025a) to streaming models with persistent temporal state. CUT3R recurrently maintains a metric-scale scene representation (Wang et al., 2025b), while LingBot-Map organizes streaming context into anchor, pose-reference, and trajectory memories and reports improved trajectory and pose accuracy over offline VGGT (Chen et al., 2026a). These models provide camera and scene representations that can serve as supervision for video generation. We therefore use LingBot-Map as the teacher and study how its recovered ego-motion state can be internalized and causally retained by a video generator.

Geometry-aware alignment and post-training. As summarized in Fig. 1, existing geometry-aware adaptation methods follow three main routes. External 3D-cache methods such as GEN3C and Voyager use point-cloud or world memories for controllable camera motion, but require additional geometric memory and may propagate cache errors (Ren et al., 2025; Huang et al., 2025b). Preference-alignment methods such as VideoGPA and Epipolar-DPO supervise generated videos through preference or epipolar feedback, without directly modifying the VGM's internal geometric state (Du et al., 2026; Kupyn et al., 2025). Direct alignment methods, including JOG3R, Geometry Forcing, and FantasyWorld, transfer geometric features into the generator but do not explicitly target causally effective layer–timestep locations (Huang et al., 2025a; Wu et al., 2026; Dai et al., 2026). Complementary WAM methods use training-only 3D/4D or action-grounded supervision to improve action representations without requiring auxiliary teachers at deployment (Li et al., 2026b; Zhang et al., 2026; Yang et al., 2026; Fu et al., 2026; Qiu et al., 2026). CausalAlign instead targets the VGM's internal ego-motion state using causal diagnosis and motion-centric delta-token alignment, without requiring test-time geometric memory.

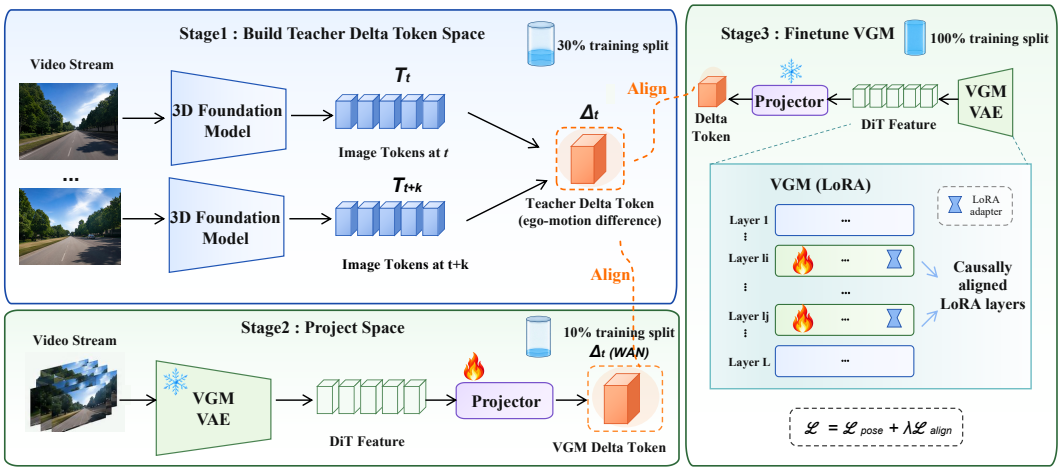

Figure 2: **Three stages of CausalAlign training.** CausalAlign transfers ego-motion supervision from a 3D teacher to the VGM through three-stage training. In Stage 1, consecutive features from a streaming 3D foundation model are encoded into a teacher delta-token space that captures inter-frame geometric transitions. In Stage 2, a projector maps VGM DiT features into the same delta-token space and is frozen after training, using 10% of the training split. In Stage 3, using the full training split, LoRA adapters are optimized only at causally targeted VGM layers using delta tokens, while the original VGM backbone, teacher model, and projector remain frozen.

3 METHOD

3.1 METHOD OVERVIEW

CausalAlign combines causal diagnosis and delta-token alignment to preserve ego-motion representations in diffusion-based VGMs. Probes and interventions identify layers and denoising timesteps where ego-motion information fades, while avoiding alignment at locations that are not implicated by the diagnosis. Alignment then proceeds through three stages: learning delta tokens from a frozen 3D foundation model, training projectors from frozen VGM transitions into this space, and updating LoRA adapters at the selected layers while freezing the teacher modules and projectors. Supervision is applied at the selected timesteps, and the original VGM backbone is otherwise kept fixed. Only the adapted VGM is retained at inference, so the teacher and auxiliary modules introduce no test-time computation.

3.2 CAUSAL DIAGNOSIS OF EGO-MOTION STATES

Decodability. Let $\mathbf{h}_t^{\ell,s}$ denote the spatially aggregated VGM state at frame t, layer ℓ, and noise-level index s. A shared operator Γ constructs the consecutive-frame transition feature

$$\mathbf{u}_{t,\ell,s} = \Gamma\left(\mathbf{h}_t^{\ell,s}, \mathbf{h}_{t+1}^{\ell,s}\right).$$

A linear probe predicts the relative camera motion

$$\widehat{\boldsymbol{\xi}}_{t,\ell,s} = \mathbf{W}_{\ell,s}\mathbf{u}_{t,\ell,s} + \mathbf{b}_{\ell,s}, \qquad \boldsymbol{\xi}_t = \mathrm{Log}\left(\mathbf{T}_t^{-1}\mathbf{T}_{t+1}\right), \tag{1}$$

where $\mathbf{T}_t \in \mathrm{SE}(3)$ is the ground-truth camera-to-world pose and Log maps the relative pose to a six-dimensional motion vector. The held-out score $D_{\ell,s}$ measures how well ego-motion is decodable at location (ℓ, s); decodability alone does not establish causal use (Meng et al., 2022; Geiger et al., 2021).

Causal intervention and target selection. At $c = (\ell, s)$, we keep the prompt, initial noise, and random seed fixed and intervene on the temporal hidden state using

$$\mathcal{A} = \{\textsc{Mask}, \textsc{Shuffle}, \textsc{Replace}, \textsc{Inject}\}.$$

These interventions remove temporal information, disrupt its order, replace it with another trajectory, or insert a pose-matched donor, respectively.

For diagnostic clip d, let $\mathbf{T}_{1:T}^{\mathrm{ref},d}$ be its ground-truth camera-to-world trajectory from the corresponding real video. A frozen pose estimator produces $\widehat{\mathbf{T}}(Y)$ from a generated video Y, and we define

$$E_{\mathrm{traj}}(Y;d) = \mathcal{E}_{\mathrm{traj}}\Big(\widehat{\mathbf{T}}(Y), \mathbf{T}_{1:T}^{\mathrm{ref},d}\Big). \tag{2}$$

The reference trajectory is used only for evaluation and is not provided as a generation condition. For intervention $a \in \mathcal{A}$, the norm-matched random-control-adjusted effect is

$$\overline{\mathcal{C}}_c^{(a)} = \eta_a \, \mathbb{E}_d \left[E_{\mathrm{traj}}\Big(Y_c^{(a)}; d\Big) - E_{\mathrm{traj}}\Big(Y_c^{(\mathrm{rand},a)}; d\Big) \right], \tag{3}$$

where $\eta_a = +1$ for the disruptive interventions and $\eta_a = -1$ for INJECT, so that larger values indicate stronger causal effects. The aggregate causal score is

$$C_{\ell,s} = \frac{1}{|\mathcal{A}|} \sum_{a \in \mathcal{A}} \overline{\mathcal{C}}_{(\ell,s)}^{(a)}. \tag{4}$$

Averaging $D_{\ell,s}$ and $C_{\ell,s}$ over the complementary axis gives layer profiles ($A = \mathrm{L}$) and timestep profiles ($A = \mathrm{S}$). We combine them as

$$J^A = \frac{1}{2}\left[\mathcal{N}(D^A) + \mathcal{N}(C^A)\right], \qquad A \in \{\mathrm{L}, \mathrm{S}\}, \tag{5}$$

where $\mathcal{N}$ independently min–max normalizes each profile. The target sets $\mathcal{L}_{\mathrm{tar}}$ and $\mathcal{S}_{\mathrm{tar}}$ are selected at locations with the largest positive decline rates of J^A. The corresponding normalized timestep declines provide the weights $\omega_{\ell,s}$ used by the alignment losses and shared across selected layers.

3.3 DELTA-TOKEN ALIGNMENT

Stage 1: learning teacher delta tokens. For a T-frame training video $\mathbf{x}_{1:T}$, the frozen 3D foundation model Φ_{T} extracts geometric states $\mathbf{z}_t^{\mathrm{T}} = \Phi_{\mathrm{T}}(\mathbf{x}_{1:t})$. A trainable encoder E_Δ maps consecutive states to a delta token:

$$\boldsymbol{\delta}_t^{\mathrm{T}} = E_\Delta(\mathbf{z}_t^{\mathrm{T}}, \mathbf{z}_{t+1}^{\mathrm{T}}) \in \mathbb{R}^{d_\Delta}, \tag{6}$$

where d_Δ is the token dimension; superscripts T and V denote teacher and VGM representations. A transition decoder reconstructs the next state from the current state and token, a temporal predictor forecasts the next token, and a camera decoder D_{cam} predicts relative pose. These auxiliary tasks encourage the token to retain transition information rather than merely compressing the teacher state. Denoting their predictions by $\widehat{\mathbf{z}}_{t+1}^{\mathrm{T}}$, $\widehat{\boldsymbol{\delta}}_{t+1}^{\mathrm{T}}$, and $\widehat{\Delta\mathbf{T}}_t^{\mathrm{T}}$, we train the encoder and auxiliary modules with

$$\begin{aligned}\mathcal{L}_{\mathrm{T}} = \mathbb{E}_t\Big[&\|\widehat{\mathbf{z}}_{t+1}^{\mathrm{T}} - \mathbf{z}_{t+1}^{\mathrm{T}}\|_2^2 + \lambda_{\mathrm{temp}} \|\widehat{\boldsymbol{\delta}}_{t+1}^{\mathrm{T}} - \boldsymbol{\delta}_{t+1}^{\mathrm{T}}\|_2^2 \Big] \\ &+ \lambda_{\mathrm{pose}}^{\mathrm{T}} \mathbb{E}_t\Big[\ell_{\mathrm{pose}}(\widehat{\Delta\mathbf{T}}_t^{\mathrm{T}}, \Delta\mathbf{T}_t) \Big].\end{aligned} \tag{7}$$

Here, $\Delta\mathbf{T}_t = \mathbf{T}_t^{-1}\mathbf{T}_{t+1}$ is the ground-truth relative pose used only for supervision, and ℓ_{pose} combines geodesic rotation distance with normalized SmoothL1 translation error. The expectation averages valid training-frame positions; λ_{temp} and $\lambda_{\mathrm{pose}}^{\mathrm{T}}$ weight temporal and pose supervision. These objectives preserve transition structure and anchor tokens to camera motion. We then freeze the encoder and auxiliary modules; decoder architectures and pose-loss details are provided in the appendix.

Stage 2: learning VGM-side projectors. With the VGM and teacher modules frozen, layer-specific projectors Q_ℓ map corresponding VGM frame-pair features to teacher-space tokens $\boldsymbol{\delta}_{t,\ell,s}^{\mathrm{V}} = Q_\ell(\mathbf{u}_{t,\ell,s})$. We minimize

$$\mathcal{L}_{\mathrm{map}} = \sum_{\ell \in \mathcal{L}_{\mathrm{tar}}} \sum_{s \in \mathcal{S}_{\mathrm{tar}}} \omega_{\ell,s} \mathbb{E}_t\left[\|\boldsymbol{\delta}_{t,\ell,s}^{\mathrm{V}} - \boldsymbol{\delta}_t^{\mathrm{T}}\|_2^2\right], \tag{8}$$

where $\omega_{\ell,s}$ weights selected timesteps by their normalized positive diagnostic-score decline rates, shared across selected layers. The projectors are frozen after training.

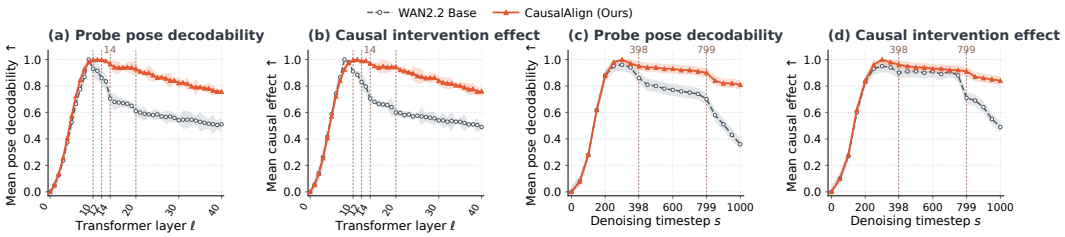

Figure 3: **Causal diagnosis of ego-motion states.** CausalAlign measures ego-motion decodability and causal intervention effect in transformer layers (a,b) and denoising timesteps (c,d). From left to right, the first two panels show probe ego-motion decodability and causal intervention effect across transformer layers, averaged over denoising timesteps; the last two panels show the corresponding measurements across denoising timesteps, averaged over layers. Dashed lines mark the selected target layers and denoising timesteps. Shading indicates the 10th–90th percentile range across the dimension being averaged.

Stage 3: adapting the VGM. We update LoRA adapters (Hu et al., 2022) at the selected layers. For an adapted transition $\widetilde{\mathbf{u}}_{t,\ell,s}$, the frozen projector produces $\widetilde{\boldsymbol{\delta}}^{\mathrm{V}}_{t,\ell,s} = Q_\ell(\widetilde{\mathbf{u}}_{t,\ell,s})$. We use two complementary objectives:

$$\mathcal{L}_\Delta = \sum_{\ell\in\mathcal{L}_{\mathrm{tar}}} \sum_{s\in\mathcal{S}_{\mathrm{tar}}} \omega_{\ell,s}\, \mathbb{E}_t \left[\left\| \widetilde{\boldsymbol{\delta}}^{\mathrm{V}}_{t,\ell,s} - \boldsymbol{\delta}^{\mathrm{T}}_t \right\|_2^2 \right], \tag{9}$$

which aligns the adapted VGM transitions with the teacher delta tokens, and

$$\mathcal{L}_{\mathrm{pose}} = \sum_{\ell\in\mathcal{L}_{\mathrm{tar}}} \sum_{s\in\mathcal{S}_{\mathrm{tar}}} \omega_{\ell,s}\, \mathbb{E}_t \left[\ell_{\mathrm{pose}} \left(D_{\mathrm{cam}} \left(\widetilde{\boldsymbol{\delta}}^{\mathrm{V}}_{t,\ell,s} \right), \Delta\mathbf{T}_t \right) \right], \tag{10}$$

which explicitly supervises the predicted relative camera motion. The final adaptation objective is

$$\mathcal{L}_{\mathrm{CausalAlign}} = \mathcal{L}_{\mathrm{pose}} + \lambda_\Delta \mathcal{L}_\Delta, \tag{11}$$

where λ_Δ controls the strength of token alignment. Gradients pass through the frozen projectors and camera decoder to the LoRA adapters, while all other parameters remain frozen. The teacher is used only to provide training targets, and no external geometric cache is required during generation.

3.4 Extension to World–Action Models

We transfer CausalAlign to Fast-WAM (Yuan et al., 2026) by initializing the video expert with the adapted Wan2.2 backbone and deriving ActionDiT from native Wan using Fast-WAM's parameter transformation. Downstream policy learning retains the video and action flow-matching objectives:

$$\mathcal{L}_{\mathrm{WAM}} = \mathcal{L}^v_{\mathrm{FM}} + \mathcal{L}^a_{\mathrm{FM}}. \tag{12}$$

Here, v and a denote the video and action experts. The three camera views are synchronized at each observation time, and action prediction starts at the current time. Future video remains a training target; inference uses the observed history to predict executable actions.

4 Experiment

4.1 Causal Diagnosis and Target Selection

Setup. We freeze Wan2.2 (Team Wan et al., 2025) and examine its hidden states across transformer layers and denoising steps. At each location $c = (\ell, s)$, a linear readout predicts relative camera motion, yielding a decodability score $\mathcal{D}_{\ell,s}$. We also apply the causal interventions introduced in Sec. 3.2 and measure their effects on the generated camera trajectory, yielding a causal effect score $\mathcal{C}_{\ell,s}$. Figure 3 summarizes the resulting layer–timestep profiles.

Diagnosis and target selection. Both geometric decodability and causal effect exhibit a consistent rise-and-fall pattern. Their increase at intermediate locations indicates that the VGM forms and uses

Table 1: CausalAlign maintains visual quality comparable to Wan2.2 across VBench-I2V and complementary visual metrics. Bold marks the best result per metric.

	VBench-I2V				Other Visual Metrics			
	Quality↑	I2V Score↑	Smooth.↑		FVD↓	CLIP↑	SSIM↑	LPIPS↓
Wan2.2	0.805	**0.958**	0.983	Wan2.2	**182.4**	0.311	0.742	**0.183**
CausalAlign	**0.806**	**0.958**	**0.985**	CausalAlign	182.7	**0.314**	**0.744**	**0.183**

an ego-motion state, while their subsequent decline suggests that this state is not retained during later denoising. Applying the decline-based target-selection rule from Sec. 3.2, we select layers 14, 10, 12, and 20 in descending order of their layer-wise decline rates. Layer 14 is the primary target, while layers 10, 12, and 20 are auxiliary targets. The same rule selects denoising timesteps $s \in \{398, 799\}$. Both timesteps are used at all selected layers, with weights proportional to their positive temporal decline rates.

4.2 Experiments on the Comprehensive Evaluation Protocol

Setup. We instantiate CausalAlign on Wan2.2-14B-I2V (Team Wan et al., 2025) using DL3DV (Ling et al., 2024) and RealEstate10K (Zhou et al., 2018), with each dataset split into 90% training and 10% held-out evaluation. Following the three-stage procedure in Fig. 2, we train the teacher delta-token modules, VGM-side projectors, and LoRA adapters on 30%, 10%, and all of the training data, respectively. Training details and hyperparameters are provided in Appendix B.2.

Visual Quality Preservation. We first evaluate whether the proposed geometric supervision preserves the visual quality of the base model. CausalAlign and Wan2.2 are evaluated with identical prompts, sampling settings, and random seeds. The left part of Table 1 reports the VBench-I2V (Huang et al., 2024; 2026) results, while the right part reports additional visual metrics (Unterthiner et al., 2018; Wang et al., 2004; Zhang et al., 2018; Radford et al., 2021) using the same evaluation method in ego-motion consistency part. CausalAlign matches the base model across all metrics, showing that the proposed alignment does not degrade the visual generation quality.

Ego-Motion Consistency. We compare CausalAlign with the original Wan2.2-14B-I2V model and representative baselines from the three categories introduced in Related Work. For each held-out clip from DL3DV and RealEstate10K, methods without explicit camera control receive only the first frame and text prompt, whereas trajectory-controlled baselines receive the camera trajectory and the freshly constructed 3D cache required by their native implementations. We use identical output length, resolution, frame rate, sampling settings, and random seeds, and process all generated videos with the same frozen pose, depth, optical-flow, and feature estimator (Wang et al., 2025a).

We group the metrics into three categories. *Motion Magnitude* measures mean translational and angular camera speed (Appendix B.1), and the resulting values are compared with the real-video validation distribution. *Geometric Consistency* includes reprojection error (Triggs et al., 2000), depth–flow consistency error on static background regions (Luo et al., 2020), and epipolar error (Lee & Civera, 2020). *Temporal Dynamics* measures trajectory smoothness and jerk, capturing motion continuity, high-frequency fluctuations, and abrupt acceleration changes (Hogan & Sternad, 2009; Flash & Hogan, 1985). We report means over multiple generated samples per input.

CausalAlign produces motion magnitudes closest to the real-video reference, fewer high-frequency and stop–start fluctuations, and lower geometric errors. It also achieves the highest smoothness and lowest jerk among the compared methods. These results indicate that causal delta-token alignment improves the intrinsic ego-motion prior and geometric self-consistency of generated videos without requiring an explicit camera trajectory at inference time. Figure 4 provides qualitative comparisons with Wan2.2, where CausalAlign shows more consistent ego-motion.

Geometric state Retention. We assess whether CausalAlign preserves the internal geometric state throughout denoising. Figure 3 shows that probe decodability and causal intervention effects in the Wan2.2 base model increase at intermediate stages and decline later, forming a *rise-and-fall* pattern. After delta-token alignment, both signals remain elevated during later denoising and exhibit a *rise-and-maintain* pattern, indicating improved retention of ego-motion information.

Table 2: **Comparison of camera ego-motion.** CausalAlign improves camera-motion consistency over Wan2.2, with the best reported results on six of seven motion, geometry, and temporal metrics. Methods without explicit camera control receive only the first frame and text prompt, whereas trajectory-controlled baselines receive the camera trajectory and the freshly constructed 3D cache required by their native implementations.

Method	Motion Magnitude		Geometric Consistency			Temporal Dynamics	
	Mean Trans. Prox. ↑	Mean Angular Prox. ↑	Reproj. (px)↓	Depth–Flow (px)↓	Epipolar (px)↓	Smoothness ↑	Jerk ↓
(a) Base Model							
Wan2.2	0.68	0.70	8.4	6.9	2.9	0.71	0.182
(b) External 3D Cache							
GEN3C	0.63	0.75	7.8	6.5	2.8	0.71	0.175
Voyager	0.69	0.73	7.1	6.8	2.8	0.73	0.179
(c) Preference Alignment							
VideoGPA	0.66	0.78	8.4	6.7	2.5	0.73	0.176
Epipolar-DPO	0.68	0.72	7.8	5.9	2.2	0.78	0.162
(d) Direct Feature Align							
JOG3R	0.70	0.75	7.2	5.9	2.5	0.79	0.158
Geometry Forcing	0.69	0.78	6.9	5.7	2.4	0.80	0.164
FantasyWorld	0.72	0.77	7.0	**5.4**	2.3	0.82	0.159
(e) Causal Delta-Token Alignment							
CausalAlign (Ours)	**0.79**	**0.89**	**6.5**	5.5	**2.1**	**0.88**	**0.151**

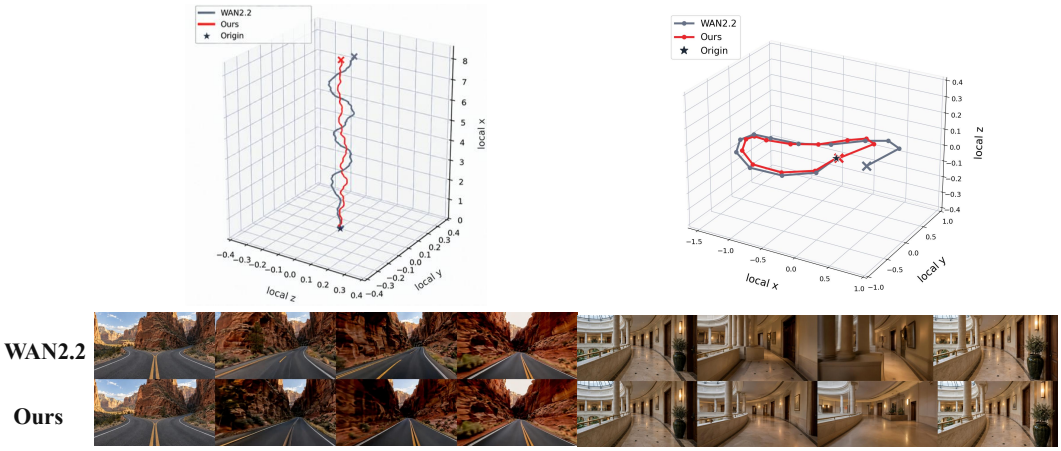

Figure 4: **Qualitative comparison with Wan2.2.** CausalAlign outperforms the base model qualitatively. On the left, CausalAlign produces fewer trajectory jitters. On the right, the intended camera motion is a loop trajectory. CausalAlign ends closer to the starting point, indicating smaller loop-closure drift and improved trajectory consistency.

To further validate the improved internal state and test whether it benefits downstream action prediction and execution, we transfer the fine-tuned video expert to Fast-WAM. We evaluate on the Atomic-Seen split of RoboCasa365 Target100 (Nasiriany et al., 2026), which contains 18 tasks in target kitchens with 50 rollouts per task. The policy is pretrained on Human300 and fine-tuned with 500 target demonstrations per task. Using the 50k fine-tuning checkpoint, the policy receives images from three cameras, and the current proprioceptive state, predicts 32 actions with ten denoising steps, and executes eight actions before replanning at 20 Hz. CausalAlign + Fast-WAM succeeds in 703 of 900 rollouts, achieving **78.1%** success (Table 3), exceeding the Fast-WAM and ABot-M0.5 scores reported in Table 3 of Chen et al. (2026b) by 19.01 and 7.51 percentage points, respectively. Appendix B.5 shows representative successful mobile-manipulation rollouts.

Table 3: CausalAlign improves Fast-WAM success rates across all three RoboCasa365 Target100 task groups, with the largest gain on Atomic-Seen (+19.0 percentage points). Success rates (%).

Method	Atomic Seen	Composite Seen	Composite Unseen	Average
GR00T-N1.5	60.6	35.0	33.3	43.7
Fast-WAM	59.1	36.4	33.2	43.5
LingBot-VA	63.5	37.3	32.1	45.1
ABot-M0.5	70.6	**44.3**	**45.6**	**54.2**
CausalAlign + Fast-WAM	**78.1**	40.8	40.6	**54.2**

Table 4: Delta-token ablations. Semantic partitioning improves motion, geometry, and appearance readouts (a), while 128D delta tokens outperform the original 4096D/8192D teacher features (b).

	Ego motion		Geometry		Visual	
Representation	Rotation (°)↓	Direction (°)↓	AbsRel↓	δ_1 ↑	SSIM↑	LPIPS↓
(a) Matched-width partitions						
Unpartitioned pair	2.453	8.582	0.2426	0.7280	0.2924	0.08816
Unpartitioned diff.	2.837	10.392	0.2301	0.7390	0.2971	0.08754
Delta regions	**2.244**	**7.565**	**0.2292**	**0.7457**	**0.3050**	**0.08643**
(b) Complete space versus original features						
Original A (8192D)	3.243	12.964	0.2698	0.6993	0.2838	0.08958
Original B (4096D)	4.276	13.451	0.2731	0.6947	0.2866	0.08958
Delta (128D)	**2.530**	**8.231**	**0.2314**	**0.7422**	**0.3028**	**0.08676**

4.3 Ablation Studies

Semantic partitioning. We compare the 64D ego-motion, 40D geometry, and 24D visual regions with training-only PCA projections of identically supervised, unpartitioned 128D pair and difference encoders. The delta regions achieve better means in all 12 paired metric comparisons, improving ego-motion, geometry, and visual readouts together, as shown in Table 4(a). Equal-width cross-probes in Appendix C.1 show that each region performs best on its designated task.

Complete-space comparison. We compare the complete 128D delta token with the original pooled LingBot-Map feature pair (A, 8192D) and its temporal difference (B, 4096D), without compressing A or B. Delta tokens achieve better mean rotation, translation-direction, depth, and appearance readouts on all six metrics, as shown in Table 4(b). This comparison evaluates the learned representation against frozen teacher features; the matched-supervision controls in panel (a) separately test the contribution of partitioning.

Alignment placement. Independently trained adapters at the layers selected through causal diagnosis achieve the largest composite gain among the evaluated layer groups. The complete layer-placement comparison and confidence intervals are reported in Appendix C.2.

5 Conclusion

We introduce **CausalAlign**, a causal alignment framework for ego-motion-consistent video generation. By combining linear probing with causal interventions, we show that video generation models form and use ego-motion state during denoising but do not retain it well at later stages, a pattern we call geometric state forgetting. CausalAlign targets the locations where geometric state forgetting is identified through linear probing and causal intervention, and aligns VGM transitions with a motion-centric delta-token space from a 3D foundation model. On Wan2.2, CausalAlign improves ego-motion consistency and temporal dynamics while preserving visual quality. It also changes the internal geometric state from rise-and-fall to rise-and-maintain and improves RoboCasa365 performance. We hope this work inspires further research on causal alignment with delta tokens for preserving geometric states in generative video models.

AI USE STATEMENT

Generative AI tools were used to assist with software implementation and debugging, experimental planning and analysis, translation, and manuscript editing. They also assisted in preparing visualization code and figure layouts. The authors take responsibility for the final content of this work, including its text, experimental claims, and artifacts, and for verifying AI-assisted outputs.

REPRODUCIBILITY STATEMENT

The Method and Experiments sections, together with the appendix, describe the proposed approach, datasets, training and evaluation settings, metrics, and ablation protocols. These descriptions provide the basis for reproducing the reported experiments. We will make the code and experiment configurations publicly available, together with data preparation instructions for the datasets used in this work.

REFERENCES

Alexandre Alahi, Kratarth Goel, Vignesh Ramanathan, Alexandre Robicquet, Li Fei-Fei, and Silvio Savarese. Social lstm: Human trajectory prediction in crowded spaces. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 961–971, 2016.

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016.

Alibaba PAI. Wan2.2-Fun-A14B-Control-Camera. Hugging Face model card, 2025. URL https://huggingface.co/alibaba-pai/Wan2.2-Fun-A14B-Control-Camera.

Sherwin Bahmani, Ivan Skorokhodov, Guocheng Qian, Aliaksandr Siarohin, Willi Menapace, Andrea Tagliasacchi, David B. Lindell, and Sergey Tulyakov. AC3D: Analyzing and improving 3d camera control in video diffusion transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22875–22889, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Bahmani_AC3D_Analyzing_and_Improving_3D_Camera_Control_in_Video_Diffusion_CVPR_2025_paper.html.

Jake Bruce, Michael D Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, Yusuf Aytar, Sarah Maria Elisabeth Bechtle, Feryal Behbahani, Stephanie C.Y. Chan, Nicolas Heess, Lucy Gonzalez, Simon Osindero, Sherjil Ozair, Scott Reed, Jingwei Zhang, Konrad Zolna, Jeff Clune, Nando De Freitas, Satinder Singh, and Tim Rocktäschel. Genie: Generative interactive environments. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 4603–4623. PMLR, 2024. URL https://proceedings.mlr.press/v235/bruce24a.html.

Lin-Zhuo Chen, Jian Gao, Shangzhan Zhang, Yihang Chen, Ka Leong Cheng, Yipengjing Sun, Liangxiao Hu, Nan Xue, Xing Zhu, Yujun Shen, Yao Yao, and Yinghao Xu. LingBot-Map: Geometric context transformer for streaming 3D reconstruction. *arXiv preprint arXiv:2604.14141*, 2026a. URL https://arxiv.org/abs/2604.14141.

Ronghan Chen, Yandan Yang, Zuojin Tang, Dongjie Huo, Tong Lin, et al. ABot-M0.5: Unified mobility-and-manipulation world action model. *arXiv preprint arXiv:2607.00678*, 2026b. URL https://arxiv.org/abs/2607.00678.

Yixiang Dai, Fan Jiang, Chiyu Wang, Mu Xu, and Yonggang Qi. Fantasyworld: Geometry-consistent world modeling via unified video and 3d prediction. In *International Conference on Learning Representations*, volume 2026, pp. 103603–103622, 2026.

Hongyang Du, Junjie Ye, Xiaoyan Cong, Runhao Li, Jingcheng Ni, Aman Agarwal, Zeqi Zhou, Zekun Li, Randall Balestriero, and Yue Wang. VideoGPA: Distilling geometry priors for 3d-consistent video generation. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2601.23286.

B. Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1): 1–26, 1979. doi: 10.1214/aos/1176344552. URL https://doi.org/10.1214/aos/1176344552.

David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. URL https://proceedings.neurips.cc/paper_files/paper/2014/hash/91c56ce4a249fae5419b90cba831e303-Abstract.html.

Tamar Flash and Neville Hogan. The coordination of arm movements: an experimentally confirmed mathematical model. *The journal of Neuroscience*, 5(7):1688–1703, 1985.

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. URL https://arxiv.org/abs/2608.10107.

Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pp. 3354–3361. IEEE, 2012.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. *Advances in neural information processing systems*, 34:9574–9586, 2021.

Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. CameraCtrl: Enabling camera control for video diffusion models. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/f98fd73d59d8494489ea970747b91fe4-Abstract-Conference.html.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/39235c56aef13fb05a6adc95eb9d8d66-Abstract-Conference.html.

Neville Hogan and Dagmar Sternad. Sensitivity of smoothness measures to movement duration, amplitude, and arrests. *Journal of motor behavior*, 41(6):529–534, 2009.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Chun-Hao Paul Huang, Niloy Mitra, Hyeonho Jeong, Jae Shin Yoon, and Duygu Ceylan. Jog3r: Towards 3d-consistent video generators. *arXiv preprint arXiv:2501.01409*, 2025a.

Tianyu Huang, Wangguandong Zheng, Tengfei Wang, Yuhao Liu, Zhenwei Wang, Junta Wu, Jie Jiang, Hui Li, Rynson Lau, Wangmeng Zuo, and Chunchao Guo. Voyager: Long-range and world-consistent video diffusion for explorable 3D scene generation. *ACM Transactions on Graphics (TOG)*, 44(6):1–15, 2025b. doi: 10.1145/3763330. URL https://doi.org/10.1145/3763330.

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. URL https://openaccess.thecvf.com/content/CVPR2024/html/Huang_VBench_Comprehensive_Benchmark_Suite_for_Video_Generative_Models_CVPR_2024_paper.html.

Ziqi Huang, Fan Zhang, Xiaojie Xu, Yinan He, Jiashuo Yu, Ziyue Dong, Qianli Ma, Nattapol Chanpaisit, Chenyang Si, Yuming Jiang, Yaohui Wang, Xinyuan Chen, Ying-Cong Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench++: Comprehensive and versatile benchmark suite for video generative models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(3):3268–3285, 2026. doi: 10.1109/TPAMI.2025.3633890. URL https://doi.org/10.1109/TPAMI.2025.3633890.

Ian T. Jolliffe and Jorge Cadima. Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150202, 2016. doi: 10.1098/rsta.2015.0202. URL https://doi.org/10.1098/rsta.2015.0202.

Tommie Kerssies, Gabriele Berton, Ju He, Qihang Yu, Wufei Ma, Daan de Geus, Gijs Dubbelman, and Liang-Chieh Chen. A frame is worth one token: Efficient generative world modeling with delta tokens. *arXiv preprint arXiv:2604.04913*, 2026.

Orest Kupyn, Théo Uscidda, Marta Tintore Gazulla, Fabian Manhardt, Federico Tombari, and Christian Rupprecht. Epipolar geometry improves video generation models. *arXiv preprint arXiv:2510.21615*, 2025.

Seong Hun Lee and Javier Civera. Geometric interpretations of the normalized epipolar error. *arXiv preprint arXiv:2008.01254*, 2020.

Lin Li, Qihang Zhang, Yiming Luo, Shuai Yang, Ruilin Wang, Luyao Zhang, Mingrui Yu, Zelin Gao, Nan Xue, Boyu Zhou, Xing Zhu, Mingyu Ding, Yujun Shen, and Yinghao Xu. Causal world modeling for robot control. In *Proceedings of Robotics: Science and Systems*, Sydney, Australia, July 2026a. doi: 10.15607/RSS.2026.XXII.016. URL https://www.roboticsproceedings.org/rss22/p016.html.

Ying Li, Xiaobao Wei, Jiajun Cao, Hao Wang, Xiaowei Chi, Chengyu Bai, Qianpu Sun, Jiajun Li, Xiaojie Zhang, Peidong Jia, Jian Tang, Sirui Han, and Shanghang Zhang. WAM4D: Fast 4d world action model via spatial register tokens. *arXiv preprint arXiv:2606.14048*, 2026b. URL https://arxiv.org/abs/2606.14048.

Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, Xuanmao Li, Xingpeng Sun, Rohan Ashok, Aniruddha Mukherjee, Hao Kang, Xiangrui Kong, Gang Hua, Tianyi Zhang, Bedrich Benes, and Aniket Bera. DL3DV-10K: A large-scale scene dataset for deep learning-based 3D vision. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22160–22169. IEEE, 2024.

Chenxu Luo, Zhenheng Yang, Peng Wang, Yang Wang, Wei Xu, Ram Nevatia, and Alan Yuille. Every pixel counts++: Joint learning of geometry and motion with 3d holistic understanding. *IEEE transactions on pattern analysis and machine intelligence*, 42(10):2624–2641, 2020. doi: 10.1109/TPAMI.2019.2930258. URL https://doi.org/10.1109/TPAMI.2019.2930258.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. *Advances in neural information processing systems*, 35:17359–17372, 2022.

Soroush Nasiriany, Sep Nasiriany, Abhiram Maddukuri, and Yuke Zhu. Robocasa365: A large-scale simulation framework for training and benchmarking generalist robots. In *International Conference on Learning Representations*, volume 2026, pp. 98643–98667, 2026.

NVIDIA Cosmos Team. Cosmos 3: Omnimodal world models for physical AI. Technical report, NVIDIA, 2026. URL https://research.nvidia.com/labs/cosmos-lab/cosmos3/technical-report.pdf.

Lu Qiu, Yizhuo Li, Yi Chen, Yuying Ge, Yixiao Ge, and Xihui Liu. Making foresight actionable: Repurposing representation alignment in world action models. *arXiv preprint arXiv:2606.12217*, 2026. URL https://arxiv.org/abs/2606.12217.

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 *International conference on machine learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

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*, pp. 6121–6132, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Ren_GEN3C_3D-Informed_World-Consistent_Video_Generation_with_Precise_Camera_Control_CVPR_2025_paper.html.

Yichao Shen, Fangyun Wei, Zhiying Du, Yaobo Liang, Yan Lu, Jiaolong Yang, Nanning Zheng, and Baining Guo. VideoVLA: Video generators can be generalizable robot manipulators. In *Advances in Neural Information Processing Systems*, volume 38, pp. 105974–105998, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/89a3b655a8b68ae1c76b768152c9c19d-Abstract-Conference.html.

Jürgen Sturm, Nikolas Engelhard, Felix Endres, Wolfram Burgard, and Daniel Cremers. A benchmark for the evaluation of rgb-d slam systems. In *2012 IEEE/RSJ international conference on intelligent robots and systems*, pp. 573–580. IEEE, 2012.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. URL https://arxiv.org/abs/2503.20314.

Bill Triggs, Philip F McLauchlan, Richard I Hartley, and Andrew W Fitzgibbon. Bundle adjustment—a modern synthesis. In *Vision Algorithms: Theory and Practice*, pp. 298–372. Springer, 2000. doi: 10.1007/3-540-44480-7_21. URL https://doi.org/10.1007/3-540-44480-7_21.

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

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual geometry grounded transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5294–5306, 2025a. URL https://openaccess.thecvf.com/content/CVPR2025/html/Wang_VGGT_Visual_Geometry_Grounded_Transformer_CVPR_2025_paper.html.

Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A. Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10510–10522, 2025b. URL https://openaccess.thecvf.com/content/CVPR2025/html/Wang_Continuous_3D_Perception_Model_with_Persistent_State_CVPR_2025_paper.html.

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*, pp. 20697–20709, 2024a. URL https://openaccess.thecvf.com/content/CVPR2024/html/Wang_DUSt3R_Geometric_3D_Vision_Made_Easy_CVPR_2024_paper.html.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

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*, pp. 1–11, 2024b. URL https://github.com/TencentARC/MotionCtrl.

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. In *International Conference on Learning Representations*, volume 2026, pp. 119186–119209, 2026.

Lishan Yang, Wenxuan Song, Xi Wang, Pingyue Sheng, Zheng Fang, Ziyang Zhou, Junjie He, Haodong Yan, Jiayi Chen, Nan Sun, Qiao Sun, Pengwei Wang, Lingqiao Liu, Yan Wang, Yuxiang Gao, Feras Dayoub, and Haoang Li. 4D-WAM: Infusing spatiotemporal awareness into world action models through trajectory fields. *arXiv preprint arXiv:2608.08023*, 2026. URL https://arxiv.org/abs/2608.08023.

Seonghyeon Ye, Yunhao Ge, Kaiyuan Zheng, Shenyuan Gao, Sihyun Yu, George Kurian, Suneel Indupuru, You Liang Tan, Chuning Zhu, Jiannan Xiang, Ayaan Malik, Kyungmin Lee, William Liang, Nadun Ranawaka, Jiasheng Gu, Yinzhen Xu, Guanzhi Wang, Fengyuan Hu, Avnish Narayan, Johan Bjorck, Jing Wang, Gwanghyun Kim, Dantong Niu, Ruijie Zheng, Yuqi Xie, Jimmy Wu, Qi Wang, Ryan Julian, Danfei Xu, Yilun Du, Yevgen Chebotar, Scott Reed, Jan Kautz, Yuke Zhu, Linxi Fan, and Joel Jang. World action models are zero-shot policies. *arXiv preprint arXiv:2602.15922*, 2026. URL https://arxiv.org/abs/2602.15922.

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. URL https://arxiv.org/abs/2603.16666.

Jianjun Zhang, Jian Zhu, Taiyi Su, Chong Ma, Zitai Huang, Yi Xu, and Hanli Wang. Learning 4d geometric priors for inference-efficient world action models. *arXiv preprint arXiv:2607.05468*, 2026. URL https://arxiv.org/abs/2607.05468.

Qihang Zhang, Shuangfei Zhai, Miguel Ángel Bautista Martin, Kevin Miao, Alexander Toshev, Joshua Susskind, and Jiatao Gu. World-consistent video diffusion with explicit 3d modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21685–21695, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Zhang_World-consistent_Video_Diffusion_with_Explicit_3D_Modeling_CVPR_2025_paper.html.

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

Zichao Zhang and Davide Scaramuzza. A tutorial on quantitative trajectory evaluation for visual (-inertial) odometry. In *2018 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pp. 7244–7251. IEEE, 2018.

Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *arXiv preprint arXiv:1805.09817*, 2018.

Fangqi Zhu, Hongtao Wu, Song Guo, Yuxiao Liu, Chilam Cheang, and Tao Kong. IRASim: A fine-grained world model for robot manipulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9834–9844, 2025. doi: 10.1109/ICCV51701.2025.00917. URL https://openaccess.thecvf.com/content/ICCV2025/html/Zhu_IRASim_A_Fine-Grained_World_Model_for_Robot_Manipulation_ICCV_2025_paper.html.

The appendix provides details needed to interpret and reproduce CausalAlign. Appendix A specifies the causal diagnosis, target selection, and three-stage delta-token alignment procedure, and provides additional evidence for geometric-state forgetting. Appendix B defines the evaluation metrics and normalization, clarifies the training and trajectory protocols, and extends the experiments to text-to-video and camera-trajectory-conditioned generation. Appendix C shows more ablation experiments to demonstrate the contributions of structured delta representations, update placement, temporal modulation, and auxiliary training objectives.

A Detailed Method Description

A.1 Causal Diagnosis and Target Selection

Diagnostic subset. We reserve a fixed diagnostic subset $\mathcal{D}_{\text{diag}}$ containing approximately 10% of the complete experimental data. The same subset is used to generate Fig. 3 and to select the target layers and denoising timesteps. Target selection is performed before adaptation using only the measurements from the frozen WAN2.2 base model (Team Wan et al., 2025). After adaptation, we re-evaluate the same subset only to visualize the change in the geometric state dynamics. The diagnostic subset is not used for the held-out evaluation or downstream WAM evaluation, and the selected target locations remain fixed in all subsequent experiments.

Noise schedule and reindexed denoising progress. Let r denote the native WAN2.2 scheduler timestep. Under the native scheduler convention, denoising starts from a large r with high noise and proceeds toward a smaller r with low noise. For diagnosis, we reverse this ordering and use a reindexed denoising-progress coordinate s. Specifically, $s = \pi(r)$ denotes the order-reversing mapping used in our implementation, so smaller s corresponds to an earlier, higher-noise state and larger s corresponds to a later, lower-noise state. All diagnostic curves, decline-rate computations, target selection, and alignment losses use this reindexed coordinate s, rather than the native scheduler timestep r.

For each diagnostic real video, we inject the native scheduler noise corresponding to $r = \pi^{-1}(s)$ and extract the VGM hidden states from the resulting denoising computation. The probe features are therefore obtained from noised real videos rather than from free-running generated videos. For the intervention analysis, we instead perform free-running generation from the same first frame, prompt, initial noise, and random seed, and intervene at the selected layer–progress locations. We follow the native WAN2.2 expert routing and noise schedule; no additional routing rule is introduced for diagnosis.

Ego-motion decodability. For each layer–timestep location $c = (\ell, s)$, a linear probe (Alain & Bengio, 2016) predicts the relative camera motion $\boldsymbol{\xi}_t$ from the VGM transition feature $\mathbf{u}_{t,\ell,s}$. Let

$$e_{\ell,s} = \mathbb{E}_{t\sim\mathcal{D}_{\text{diag}}}\left[\ell_{\text{pose}}\left(\widehat{\boldsymbol{\xi}}_{t,\ell,s}, \boldsymbol{\xi}_t\right)\right] \tag{13}$$

denote the held-out probe error, and let e_{chance} denote the error of a mean-motion predictor. We convert the error into a higher-is-better decodability score:

$$D_{\ell,s} = \left[1 - \frac{e_{\ell,s}}{e_{\text{chance}} + \epsilon}\right]_{[0,1]}, \qquad [x]_{[0,1]} = \min\{1, \max\{0, x\}\}, \tag{14}$$

where ϵ is a small numerical constant. The raw values of $D_{\ell,s}$ are retained for reporting; normalization is applied only when constructing the target-selection profiles.

Causal intervention. At each location $c = (\ell, s)$, we keep the prompt, initial noise, and random seed fixed. Let $a \in \mathcal{A}$ denote one of four interventions: *Mask*, *Shuffle*, *Replace*, and *Inject*. Mask removes the temporal state, Shuffle permutes its temporal order, Replace substitutes the corresponding state from another clip, and Inject inserts a pose-matched donor state.

For each target clip, donor clips are sampled from other clips in $\mathcal{D}_{\text{diag}}$ with the same frame length and valid pose estimates. For Inject, the donor is selected according to the closest relative-pose

distance; for Replace, the donor is sampled without pose matching. Let $\mathbf{h}_c$ be the original hidden state and let $\Delta\mathbf{h}_c^{(a)}$ be the change induced by intervention a. The relative intervention magnitude is

$$r_c^{(a)} = \frac{\left\|\Delta\mathbf{h}_c^{(a)}\right\|_2}{\|\mathbf{h}_c\|_2 + \epsilon}. \tag{15}$$

The random control is constructed from the same donor pool and rescaled to match the intervention magnitude:

$$\Delta\mathbf{h}_c^{(\mathrm{rand},a)} = \frac{\left\|\Delta\mathbf{h}_c^{(a)}\right\|_2}{\left\|\Delta\mathbf{h}_c^{(\mathrm{donor})}\right\|_2 + \epsilon}\Delta\mathbf{h}_c^{(\mathrm{donor})}. \tag{16}$$

Let $Y_c^{(a)}$ and $Y_c^{(\mathrm{rand},a)}$ denote the videos generated under the intervention and its norm-matched random control. The random-control-adjusted causal effect is

$$C_c^{(a)} = \eta_a \left[\mathbb{E}\left[E_{\mathrm{traj}}\left(Y_c^{(a)}\right)\right] - \mathbb{E}\left[E_{\mathrm{traj}}\left(Y_c^{(\mathrm{rand},a)}\right)\right]\right], \tag{17}$$

where $E_{\mathrm{traj}}(Y; d)$ measures the trajectory error of the generated video Y against the ground-truth camera trajectory $\mathbf{T}_{1:T}^{\mathrm{ref},d}$ associated with diagnostic clip d. This trajectory is an evaluation reference rather than a generation condition. We set $\eta_a = +1$ for disruptive interventions and $\eta_a = -1$ for Inject, so that a larger value always indicates a stronger causal effect. The aggregate causal score is

$$C_{\ell,s} = \frac{1}{|\mathcal{A}|}\sum_{a\in\mathcal{A}} C_{(\ell,s)}^{(a)}. \tag{18}$$

For $A \in \{L, S\}$, we average over the complementary axis:

$$D_i^A = \mathrm{Avg}_j\, D_{i,j}, \qquad C_i^A = \mathrm{Avg}_j\, C_{i,j}. \tag{19}$$

Each profile is independently min–max normalized:

$$\mathcal{N}(x_i) = \frac{x_i - \min_j x_j}{\max_j x_j - \min_j x_j + \epsilon}. \tag{20}$$

The joint profile used for target selection is

$$J_i^A = \frac{1}{2}\left[\mathcal{N}(D_i^A) + \mathcal{N}(C_i^A)\right]. \tag{21}$$

Let q_i denote the actual layer coordinate or the reindexed denoising-progress coordinate s. For timestep profiles, the coordinates are ordered by increasing s, from earlier to later denoising states. The positive decline rate is

$$g_i^A = \left[\frac{J_i^A - J_{i+1}^A}{q_{i+1} - q_i}\right]_+, \qquad [x]_+ = \max\{0, x\}. \tag{22}$$

The resulting target sets are

$$\mathcal{L}_{\mathrm{tar}} = \{10, 12, 14, 20\}, \qquad \mathcal{S}_{\mathrm{tar}} = \{398, 799\}.$$

The timestep set $\mathcal{S}_{\mathrm{tar}}$ is expressed in the reindexed progress coordinate s, so $s = 398$ is earlier than $s = 799$ in the denoising process. Layer 14 is the primary target, while layers $\{10, 12, 20\}$ are auxiliary targets. The alignment weights are proportional to the positive temporal decline rates:

$$\omega_{\ell,s} = \frac{g_s^S}{\sum_{s'\in\mathcal{S}_{\mathrm{tar}}} g_{s'}^S + \epsilon}, \qquad \ell \in \mathcal{L}_{\mathrm{tar}}. \tag{23}$$

We report the raw $D_{\ell,s}$ and $C_{\ell,s}$ values, the intervention magnitudes $r_c^{(a)}$, and their uncertainty estimates. Specifically, we bootstrap diagnostic clips (Efron, 1979) while preserving all intervention variants belonging to the same clip. For a quantity q, the 95% confidence interval is computed from the bootstrap distribution as

$$\mathrm{CI}_{95}(q) = \left[Q_{0.025}(q^{(1:B)}), Q_{0.975}(q^{(1:B)})\right], \tag{24}$$

where B is the number of bootstrap replicates and Q_α denotes the α-quantile.

A.2 Additional Details of Delta-Token Alignment

The frozen streaming 3D foundation model (Chen et al., 2026a) produces geometric states

$$\mathbf{z}_t^{\mathrm{T}} = \Phi^{\mathrm{T}}(x_{1:t}), \tag{25}$$

where $x_{1:t}$ denotes the observed video stream up to frame t. The teacher delta token is encoded from two consecutive geometric states:

$$\boldsymbol{\delta}_t^{\mathrm{T}} = E_{\Delta}\left(\mathbf{z}_t^{\mathrm{T}}, \mathbf{z}_{t+1}^{\mathrm{T}}\right). \tag{26}$$

The teacher delta-token modules are trained with transition reconstruction, temporal prediction, and relative-pose supervision:

$$\mathcal{L}_{\mathrm{T}} = \mathcal{L}_{\mathrm{tr}} + \lambda_{\mathrm{temp}}\mathcal{L}_{\mathrm{temp}} + \lambda_{\mathrm{pose}}^{T}\mathcal{L}_{\mathrm{pose}}^{T}. \tag{27}$$

Here, $\mathcal{L}_{\mathrm{tr}}$ reconstructs the next teacher state, $\mathcal{L}_{\mathrm{temp}}$ predicts the next delta token, and $\mathcal{L}_{\mathrm{pose}}^{T}$ supervises the relative camera motion $\Delta\mathbf{T}_t = \mathbf{T}_t^{-1}\mathbf{T}_{t+1}$. After this stage, the teacher encoder, delta encoder, transition decoder, temporal predictor, and camera decoder are frozen.

For a selected VGM layer ℓ, the VGM transition feature is mapped into the teacher delta-token space by

$$\boldsymbol{\delta}_{t,\ell,s}^{\mathrm{V}} = Q_{\ell}\left(\mathbf{u}_{t,\ell,s}\right), \tag{28}$$

where Q_{ℓ} is a layer-specific projector. With the VGM and teacher modules frozen, the projectors are trained using

$$\mathcal{L}_{\mathrm{map}} = \sum_{\ell\in\mathcal{L}_{\mathrm{tar}}}\sum_{s\in\mathcal{S}_{\mathrm{tar}}}\omega_{\ell,s}\mathbb{E}_t\left[\left\|\boldsymbol{\delta}_{t,\ell,s}^{\mathrm{V}} - \boldsymbol{\delta}_t^{\mathrm{T}}\right\|_2^2\right]. \tag{29}$$

The projectors are frozen after training. In the final stage, LoRA adapters (Hu et al., 2022) are updated only at $\mathcal{L}_{\mathrm{tar}}$ and $\mathcal{S}_{\mathrm{tar}}$ using the pose-supervision and delta-token matching losses defined in Sec. 3.3. The teacher modules and projectors are used only during training and introduce no computation during inference.

A.3 More Evidence for geometric state Forgetting

The original decodability and intervention analyses may admit several alternative explanations. A late decrease in linear decodability could result from nonlinear re-encoding, while whole-state interventions may change appearance and scene content together with ego-motion. Moreover, camera geometry may already have been committed to the evolving video latent before the measured hidden-state signals decline, as suggested by the early emergence of camera motion reported in AC3D (Bahmani et al., 2025). We therefore conduct three complementary analyses. The results are summarized in Fig. 5.

Linear and nonlinear decodability. We evaluate ego-motion decodability using both a linear probe and a small MLP probe. As shown in Fig. 5(a), both probes exhibit a consistent rise-and-fall pattern over denoising. Thus, the late-stage decrease is not specific to the linear readout used in the main diagnosis. The result supports reduced accessibility of the ego-motion state under both tested probe families, while making no assumption that all possible nonlinear representations are unrecoverable.

Motion-specific causal intervention. To isolate the causal role of ego-motion from appearance and scene content, we compare an intervention applied to the ego-motion-related delta-token component with a norm-matched intervention applied to non-motion content. Let $C_{\mathrm{motion}}(s)$ denote the change in the generated camera trajectory caused by the motion-only intervention, and let $C_{\mathrm{content}}(s)$ denote the corresponding effect of the content control. Figure 5(b) shows that

$$C_{\mathrm{motion}}(s) \gg C_{\mathrm{content}}(s)$$

throughout the main diagnostic range, indicating that the measured intervention effect is primarily attributable to the motion-specific component rather than to a generic perturbation of the hidden state. The motion-specific causal effect also decreases during late denoising.

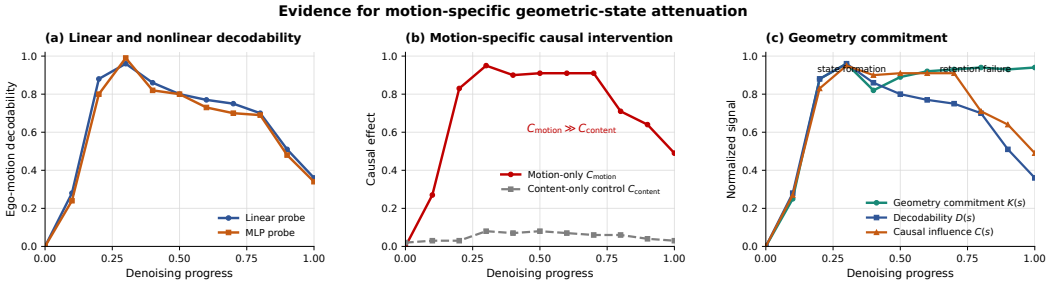

Figure 5: Additional evidence for geometric state forgetting. (a) Linear and MLP probes both show a rise-and-fall pattern in ego-motion decodability. (b) Motion-only interventions produce substantially larger causal effects than norm-matched content-only controls. (c) Geometry commitment $K(s)$ rises early and remains high, while the decodability $D(s)$ and motion-specific causal effect $C(s)$ reused from (a) and (b), respectively, decline during late denoising.

Geometry commitment and late attenuation. Finally, we compare the geometry-commitment score $K(s)$ of the evolving video latent with the internal ego-motion signals. Here, $D(s)$ is the same linear-probe decodability curve reported in Fig. 5(a), and $C(s)$ is the same motion-only causal-effect curve reported in Fig. 5(b). This shared construction allows the three quantities to be compared directly. As shown in Fig. 5(c), $K(s)$ rises rapidly during early denoising and remains high thereafter, whereas both $D(s)$ and $C(s)$ decline at later stages. This temporal separation indicates that the geometric state has already formed in the evolving latent when the measured hidden-state accessibility and causal effect begin to attenuate. It therefore supports our interpretation of the late-stage decline as a retention failure of the internal ego-motion state, rather than a failure to form camera geometry in the first place.

B Extended Evaluation of CausalAlign

This section first clarifies the metric definitions and training settings used in the I2V evaluation in the main paper. We then explain why the unconstrained I2V protocol does not use direct ground-truth trajectory error. Finally, we report two complementary experiments: text-to-video generation and camera-trajectory-conditioned generation.

B.1 Metric Definitions and Normalization

For generated clip i, let $\mathbf{p}_{i,t}$ and $\mathbf{R}_{i,t}$ denote the estimated camera position and rotation at frame t. Let $\Delta\tau$ be the frame interval and let $d_{i,t}(u)$ be the estimated depth at pixel u and frame t. The clip-level median scene depth is

$$d_i = \mathrm{median}_{t,u}\, d_{i,t}(u). \tag{30}$$

The depth-normalized translational speed and angular speed are

$$v_{i,t} = \frac{\|\mathbf{p}_{i,t+1} - \mathbf{p}_{i,t}\|_2}{d_i \Delta\tau}, \qquad \omega_{i,t} = \frac{\left\|\log\left(\mathbf{R}_{i,t}^\top \mathbf{R}_{i,t+1}\right)\right\|_2}{\Delta\tau}. \tag{31}$$

For $m \in \{v, \omega\}$, let

$$\overline{m}_i = \frac{1}{T-1}\sum_{t=1}^{T-1} m_{i,t} \tag{32}$$

denote the clip-level mean. Let $(\mu_m^{\mathrm{real}}, \sigma_m^{\mathrm{real}})$ denote the mean and standard deviation of metric m on the real-video validation distribution. We define the proximity score

$$S_m(i) = \exp\left(-\frac{\left|\overline{m}_i - \mu_m^{\mathrm{real}}\right|}{\sigma_m^{\mathrm{real}} + \epsilon}\right), \tag{33}$$

and report its dataset-level average:

$$M_m = \frac{1}{N}\sum_{i=1}^{N} S_m(i). \tag{34}$$

Thus, "Mean Trans. Prox." and "Mean Angular Prox." correspond to M_v and M_ω, respectively. These scores measure proximity to the real-data mean of the corresponding motion magnitude. A value closer to one indicates that the generated clip-level mean is closer to the real-data mean; it does not by itself measure agreement with the full motion distribution.

For the composite gain used in the update-placement experiments, let $m_{i,k}$ denote the value of metric k for clip i. Let a_k and b_k be fixed normalization bounds computed over the comparison set. The direction-corrected normalized score is

$$q_{i,k} = \begin{cases} \left[\dfrac{m_{i,k} - a_k}{b_k - a_k + \epsilon}\right]_{[0,1]}, & k \in \mathcal{K}_\uparrow, \\ \left[\dfrac{b_k - m_{i,k}}{b_k - a_k + \epsilon}\right]_{[0,1]}, & k \in \mathcal{K}_\downarrow, \end{cases} \tag{35}$$

where $[x]_{[0,1]} = \min\{1, \max\{0, x\}\}$. The dataset-level score for method M is

$$q_k^{(M)} = \frac{1}{N}\sum_{i=1}^{N} q_{i,k}^{(M)}. \tag{36}$$

The composite gain of an adapted model over the base model is

$$G = 100\left[\sum_k w_k q_k^{(\text{adapted})} - \sum_k w_k q_k^{(\text{base})}\right], \qquad \sum_k w_k = 1. \tag{37}$$

For the update-placement study, the composite contains rotation RPE, translation-direction RPE, Sim(3) ATE, endpoint error, speed-profile error, and depth AbsRel.

B.2 Training and Inference Details

We instantiate CausalAlign on WAN2.2-14B-I2V (Team Wan et al., 2025) using DL3DV (Ling et al., 2024) and RealEstate10K (Zhou et al., 2018). Each dataset is divided into 90% training data and 10% held-out evaluation data. Following the three-stage procedure in Fig. 2, the three stages use fixed 30%, 10%, and 100% subsets of the training split, respectively.

In Stage 1, the teacher delta-token modules, including the delta encoder, transition decoder, temporal predictor, and camera decoder, are trained with

$$\lambda_{\text{temp}} = 0.1, \qquad \lambda_{\text{pose}}^{T} = 0.2. \tag{38}$$

In Stage 2, the VGM-side projectors are trained with $\mathcal{L}_{\text{map}}$ while the VGM and all teacher modules remain frozen. In Stage 3, LoRA adapters are optimized only at

$$\mathcal{L}_{\text{tar}} = \{10, 12, 14, 20\}, \qquad \mathcal{S}_{\text{tar}} = \{398, 799\}, \tag{39}$$

where layer 14 is the primary target and layers $\{10, 12, 20\}$ are auxiliary targets. The final objective is

$$\mathcal{L}_{\text{pose}} + \lambda_\Delta \mathcal{L}_\Delta, \qquad \lambda_\Delta = 0.2. \tag{40}$$

The original VGM backbone, LingBot-Map, teacher modules, and VGM-side projectors remain frozen during the final stage. Only the adapted VGM is retained at inference, and no teacher or external geometric cache is used.

Why the main I2V protocol does not use direct ground-truth pose error. In the main first-frame-and-prompt I2V protocol, the model receives only the first frame and a text prompt. The future video is therefore one-to-many: multiple camera trajectories can be valid continuations of the same input. Directly comparing an unconstrained generated trajectory with a single ground-truth trajectory can consequently penalize valid alternative motions.

This differs from the causal diagnosis in Sec. 3.2. There, the ground-truth trajectory of the corresponding diagnostic video is used only as a fixed measurement reference for $E_{\mathrm{traj}}(Y; d)$. It is never provided to the generator. The diagnostic score measures how an intervention changes the generated trajectory error relative to a norm-matched random control, thereby measuring the sensitivity of the generation process to the intervened internal state. It does not require the generated video to reproduce the reference trajectory.

For the main unconstrained I2V evaluation, methods whose native inference does not require an explicit camera trajectory receive only the first frame and the text prompt. These methods are evaluated with matched output length, resolution, frame rate, sampling budget, and random seeds. For methods whose standard inference requires an explicit camera trajectory and a 3D cache, such as GEN3C and Voyager (Ren et al., 2025; Huang et al., 2025b), we follow their native inference interfaces: the ground-truth camera trajectory is provided as the camera-control condition, and a fresh 3D cache is constructed from the available input frame(s), without using future video frames. These methods are reported as trajectory-conditioned baselines rather than as unconstrained generators.

Regardless of the input interface, every generated video is processed by the same frozen, independent reconstruction pipeline. For the main intrinsic comparison, trajectory analysis is based only on the reconstructed video and uses motion magnitude, geometric self-consistency, temporal dynamics, and visual quality. Neither the conditioning trajectory nor its ground-truth trajectory is used as the target for these intrinsic metrics. Thus, the trajectory-conditioned baselines are not evaluated by a ground-truth pose matching loss, while the evaluator and intrinsic metrics remain shared across methods. Results from unconstrained and trajectory-conditioned protocols are therefore interpreted separately.

To complement this intrinsic evaluation, we introduce a separate camera-trajectory-conditioned protocol. In this setting, the reference camera trajectory $\mathbf{T}^*_{1:T}$ is provided as an explicit condition, and direct comparison with the target trajectory is well-defined. The results are reported in Sec. B.4.

We next evaluate CausalAlign in two additional settings. The first tests whether the learned alignment transfers to text-to-video generation. The second evaluates direct trajectory accuracy under explicit camera conditioning. Within each setting, methods share evaluation inputs, output length, resolution, frame rate, sampling budgets, and evaluation procedures.

B.3 Text-to-Video Generation

We adapt Wan2.2-14B-T2V with the same training hyperparameters and evaluate it on the same test split as the main I2V experiment. Text prompts are generated from RealEstate10K clips (Zhou et al., 2018) following the captioning procedure used by CameraCtrl (He et al., 2025). Both the native and adapted models receive only these prompts.

Visual quality. Table 5 reports VBench quality, semantic, and smoothness scores (Huang et al., 2024), together with FVD (Unterthiner et al., 2018), CLIP (Radford et al., 2021), SSIM (Wang et al., 2004), and LPIPS (Zhang et al., 2018). CausalAlign improves quality, semantic score, FVD, and LPIPS, matches smoothness, and has lower CLIP and SSIM scores.

Table 5: Visual quality on Wan2.2-14B-T2V. Bold marks the best value in each column, including ties.

	VBench			Other visual metrics			
Method	Quality↑	Semantic↑	Smooth.↑	FVD↓	CLIP↑	SSIM↑	LPIPS↓
Wan2.2-14B-T2V	0.850	0.729	**0.979**	183.6	**0.331**	**0.728**	0.181
CausalAlign	**0.852**	**0.730**	**0.979**	**183.4**	0.320	0.726	**0.175**

Intrinsic ego motion. The frozen pose, depth, optical-flow, and feature estimators from the main evaluation process all generated videos. Motion magnitudes are compared with the real-video reference distribution, whereas geometric consistency and temporal dynamics measure internal video consistency. Table 6 shows improved motion magnitudes, reprojection, depth–flow, and epipolar errors, together with higher smoothness and lower jerk.

Table 6: Intrinsic ego motion under text-only conditioning. Mean Trans. Prox. and Mean Angular Prox. measure the proximity of generated clip-level mean speeds to the corresponding real-data means; higher values are better.

	Motion magnitude		Geometric consistency			Temporal dynamics	
Method	Mean Trans. Prox.↑	Mean Angular Prox.↑	Reproj. (px)↓	Depth–flow (px)↓	Epipolar (px)↓	Smooth.↑	Jerk↓
Wan2.2-14B-T2V	0.63	0.68	8.2	7.1	5.4	0.78	0.181
CausalAlign	**0.78**	**0.90**	**6.9**	**5.3**	**3.4**	**0.86**	**0.154**



Figure 6: Successful mobile manipulation on RoboCasa365. The left group shows the robot collecting food in a bowl, turning toward the refrigerator, and placing the bowl inside. The right group shows the robot carrying a bread basket from the cabinet area to the dining counter. Each group corresponds to one actual policy rollout, with frames arranged from left to right in temporal order.

B.4 Camera-Trajectory-Conditioned Generation

We apply CausalAlign to Wan2.2-Fun-A14B-Control-Camera (Alibaba PAI, 2025) and compare it with the native model and FantasyWorld (Dai et al., 2026). Each method receives the same first frame, ground-truth camera trajectory, and text prompt. Generated trajectories are recovered with a shared frozen pose estimator and compared with the conditioning trajectories using identical temporal correspondence, coordinate conventions, and trajectory alignment.

Table 7 reports relative rotation and translation errors (RRE (Zhang & Scaramuzza, 2018) and RTE (Geiger et al., 2012)), absolute trajectory error (ATE (Sturm et al., 2012)), final displacement error (FDE (Alahi et al., 2016)), drift (Geiger et al., 2012), smoothness (Hogan & Sternad, 2009), and jerk (Flash & Hogan, 1985). CausalAlign gives the best values on six of the seven metrics; FantasyWorld has the lowest RTE.

B.5 More Visualization For Mobile Manipulation Rollouts

Here we give two more successful examples on RoboCasa365 (Nasiriany et al., 2026) in Fig. 6. Meanwhile, we provide our experiment setting in Table 13.

Table 7: Camera-conditioned video generation. The native model is Wan2.2-Fun-A14B-Control-Camera. Bold marks the best value per column.

	Relative pose		Global trajectory			Temporal dynamics	
Method	RRE (°)↓	RTE (%)↓	ATE (m)↓	FDE (m)↓	Drift (%)↓	Smooth.↑	Jerk↓
Native	3.82	7.70	0.214	0.263	13.6	0.69	0.189
FantasyWorld	3.72	**6.60**	0.169	0.208	10.5	0.82	0.159
CausalAlign	**3.21**	6.80	**0.167**	**0.201**	**10.2**	**0.85**	**0.155**

C Additional Ablation Studies

C.1 Structured Delta-Space Representation

Representations and Stage-1 supervision. Table 4 evaluates the frozen teacher delta-token representation learned in Stage 1. The structured encoder produces

$$\boldsymbol{\delta}_t^{\mathrm{T}} = \left[\boldsymbol{\delta}_t^{\mathrm{ego}}; \boldsymbol{\delta}_t^{\mathrm{geo}}; \boldsymbol{\delta}_t^{\mathrm{vis}}\right], \tag{41}$$

where the three regions encode relative camera motion, camera-local geometric changes such as depth and surface-normal differences, and RGB changes, respectively.

These regions are structured target channels rather than three additional losses. The camera decoder uses the pose target to provide explicit ego-motion supervision through $\mathcal{L}_{\mathrm{pose}}^{T}$. The transition decoder reconstructs the next teacher state, so $\mathcal{L}_{\mathrm{tr}}$ supervises the transition-related geometric and visual content in that state. The temporal predictor forecasts the complete next delta token, and $\mathcal{L}_{\mathrm{temp}}$ is therefore applied to the concatenated token, including all three regions. Thus, the pose, depth/normal, and RGB targets jointly implement the Stage-1 objective in Eq. (7), rather than defining separate objectives.

For the unpartitioned controls, the pair and difference encoders also produce 128-dimensional tokens and are trained with the same three loss terms and the same pose, geometric, and visual targets, but without explicitly separating the token into semantic regions. All encoders receive 40,000 updates: 10,000 on 4,000 training scenes and 30,000 on 8,312 scenes. Camera poses and target geometry provide the supervision labels.

Readout protocol. Independent two-layer, width-256 probes are trained for 3,000 updates with batch size 128 on 4,000 scenes. Learning rates are selected from $\{10^{-5}, 10^{-4}, 10^{-3}\}$ using 150-step validation runs. Evaluation covers 256 development scenes and three seeds. Each scene contributes nine frame pairs, combining source indices $\{0, 4, 8\}$ with gaps $\{1, 2, 4\}$ in a 21-frame, stride-4 feature sequence. Scores are averaged over frame pairs within each scene, then over scenes and seeds.

Targets and metrics. Ego-motion readouts are evaluated against DL3DV poses (Ling et al., 2024) using rotation and translation-direction angular errors. Geometry probes predict log-depth changes on a 16×28 grid, which are added to the common source log-depth and compared with frozen LingBot-Map estimates (Chen et al., 2026a) on shared valid pixels. We report $\mathrm{AbsRel} = \langle |\hat{d} - d|/d \rangle$ and $\delta_1 = \langle \mathbf{1}[\max(\hat{d}/d, d/\hat{d}) < 1.25] \rangle$ (Eigen et al., 2014). Visual probes predict RGB changes on a 32×56 grid; reconstruction from the common source image is evaluated with SSIM (Wang et al., 2004) and AlexNet LPIPS (Zhang et al., 2018). Copying the source image gives SSIM 0.29397 and LPIPS 0.08924, versus 0.30283 and 0.08676 for the complete delta token.

Partition specialization. For the matched-width comparisons in Table 4(a), PCA (Jolliffe & Cadima, 2016) fitted on training data projects the supervised unpartitioned encoders to 64, 40, and 24 dimensions. Delta regions lead all 12 paired metric comparisons. Table 8 further evaluates every region on every task at a common width of 24 dimensions, projecting the ego and geometry regions with training-only PCA. Each region performs best on its designated task.

Table 8: Equal-width cross-probes. All regions use 24 dimensions. Values are means over three seeds on 256 development scenes; bold marks the best region per metric.

	Ego motion		Geometry		Visual	
Region	Rotation (°)↓	Direction (°)↓	AbsRel↓	δ_1 ↑	SSIM↑	LPIPS↓
Ego	**2.429**	**7.558**	0.3115	0.6764	0.2881	0.08964
Geometry	13.070	49.822	**0.2292**	**0.7448**	0.2908	0.08894
Visual	14.682	65.830	0.3188	0.6620	**0.3050**	**0.08643**

Original features and extraction controls. The original LingBot-Map frame representation concatenates the final camera token and the mean of 777 patch tokens, yielding 4096 dimensions. The

pooled pair A concatenates two frame vectors (8192D), while B takes their difference (4096D). Both retain their original width, with coordinate-wise standardization fitted on probe-training data. Table 4(b) compares these frozen teacher features with the complete learned delta representation; panel (a) isolates partitioning under matched supervision. Delta leads all six readout metrics with the original eight-frame streaming context, with whole-sequence extraction, and with whole-sequence extraction plus four-layer averaging. A full-spatial-grid difference provides a separate 133,120D control: its AbsRel 0.2109, SSIM 0.3533, and LPIPS 0.08575 improve on the corresponding delta readouts.

C.2 Update Placement and Temporal Modulation

Independent adapter placement. We independently train a LoRA adapter for each of five layer sets in the second transformer of Wan2.2-14B. All runs use the same training data, objective, LoRA configuration, optimization budget, and inference protocol, and the resulting updates are matched in effective global norm. The fixed-source video-to-video evaluation uses 32 scenes, one generation seed, ten inference steps, and identical sources, prompts, scheduler, and initial noise. The composite combines rotation RPE, translation-direction RPE, Sim(3) ATE, endpoint error, speed-profile error, and depth AbsRel, using frozen LingBot-Map estimates from source videos as references. Positive values denote improvements; 95% confidence intervals use 50,000 scene-bootstrap resamples (Efron, 1979).

Table 9(a) reports the complete placement comparison. The independently trained adapter at layers 10, 12, 14, and 20 gives the largest gain over the native model (3.109%). Its gain over a norm-matched random update is 2.224%. This matched-adapter comparison isolates the effect of target-layer placement.

Temporal modulation. Using 32 scenes and three seeds, we compare timestep-conditioned, static, time-averaged, and ensemble updates under the same fixed-source protocol. Table 9(b) reports all three contrasts. The ensemble has the largest mean gain over the static update; each interval includes zero.

Table 9: Matched-adapter placement and temporal-modulation experiments on Wan2.2-14B. Gains are percentages; brackets give 95% scene-bootstrap confidence intervals. Placement uses one seed, and temporal modulation uses three seeds.

(a) Layer placement

	Versus native		Versus random update	
Layer set	Gain	95% CI	Gain	95% CI
10, 12, 14, 20	+3.109	[+0.808, +5.373]	+2.224	[+0.302, +4.022]
16, 18, 20, 26	+0.250	[-2.483, +2.769]	-1.045	[-4.391, +1.961]
20, 22, 24, 30	-2.194	[-5.552, +0.760]	-1.731	[-5.315, +1.611]
10, 14, 20, 24	-0.513	[-3.595, +2.441]	-0.738	[-4.136, +2.232]
10, 12, 14, 16, 18, 20	-0.414	[-2.919, +1.883]	+1.652	[-0.348, +3.571]

(b) Temporal modulation

Comparison	Gain	95% CI
Dynamic vs. static	+0.694	[-0.949, +2.405]
Dynamic vs. time-averaged	-0.238	[-2.055, +1.608]
Ensemble vs. static	+2.130	[-0.057, +4.467]

C.3 Auxiliary Objectives and Update Strength

We compare local relative-pose supervision (RPE), kinematic supervision (speed, acceleration, and jerk), and their combination on Wan2.2-TI2V-5B. The three runs continue from the same parent checkpoint to the same endpoint, step 48. Local RPE gives the largest trajectory gain (3.525%), with improvements concentrated in endpoint and motion-dynamics errors. Combining objectives reduces this gain to 1.299% while improving geometry and temporal preservation (Table 10).

In matched step-24 continuations, reducing the later-layer update scale from 0.50 to 0.25 increases trajectory gain from 1.750% to 2.209% and geometry gain from -0.506% to 0.761%. Temporal and visual gains decrease from 0.218% and 0.059% to 0.086% and 0.021%, respectively. Both comparisons show that stronger or more numerous constraints need not improve all metric groups together.

Table 10: Auxiliary motion objectives on Wan2.2-TI2V-5B at the matched step-48 endpoint. Values are improvements (%) over the native backbone; positive is better.

Metric	Local RPE	Kinematic	RPE + kinematic
Local pose and global trajectory			
Rotation RPE	+0.132	+0.065	+0.078
Translation-direction RPE	+0.399	+2.002	+1.363
Sim(3) ATE	+2.024	+2.521	+1.074
Endpoint error	+5.602	+4.952	+3.089
Path-ratio error	-0.031	-0.048	-0.024
Motion dynamics			
Speed-profile error	+6.629	+3.536	+2.035
Acceleration error	+6.254	+1.676	+1.424
Jerk error	+7.193	+2.595	+1.352
Group summaries			
Trajectory	+3.525	+2.162	+1.299
Geometry	-0.468	+0.475	+0.643
Temporal	-0.083	-0.033	+0.025
Visual	+0.065	+0.058	+0.047

C.4 Additional 5B Configurations

Tables 11 and 12 collect the 27 evaluated configurations, grouped by update design and training objective. They report trajectory, geometry, temporal, and visual improvements, their mean, and its 95% confidence interval. These recipe comparisons include different continuation lengths; the matched controls are described above. Q/V and Q/K/V/O specify the self-attention projections being updated, and L22 denotes layer 22. Short configuration identifiers distinguish the remaining training recipes.

Table 11: 5B update and weight-averaging configurations. Values are improvements (%) over the native backbone. Q/V–L22 blends use the indicated mixing coefficient.

Configuration	Traj.	Geom.	Temp.	Visual	Mean	95% CI
Q/V updates and layer scaling						
Q/V balanced	+2.223	-0.609	+0.153	+0.036	+0.451	[-0.217, +1.090]
Q/V, L22 scale 0.25	+2.209	+0.761	+0.086	+0.021	+0.769	[+0.090, +1.454]
Q/V, L22 scale 0.50	+1.750	-0.506	+0.218	+0.059	+0.380	[-0.422, +1.174]
Q/V, long, unscaled	+0.292	-0.378	+0.150	+0.045	+0.027	[-0.679, +0.721]
Attenuated L22	+1.025	-0.508	+0.139	+0.048	+0.176	[-0.611, +0.989]
Attenuated L22, long	+0.891	+0.255	+0.262	+0.045	+0.363	[-0.248, +0.969]
Weight averaging						
Q/V–L22 blend, 0.25	+1.898	-1.348	+0.107	+0.034	+0.173	[-0.547, +0.885]
Q/V–L22 blend, 0.50	+2.615	-0.624	+0.241	+0.062	+0.574	[-0.110, +1.193]
Q/V–L22 blend, 0.75	+1.915	-0.758	+0.077	+0.031	+0.316	[-0.497, +1.114]
`soup12_75`	+1.420	-0.342	+0.143	+0.077	+0.324	[-0.567, +1.224]

Table 12: 5B objective and continuation configurations. Values follow Table 11. Matched step-48 motion objectives are detailed in Table 10.

Configuration	Traj.	Geom.	Temp.	Visual	Mean	95% CI
Motion objectives						
Local RPE, step 48	+3.525	-0.468	-0.083	+0.065	+0.760	[-0.142, +1.626]
Kinematic, step 48	+2.162	+0.475	-0.033	+0.058	+0.666	[-0.078, +1.328]
RPE + kinematic, step 48	+1.299	+0.643	+0.025	+0.047	+0.503	[-0.148, +1.150]
Aligned RPE, step 72	+1.728	+1.987	+0.172	+0.040	+0.982	[+0.150, +1.796]
Aligned kinematic, step 72	+0.249	-1.373	-0.008	+0.042	-0.273	[-1.024, +0.489]
Aligned all, step 72	+0.236	+0.095	+0.199	+0.039	+0.142	[-0.486, +0.843]
Geometry, appearance, and other objectives						
Geometry–visual, s12	+1.763	-0.430	-0.059	+0.053	+0.332	[-0.500, +1.149]
High-frequency, light	+1.020	-0.277	+0.071	+0.042	+0.214	[-0.630, +0.982]
High-frequency, s8	-0.209	-0.519	+0.074	+0.079	-0.144	[-0.852, +0.525]
`current`	+0.951	+0.137	-0.277	-0.115	+0.174	[-0.636, +0.967]
`gt_full`	+1.310	-0.390	-0.200	-0.049	+0.168	[-0.621, +0.928]
`gt_half`	+0.089	-0.804	-0.205	-0.037	-0.239	[-0.965, +0.473]
`gt_preserve6k`	-0.774	-1.392	-0.538	-0.087	-0.698	[-1.393, +0.033]
`dual_short`	+0.941	-0.553	-0.184	-0.001	+0.051	[-0.812, +0.913]
`dual_flowmatch1`	+0.451	-0.868	-0.022	-0.042	-0.120	[-0.909, +0.628]
`dual_aux015`	+0.306	-0.639	-0.412	-0.069	-0.203	[-1.018, +0.556]
Dual Q/K/V/O	-0.895	-0.978	-0.025	+0.048	-0.462	[-1.318, +0.377]

Table 13: Fast-WAM transfer configuration on RoboCasa365 Target100.

Setting	Configuration
Pretraining data	Human300, 100 demonstrations per task
Target demonstrations	500 per task (Target100)
Evaluation scenes	Held-out target kitchens
Pretraining / evaluated fine-tuning step	60k / 50k
Fine-tuning global batch	128
Fine-tuning optimizer	AdamW, weight decay 0.01, gradient clipping 1.0
Fine-tuning learning rate	3×10^{-5}, 3000-step warmup, cosine decay
Action / state dimension	12 / 16
Input cameras	3
Predicted action chunk / executed prefix	32 / 8
Denoising steps	10
Quantitative evaluation	18 Atomic-Seen tasks, 50 rollouts per task
Evaluation seeds	10000–10049
Simulation renderer	CPU llvmpipe