LEARNING FACTORIZED WORLD TRANSITIONS FOR GEOMETRY-AWARE VIDEO WORLD MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Video world models must preserve both physical structure and visual detail when predicting future observations. Pretrained geometry models provide valuable structural knowledge, but geometry alone does not specify the full visual prediction. This raises the question of how complementary representations should be organized for world-transition modeling. We introduce UPCAST, a framework that learns a geometry-anchored shared physical state together with a private appearance state. Cross-modal reconstruction relates transitions from different representations, while invertible coupling combines their shared and complementary information. A conditional teacher and selective knowledge transfer consolidate the learned representation into the original autoregressive video backbone, without auxiliary modules at inference. On camera-conditioned RealEstate10K, UPCAST reduces 64-frame FVD from 534.9 to 497.8 and JEDi from 4.40 to 3.88 relative to Geometry Forcing. Zero-shot ARKitScenes evaluation shows improved RGB fidelity with mixed sensor-referenced geometric outcomes under domain shift. An action-conditioned Minecraft evaluation further demonstrates applicability beyond camera control. Representation probes and training-stage analysis provide complementary evidence for the learned factorization and its transfer to standalone generation. Together, these results support organizing shared structure and complementary visual information within a predictive transition representation. Project page: https://upcast-anon.github.io/

1 INTRODUCTION

A central goal of world models is to predict how the world changes from its observations. For video generation, this requires maintaining spatial structure while preserving appearance across successive predictions. Visual and geometric representations describe different aspects of this process: geometry provides structure and correspondence, while visual features retain texture, illumination, and semantic detail. The representation used for prediction must organize these aspects together, even when no single observation fully specifies the underlying transition.

Recent geometry-aware video models introduce structural knowledge through explicit 3D outputs or training-time representation supervision (Zhang et al., 2025a; Zhen et al., 2025; Wu et al., 2025a). Geometry Forcing, for example, aligns projected diffusion features with the directional and scale-related information of a pretrained geometry model. This establishes the value of geometric supervision without additional inference modules. Flow Matching continues to supervise visual generation, but the standard formulation leaves the allocation of geometry-compatible structure and complementary visual information implicit.

In this work, we ask: *How should complementary observations be organized into a predictive world-transition representation?* We study this question using video features and geometry-derived representations of the same frame pair. These representations provide complementary views of a transition, with different pretrained knowledge and predictive roles. Our objective is to learn a common structure across them while retaining useful information specific to the visual prediction. Factorization provides a way to organize this information without requiring every aspect of a generated future to pass through the same geometric representation.

We propose UPCAST (Unified Physical and Complementary Appearance State Transitions), a factorized world-transition framework with a shared physical state and a private appearance state (Fig. 1). A geometry-anchored codebook gives physical transitions a common discrete vocabulary,

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.

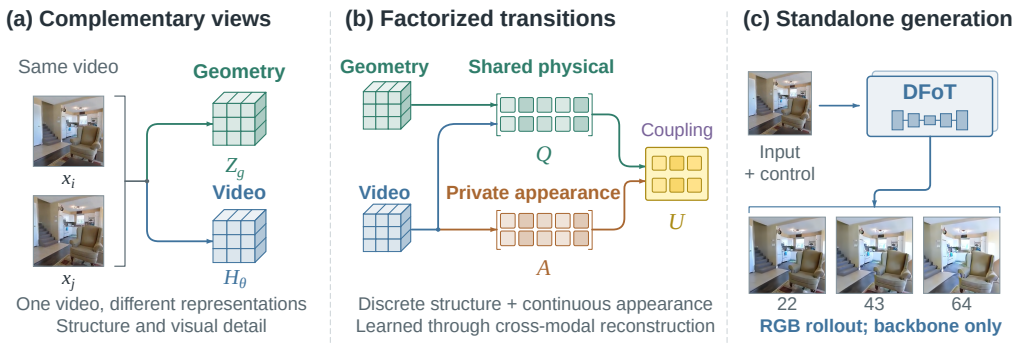


Figure 1: Organizing complementary representations for prediction. (a) The same video provides visual features and geometry-derived structure. (b) Cross-modal reconstruction organizes shared physical and private appearance states; invertible coupling forms the unified state $\mathbf{U}$ (yellow). (c) Knowledge transfer retains their predictive benefit in a standalone video backbone. Colored representations are schematic; RGB thumbnails are actual input and generated frames.

while a continuous visual pathway retains complementary appearance information. Cross-modal reconstruction requires the video, geometry-semantic, and fused branches to predict the same future targets. Invertible coupling then organizes interactions between the two states without an additional dimensional bottleneck. The factorization is functional: the states are defined by their supervision and predictive roles, rather than by an assumption of strict statistical independence.

Learning this auxiliary space is only part of the problem: its predictive benefit must remain after the auxiliary modules are removed. We therefore organize training into representation acquisition, conditional teaching, and backbone consolidation. The video model participates in learning both physical and appearance transitions. A conditional teacher subsequently uses fused physical evidence to produce predictive corrections, and selective distillation transfers useful corrections into the original backbone. Final generation retains the standard autoregressive video interface.

We evaluate the formulation through representation probes, training-stage analysis, and video generation. On RealEstate10K (Zhou et al., 2018), 64-frame rollouts span four times the training horizon and improve FVD, FID, PSNR, JEDi (Luo et al., 2024), and RPE over Geometry Forcing; 128/256-frame rollouts examine degradation under more extreme extrapolation. Zero-shot ARKitScenes (Baruch et al., 2021) evaluation probes the balance between RGB fidelity and sensor-referenced geometry under domain shift, while Minecraft tests action-conditioned generation. These evaluations connect the organization of the training-time representation with the behavior of the deployed video model.

Our contributions are summarized as follows:

- We formulate geometry-aware video modeling as organizing shared transition structure together with complementary visual information for prediction.
- We introduce a geometry-anchored factorized representation, learned through cross-modal reconstruction and coupled interaction, and consolidate its predictive benefit into a standalone video backbone.
- We evaluate cross-modal correspondence, functional roles, and backbone transfer alongside long-horizon generation, zero-shot RGB-D geometry, and action conditioning.

2 Related Work

2.1 Autoregressive Video World Models

World models study both latent dynamics and observation prediction. Dreamer learns behavior through imagined latent trajectories (Hafner et al., 2020), while VideoGPT, Genie, and GameNGen model video under latent or explicit controls (Yan et al., 2021; Bruce et al., 2024; Valevski et al., 2024). Diffusion Forcing and DFoT support history-conditioned generation with per-token noise

levels (Chen et al., 2024; Song et al., 2025). Self-generated training, longer context, retrieval, and memory address exposure to generated history (Huang et al., 2025; Po et al., 2025; Xiao et al., 2025; Wu et al., 2025b). Our focus is the organization of complementary representations used to train such a predictive model.

2.2 Geometry-Aware World Modeling

Explicit approaches generate RGB with depth, point maps, or persistent 3D state (Zhang et al., 2025a; Zhen et al., 2025; Jiang et al., 2025; Ren et al., 2025; Liu et al., 2025). Implicit approaches supervise video features during training: Geometry Forcing aligns geometric features, GEM-4D distills 4D correspondence, and PhysisForcing aligns trajectories and inter-region relations (Wu et al., 2025a; Zhou et al., 2026; Zhang et al., 2026). Our method jointly organizes geometry-anchored transitions and complementary visual information, with reconstruction defining their predictive roles.

Geometry foundation models provide access to learned 3D knowledge without target-domain 3D labels. VGGT estimates cameras, depth, and point maps, while PAGE-4D extends geometric representations to dynamic scenes (Wang et al., 2025; Zhou et al., 2025). In our framework, this knowledge anchors a training-time transition space without requiring explicit 3D reconstruction at deployment.

2.3 Representation Alignment and Unified Latent Spaces

REPA and VideoREPA align generative features with pretrained encoders, while representation autoencoders use semantic features as generative latents (Yu et al., 2024; Zhang et al., 2025b; Zheng et al., 2025). DINOv2 and V-JEPA study visual representations through image structure and video feature prediction, respectively (Oquab et al., 2023; Bardes et al., 2024). Shared/private representations also appear in domain adaptation (Bousmalis et al., 2016), and multimodal models organize heterogeneous inputs through shared token spaces (Mizrahi et al., 2023; Chameleon Team, 2024). UniT learns a visual-anchored physical language for cross-embodiment action transfer through cross-reconstruction (Chen et al., 2026). We study a different requirement: preserving both shared structure and complementary visual detail in generative transitions, then internalizing their predictive benefit in an auxiliary-free video backbone.

3 Factorized World-Transition Modeling

3.1 Method Overview

Predictive Transition Representation. Given an ordered frame pair and controls, we learn a transition representation for future visual and geometric prediction. We build on DFoT (Song et al., 2025) with Flow Matching (Lipman et al., 2023). Frozen PAGE-4D and DINOv2 provide complementary learned views of the same RGB video, not additional sensor inputs; V-JEPA adds predictive regularization. The pairwise representation is training-only, while DFoT remains the autoregressive predictor. See Appendix A for preliminaries.

Learning and Transfer. Figure 2 summarizes three stages. We *acquire* geometry-anchored shared codes and private appearance through cross-modal reconstruction, *teach* a conditional model to use fused physical evidence, then *consolidate* useful corrections into the original backbone before rollout refinement. Thus training organizes complementary knowledge without changing the deployment interface. Appendix C gives the schedule.

3.2 Shared Physical and Private Appearance States

Pairwise Spatial Transition Encoding. We sample an ordered pair (i, j) with $i < j$ and displacement $\Delta = j - i$ for every branch. Let $\mathbf{H}_i, \mathbf{H}_j$ be intermediate noisy DFoT features and ℓ_i, ℓ_j their log signal-to-noise ratios. Video and geometry transitions are

$$\begin{aligned} \mathbf{S}_v &= E_v(\mathbf{H}_i, \mathbf{H}_j, \Delta, \ell_i, \ell_j), \\ \mathbf{S}_g &= E_g(\mathbf{Z}_{g,i}, \mathbf{P}_i, \mathbf{Z}_{g,j}, \mathbf{P}_j, \Delta), \end{aligned} \tag{1}$$

where $\mathbf{S}_v, \mathbf{S}_g \in \mathbb{R}^{M \times D}$ are physical-transition tokens. Each encoder combines learned queries and both frames to model global camera motion and correspondence. At each spatial location, an

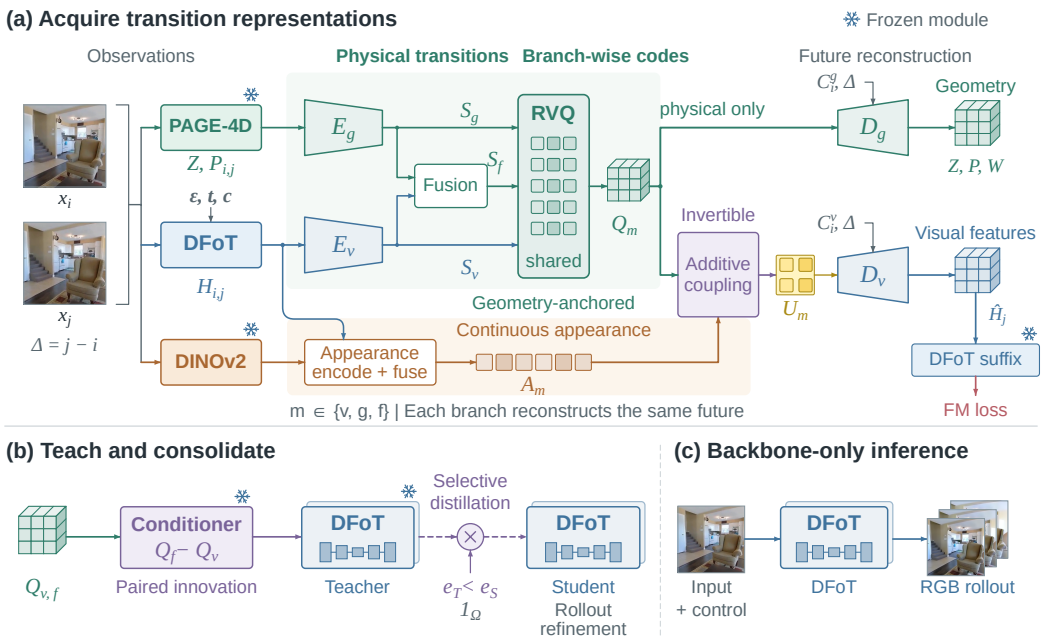


Figure 2: Learning and deploying factorized transitions. (a) Acquire: physical branches share a geometry-anchored RVQ; continuous appearance bypasses it. DFoT features participate in both pathways. Coupling produces $\mathbf{U}_m$ (yellow) for visual reconstruction, while geometry reconstruction reads $\mathbf{Q}_m$. Current anchors and Δ condition both decoders. (b) Teach and consolidate: a conditioner learns to use $\mathbf{Q}_f - \mathbf{Q}_v$; the trained teacher is frozen for selective distillation, followed by rollout refinement. The gate uses Flow Matching advantage $e_T < e_S$ and valid signal $\mathbf{1}_\Omega$. (c) Inference retains only DFoT. Snowflakes mark frozen modules; token counts are schematic.

MLP projects the current feature, future feature, and their difference into a residual added to the corresponding query output before normalization, preserving local change.

The visual branch obtains $\mathbf{A}_v$ from DFoT and $\mathbf{A}_s$ from DINOv2 patches. Modality-fusion encoders form

$$\mathbf{S}_f = F_p(\mathbf{S}_v, \mathbf{S}_g), \qquad \mathbf{A}_f = F_a(\mathbf{A}_v, \mathbf{A}_s). \tag{2}$$

An explicit residual preserves the geometry transition in F_p, while video evidence enriches it. DFoT therefore participates in both states, beyond receiving external supervision.

Shared Physical State. The video, geometry, and fused physical transitions are mapped into a shared residual vector-quantized space:

$$\mathbf{Q}_m = \pi_\uparrow \mathcal{Q}\left(\pi_\downarrow(\mathbf{S}_m)\right), \qquad m \in \{v, g, f\}, \tag{3}$$

where $\mathcal{Q}$ is a shared residual vector quantizer and $\pi_\downarrow, \pi_\uparrow$ map to and from its compact code dimension. Geometry initializes the codebook before video and fused branches learn its hard, soft, and pre-quantization structure. The resulting common vocabulary is shared across branches, not invariant to the transition.

Private Appearance State. Geometry alone does not specify texture, illumination, and semantic detail. A continuous private state carries this information outside the physical codebook. For the video, geometry-semantic, and fused branches, $\mathbf{A}_m \in \{\mathbf{A}_v, \mathbf{A}_g, \mathbf{A}_f\}$ with $\mathbf{A}_g = \mathbf{A}_s$. Invertible additive coupling combines it with $\mathbf{Q}_m$:

$$\begin{aligned} \overline{\mathbf{Q}}_m &= \mathbf{Q}_m + r_{a\to p}(\mathbf{A}_m), \\ \overline{\mathbf{A}}_m &= \mathbf{A}_m + r_{p\to a}(\overline{\mathbf{Q}}_m), \\ \mathbf{U}_m &= \left[\overline{\mathbf{Q}}_m; \overline{\mathbf{A}}_m\right]. \end{aligned} \tag{4}$$

The inverse subtracts the residuals in reverse order, preserving joint information without dimensional reduction. Unlike cross-source fusion in Eq. 2, coupling acts within a branch. The state names indicate supervision and function, not strict statistical disentanglement.

Cross-Modal Transition Reconstruction. Reconstruction conditions on the current observation, leaving transition tokens to describe change. The visual anchor $\mathbf{C}_i^v$ combines DINOv2 patches with a trainable pixel stem; the geometry anchor $\mathbf{C}_i^g$ combines PAGE-4D patches and normalized point maps. The decoders predict

$$\begin{aligned} D_v(\mathbf{U}_m, \mathbf{C}_i^v, \Delta) &= (\widehat{\mathbf{Z}}_{d,j}^m, \widehat{\mathbf{H}}_j^m), \\ D_g(\mathbf{Q}_m, \mathbf{C}_i^g, \Delta) &= (\widehat{\mathbf{Z}}_{g,j}^m, \widehat{\mathbf{P}}_j^m, \widehat{\mathbf{W}}_j^m). \end{aligned} \tag{5}$$

All branches reconstruct the same future targets. The geometry decoder reads only $\mathbf{Q}_m$; the visual decoder reads $\mathbf{U}_m$, defining their functional roles. Visual losses compare DINOv2 features and apply Flow Matching through a frozen DFoT suffix. Geometry losses cover PAGE-4D features, confidence-weighted points, gradients, normals, and confidence. Current-observation dropout and transition-only reconstruction discourage shortcuts.

The unified-space objective is summarized as

$$\mathcal{L}_{\text{space}} = \lambda_v \mathcal{L}_{\text{visual}} + \lambda_g \mathcal{L}_{\text{geometry}} + \lambda_t \mathcal{L}_{\text{tail}} + \lambda_q \mathcal{L}_{\text{code}} + \lambda_r \mathcal{L}_{\text{reg}}, \tag{6}$$

Here $\mathcal{L}_{\text{code}}$ covers shared codes and commitment; $\mathcal{L}_{\text{reg}}$ covers retrieval, shuffled pairs, transition-only prediction, and variance. Frozen DINOv2 and V-JEPA additionally supervise visual structure and predictive changes during joint refinement (Appendix C).

3.3 Knowledge Transfer to Autoregressive Video Models

Teaching with Fused Physical Evidence. The complete model provides video-only $\mathbf{Q}_v$ and fused $\mathbf{Q}_f$ physical transitions. After freezing it, we train a conditioner at the same DFoT layer. It reads a noisy physical state and paired tokens, models the innovation $\mathbf{Q}_f - \mathbf{Q}_v$, and applies a bounded, learned-gate residual update:

$$\widetilde{\mathbf{H}} = \mathbf{H} + \gamma_\phi C_\phi(\mathbf{H}, \widetilde{\mathbf{Q}}_f, \mathbf{Q}_v, \mathbf{Q}_f, \ell_j). \tag{7}$$

The conditioned feature is processed by the original DFoT suffix. Let e_b, e_c, and e_s denote the future-frame Flow Matching errors of the unconditioned prediction, the correctly conditioned prediction, and a prediction using shuffled fused tokens. Writing $\bar{e}_k = e_k / \max(e_b, \varepsilon)$, the principal teacher terms are

$$\mathcal{L}_{\text{teacher}} = \bar{e}_c + \lambda_b[\bar{e}_c - \bar{e}_b + \mu_b]_+ + \lambda_s[\bar{e}_c - \bar{e}_s + \mu_s]_+ + \mathcal{L}_{\text{aux}}, \tag{8}$$

The rank terms favor the correct physical transition over both base and shuffled conditions; $\mathcal{L}_{\text{aux}}$ controls shuffled-condition corrections and pair compatibility. Appendix C gives the normalization and active terms. This rewards transition-specific evidence over an unconditional residual.

Backbone Consolidation. After the conditional teacher is trained, we freeze it and transfer its advantage into a student without the physical conditioner. Standard camera or action controls are retained. For each sample, we compute

$$a = \text{clip}\left(\frac{e_{\text{student}} - e_{\text{teacher}}}{\max(e_{\text{student}}, \varepsilon)}, 0, a_{\max}\right), \tag{9}$$

and distill the teacher output only when it improves the student and the sampled noise level contains sufficient signal. The distillation loss is

$$\mathcal{L}_{\text{KT}} = \mathbb{E}\left[\mathbf{1}_\Omega a \, w_{\text{motion}} \, \|v_{\text{student}} - v_{\text{teacher}}\|_{\text{RMS}}\right], \tag{10}$$

Here Ω is the valid signal range and w_{motion} shifts a fixed distillation budget toward larger visual or camera changes. Selection avoids imitating uninformative corrections.

Overall Training and Inference. The student combines Flow Matching with selective transfer, backbone preservation, and generated-history rollout supervision; the complete objective is given in Appendix D. At inference, PAGE-4D, DINOv2, V-JEPA, the RVQ, fusion modules, decoders, and conditioner are removed, and sampling uses only the consolidated DFoT backbone.

4 Experiments

Our evaluation examines the deployed model and the representation learned during training. RealEstate10K (Zhou et al., 2018) tests camera-conditioned long-horizon generation; ARKitScenes (Baruch et al., 2021) probes zero-shot RGB fidelity and sensor-referenced geometry under domain shift; Minecraft tests action conditioning. Training-stage analysis and representation

Table 1: Camera-conditioned RealEstate10K generation. The 64-frame test uses 100 matched videos (RPE on its first 12); 128/256-frame tests use a separate 12-video cohort. All models train on 16-frame clips. Best and second-best values per horizon are bold and underlined, respectively.

Method	Frames	FVD ↓	FID ↓	LPIPS ↓	SSIM ↑	PSNR ↑	JEDi ↓	RPE ↓
DFoT	64	654.0	69.46	0.440	0.474	<u>13.64</u>	5.65	<u>0.176</u>
REPA	64	643.6	69.42	0.457	0.463	13.33	6.77	0.283
VideoREPA	64	795.4	76.09	0.485	0.410	11.48	9.98	0.282
Geometry Forcing	64	<u>534.9</u>	<u>65.15</u>	**0.391**	**0.522**	12.44	<u>4.40</u>	0.192
UPCAST	64	**497.8**	**61.26**	<u>0.396</u>	<u>0.498</u>	**13.94**	**3.88**	**0.161**
DFoT	128	1136.3	153.9	0.608	0.362	<u>10.91</u>	–	–
REPA	128	1258.5	162.2	0.632	0.312	10.31	–	–
VideoREPA	128	1068.1	142.9	0.619	0.290	9.75	–	–
Geometry Forcing	128	<u>924.7</u>	**128.1**	<u>0.549</u>	**0.408**	9.13	–	–
UPCAST	128	**803.6**	<u>136.5</u>	**0.541**	<u>0.390</u>	**11.28**	–	–
DFoT	256	1711.5	214.0	0.717	0.298	<u>9.32</u>	18.88	–
REPA	256	2065.9	220.0	0.750	0.238	8.77	32.68	–
VideoREPA	256	1378.0	201.7	0.705	0.237	8.88	16.22	–
Geometry Forcing	256	<u>1221.7</u>	**175.1**	**0.643**	**0.373**	7.73	<u>15.45</u>	–
UPCAST	256	**1217.6**	<u>189.7</u>	<u>0.664</u>	<u>0.340</u>	**9.59**	**12.83**	–

probes then examine knowledge transfer, cross-modal correspondence, and the functional roles of the shared and private states.

Implementation Details. We train DFoT on 16-frame, 256×256 clips conditioned on the first frame and camera poses. PAGE-4D, DINOv2, and V-JEPA remain frozen. Generation uses the final backbone, whereas probes use the complete model before consolidation. Appendix C gives the schedule and checkpoint roles.

Evaluation Metrics. We report FVD (Unterthiner et al., 2018), FID (Heusel et al., 2017), LPIPS (Zhang et al., 2018), SSIM (Wang et al., 2004), PSNR, JEDi (Luo et al., 2024), and RPE (Wu et al., 2025a) under matched conditions. FID, LPIPS, SSIM, PSNR, and JEDi omit the conditioning frame. The primary 64-frame test has 100 videos (RPE on its first 12); separate 128/256-frame tests use 12 clips. We also report an Ours-only **V-JEPA feature F-Score** on the 100 primary rollouts, the harmonic mean of $k = 3$ nearest-neighbor manifold precision and recall in normalized feature space (Kynkäänniemi et al., 2019). It measures neither pixel nor geometric accuracy; Appendix E details the protocols.

4.1 Main Results

We compare UPCAST with DFoT, REPA, VideoREPA, and Geometry Forcing using matched generation conditions.

Camera View-Conditioned Long-Horizon Video Generation. Starting from a single image, each model recursively generates future clips using the same per-frame camera poses. The primary evaluation uses 64 frames, four times the training horizon. Table 1 also reports 128/256-frame stress tests to examine behavior under more extreme autoregressive extrapolation.

At 64 frames, UPCAST achieves the best FVD, FID, PSNR, JEDi, and RPE among the evaluated methods. Relative to Geometry Forcing, FVD decreases from 534.9 to 497.8 and JEDi from 4.40 to 3.88, while Geometry Forcing retains better LPIPS and SSIM. The improvements therefore span distribution quality, pixel fidelity, and predictive-feature consistency without uniform dominance of all perceptual metrics. The Ours-only **V-JEPA feature F-Score is 90.48%** (89.0% precision, 92.0% recall; Appendix E).

On the first 12 main-test videos, the lower 64-frame RPE provides complementary evidence of cross-view consistency. At 128/256 frames, UPCAST maintains the lowest FVD and PSNR is consistently higher on the separate stress cohort, but perceptual metrics remain mixed. In particular, 256-frame FVD is close to Geometry Forcing, whereas JEDi is lower at 12.83 versus 15.45. Per-frame camera poses remain available throughout each rollout, whereas appearance and scene identity

Table 2: Zero-shot ARKitScenes evaluation on 12 motion-filtered 64-frame clips. Metrics exclude the conditioning frame. Frozen-estimator depth uses one scale per clip; sensor-anchored Pose R@10 uses 10 cm/10° and counts PnP failures as misses. Best values are bold.

	RGB	Depth			Surface	Point Cloud			Sensor-Anchored Pose
Method	PSNR ↑	AbsRel ↓	δ_1 ↑	Scale Drift ↓	Normal MAE (°) ↓	Chamfer-L_1 (m) ↓	F@5 cm ↑	F@10 cm ↑	Pose R@10 ↑
Geometry Forcing	11.85	0.216	0.670	0.177	30.31	**0.101**	0.425	0.621	**0.317**
UPCAST	**14.19**	**0.194**	**0.699**	**0.160**	**29.13**	0.105	**0.439**	**0.627**	0.265

Table 3: Action-conditioned generation on DFoT Minecraft quickstart: 10 training and 10 validation trajectories, 100 matched 16-frame videos. A held-out inverse-dynamics probe (85.73% balanced accuracy) measures action BAcc.; response is the fraction of changed action intents yielding different predicted intents with matched initial frames and seeds. Both variants use backbone-only inference.

Method	FVD ↓	FID ↓	LPIPS ↓	SSIM ↑	PSNR ↑	Action BAcc. (%) ↑	Response (%) ↑
Action-DFoT	245.0	39.51	0.431	0.524	15.83	85.10	86.89
Action-DFoT + UPCAST	**221.7**	**37.55**	**0.424**	**0.528**	**15.97**	**85.35**	**89.78**

must be carried through generated history; this distinction may contribute to different long-horizon behavior of camera control and visual consistency. These stress tests characterize long-horizon degradation rather than serving as the primary performance comparison.

Sensor-Referenced Evaluation on ARKitScenes. RealEstate10K lacks synchronized metric depth for target views. We therefore evaluate the same checkpoints zero-shot on ARKitScenes, which supplies RGB, LiDAR depth, intrinsics, and poses. Before generation, sensor motion filters select one 64-frame clip from each of 12 videos within the RealEstate10K motion envelope. Both methods share the initial frame, camera path, and seed; neither is fine-tuned. All metrics exclude the conditioning frame.

PSNR measures RGB fidelity. A frozen metric-indoor Depth Anything V2 model (Yang et al., 2024) estimates generated-frame depth against synchronized sensor depth; one scale per clip preserves temporal scale variation for AbsRel, δ_1, and Scale Drift. Back-projected depth yields normal error, symmetric Chamfer-L_1, and point-cloud F-scores. Sensor-depth keypoints and PnP-RANSAC test whether generated views realize the target pose; Pose R@10 counts failures as misses. Appendix E specifies the metrics and thresholds.

On these clips, UPCAST improves PSNR from 11.85 to 14.19 dB and AbsRel from 0.216 to 0.194; δ_1, scale drift, normal error, and point-cloud F-scores also improve. Applying the same frozen estimator to real RGB gives AbsRel 0.125 and F@10 cm 0.752, contextualizing this cross-domain evaluation.

Geometry Forcing retains lower Chamfer distance and higher pose recall. Thus RGB and depth gains coexist with mixed geometric outcomes under domain shift, rather than uniform geometric superiority.

Action-Conditioned Video Generation. To test applicability beyond camera control, we apply unified-space transfer to Action-DFoT on the official DFoT Minecraft quickstart split (Baker et al., 2022). Table 3 shows improvements in all reported generation metrics, action balanced accuracy, and response under paired action interventions. This small-scale extension supports using the transition-learning framework with action-conditioned dynamics while retaining backbone-only inference.

4.2 Qualitative Results

Figure 3 presents two matched examples on RealEstate10K. As autoregressive generation proceeds, the baseline methods exhibit accumulated color shifts, overexposure, and local structural artifacts. In contrast, UPCAST maintains more stable object appearance and scene structure through the final generated frame. These results further demonstrate the benefit of unified physical transition modeling for long-horizon scene generation.



Figure 3: Qualitative comparison of camera view-conditioned 64-frame video generation on RealEstate10K. All methods are generated from the same input frame using matched camera trajectories and sampling seeds. Compared with the baselines, UPCAST better preserves scene layout and appearance while reducing accumulated color and texture drift.

Table 4: Training-stage analysis on 25 matched 64-frame rollouts. The stages follow one checkpoint lineage, and all variants use the same backbone-only inference path.

Training stage	FVD ↓	FID ↓	LPIPS ↓	SSIM ↑	PSNR ↑	JEDi ↓
Video backbone	869.8	88.30	0.414	0.505	13.78	7.88
Unified-space learning	**674.4**	**74.69**	0.373	0.522	14.22	6.59
Backbone consolidation	708.9	78.18	**0.361**	**0.534**	14.29	6.05
Rollout refinement	707.5	77.87	**0.361**	**0.534**	**14.30**	**5.90**

4.3 Training-Stage Analysis

Table 4 follows one checkpoint lineage on matched clips, trajectories, and seeds under backbone-only inference. It separates transition-space learning, consolidation, and generated-history adaptation in the deployed model.

Does a Unified Physical Space Improve Long-Horizon Generation? Cross-modal reconstruction organizes video, geometry, and fused transitions. Unified-space learning lowers FVD from 869.8 to 674.4 and JEDi from 7.88 to 6.59, improving every reported metric over the original backbone. Because evaluation already uses the backbone alone, these gains appear before conditional distillation and rollout refinement.

Can the Unified Representation Be Consolidated into the Video Backbone? Consolidation transfers fused-evidence corrections into the standalone backbone. LPIPS, SSIM, PSNR, and JEDi improve, while FVD and FID worsen: JEDi falls from 6.59 to 6.05 as FVD rises from 674.4 to 708.9. Thus transfer preserves benefits without a uniform distribution-quality gain or test-time auxiliary modules.

Does Motion-Aware Refinement Reduce Autoregressive Drift? Generated-history refinement lowers JEDi from 6.05 to 5.90 and slightly improves FVD, FID, and PSNR; LPIPS and SSIM are unchanged at reported precision. Most FVD improvement precedes this stage, which adapts the consolidated backbone to recursive generation without changing inference.

Table 5: Unified-space probes on 192 held-out transitions. V_p, G_p, and F_p denote video, geometry, and fused physical states; V_a, S_a, and F_a denote video, semantic, and fused private states.

State	Retrieval direction	Top-1 (%) ↑	Top-5 (%) ↑	CKA ↑	Span Top-1 (%) ↑
Physical	$V_p \to G_p$	13.54	43.23	0.676	46.35
	$V_p \to F_p$	13.02	53.65	0.679	38.54
	$G_p \to F_p$	35.94	96.35	0.967	42.71
Private	$V_a \to S_a$	70.31	94.27	0.388	100.0
	$V_a \to F_a$	71.35	94.79	0.390	100.0
	$S_a \to F_a$	100.0	100.0	0.994	100.0
Chance	n/a	0.52	2.60	n/a	33.33

4.4 Representation Analysis

Probes on the frozen complete model before consolidation test cross-modal correspondence, complementary appearance, and temporal content. They characterize functional organization, not strict statistical disentanglement; generation uses the final backbone alone.

Cross-Modal Correspondence. Pooled transition tokens retrieve the same video and temporal span across modalities. The 64 videos and three spans give 192 candidates and 0.52% chance Top-1 accuracy.

Table 5 shows above-chance physical correspondence: video-to-geometry retrieval reaches 13.54% Top-1 and CKA 0.676, while geometry-to-fused CKA reaches 0.967. These results link separately encoded transitions and the geometry-anchored fusion design.

Complementary Private Appearance State. Private-state retrieval reaches 70.31%–100.0% Top-1. Under photometric perturbations, its relative feature change is 2.1 times the physical state's, with shared-to-private CKA of 0.242–0.270. Removing private versus physical input raises frozen visual-decoder loss by 45.1% versus 4.7%. Thus the private pathway retains distinct, visually useful information.

Temporal Transition Information. With scene identity fixed, all three physical retrieval directions exceed 33.33% chance; video-to-geometry reaches 46.35%. This same-scene control tests temporal-span correspondence without relying on differences between scenes. A linear probe predicts held-out camera transitions from video physical tokens with $R^2 = 0.468$. These results complement the backbone-only generation analysis.

Taken together, cross-modal and temporal retrieval connect the physical states to corresponding transitions, while photometric sensitivity and decoder interventions indicate that the private state retains complementary visual information. Table 4 further shows that improvements emerge before final rollout refinement and persist after consolidation into the standalone backbone.

5 Conclusion

We presented UPCAST, a framework for organizing complementary visual and geometric representations into predictive world transitions. Geometry-anchored shared codes and continuous private appearance codes are learned through cross-modal reconstruction, coupled for visual prediction, and consolidated into the original video backbone. Representation probes characterize their cross-modal correspondence and functional roles, while generation and training-stage evaluations show benefits that remain under backbone-only inference.

The results support treating geometric knowledge as part of a predictive representation together with complementary visual information. Long-horizon perceptual metrics and zero-shot geometric measures exhibit trade-offs, and the sensor-grounded evaluation remains limited to indoor scenes and RGB-based estimators. Extending this formulation to broader scene dynamics and additional observations or controls is a direction for future work; the present study establishes its realization for camera- and action-conditioned video generation.

Ethics Statement

This study uses existing scene datasets and collects no new human-subject data. Indoor video can contain private or identifying details; any release of visual examples or artifacts should respect the source datasets' terms and avoid exposing such details. Camera- and action-conditioned video generators could also be used to create misleading synthetic footage. Our experiments examine predictive modeling and do not present generated videos as authentic records.

AI Use Statement

In this work, we used generative AI tools to assist with method and evaluation code, dataset preparation, related-work discovery, manuscript drafting, translation and language editing, and scientific figure preparation and revision. We did not use these tools to generate synthetic datasets; formal proof-related tasks were not applicable to this empirical study. Additionally, we used these tools to refine the conceptual framing and hypotheses, evaluation design, and interpretation of results. We reviewed all AI-assisted work, checking literature suggestions against the original sources, code and figures against the intended method, and numerical statements against evaluation records. The authors made the final research decisions and take responsibility for the claims, results, and artifacts in this paper.

Reproducibility Statement

Section 3 and Appendix D describe the model formulation and training objectives. Appendix C records the phase-wise optimization schedule and checkpoint roles. Section 4 and Appendix E document the matched generation conditions, clip selection, and evaluation protocol.

REFERENCES

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. *arXiv preprint arXiv:2301.08243*, 2023.

Bowen Baker, Ilge Akkaya, Peter Zhokhov, Joost Huizinga, Jie Tang, Adrien Ecoffet, Brandon Houghton, Raul Sampedro, and Jeff Clune. Video pretraining (VPT): Learning to act by watching unlabeled online videos. *arXiv preprint arXiv:2206.11795*, 2022.

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

Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Tal Dimry, Yuri Feigin, Peter Fu, Thomas Gebauer, Brandon Joffe, Daniel Kurz, Arik Schwartz, and Elad Shulman. ARKitScenes: A diverse real-world dataset for 3d indoor scene understanding using mobile RGB-D data. In *NeurIPS Datasets and Benchmarks Track*, 2021.

Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan. Domain separation networks. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

Jake Bruce, Michael Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, Yusuf Aytar, Sarah Bechtle, Feryal Behbahani, Stephanie 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*, 2024.

Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024.

Boyu Chen, Yi Chen, Lu Qiu, Jerry Bai, Yuying Ge, and Yixiao Ge. UniT: Toward a unified physical language for human-to-humanoid policy learning and world modeling. *arXiv preprint arXiv:2604.19734*, 2026.

Boyuan Chen, Diego Martí Monsó, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion forcing: Next-token prediction meets full-sequence diffusion. In *Advances in Neural Information Processing Systems*, volume 37, pp. 24081–24125, 2024.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. *arXiv preprint arXiv:1912.01603*, 2020.

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

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. *arXiv preprint arXiv:2506.08009*, 2025.

Zeren Jiang, Chuanxia Zheng, Iro Laina, Diane Larlus, and Andrea Vedaldi. Geo4D: Leveraging video generators for geometric 4D scene reconstruction. *arXiv preprint arXiv:2504.07961*, 2025.

Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

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

Zeyi Liu, Shuang Li, Eric Cousineau, Siyuan Feng, Benjamin Burchfiel, and Shuran Song. Geometry-aware 4D video generation for robot manipulation. *arXiv preprint arXiv:2507.01099*, 2025.

Ge Ya Luo, Gian Favero, Zhi Hao Luo, Alexia Jolicoeur-Martineau, and Christopher Pal. Beyond FVD: Enhanced evaluation metrics for video generation quality. *arXiv preprint arXiv:2410.05203*, 2024.

David Mizrahi, Roman Bachmann, Oğuzhan Fatih Kar, Teresa Yeo, Mingfei Gao, Afshin Dehghan, and Amir Zamir. 4M: Massively multimodal masked modeling. *arXiv preprint arXiv:2312.06647*, 2023.

Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Ryan Po, Yotam Nitzan, Richard Zhang, Berlin Chen, Tri Dao, Eli Shechtman, Gordon Wetzstein, and Xun Huang. Long-context state-space video world models. *arXiv preprint arXiv:2505.20171*, 2025.

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. *arXiv preprint arXiv:2503.03751*, 2025.

Johannes L. Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4104–4113, 2016.

Kiwhan Song, Boyuan Chen, Max Simchowitz, Yilun Du, Russ Tedrake, and Vincent Sitzmann. History-guided video diffusion. *arXiv preprint arXiv:2502.06764*, 2025.

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

Dani Valevski, Yaniv Leviathan, Moab Arar, and Shlomi Fruchter. Diffusion models are real-time game engines. *arXiv preprint arXiv:2408.14837*, 2024.

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, 2025.

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.

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*, 2025a.

Tong Wu, Shuai Yang, Ryan Po, Yinghao Xu, Ziwei Liu, Dahua Lin, and Gordon Wetzstein. Video world models with long-term spatial memory. *arXiv preprint arXiv:2506.05284*, 2025b.

Zeqi Xiao, Yushi Lan, Yifan Zhou, Wenqi Ouyang, Shuai Yang, Yanhong Zeng, and Xingang Pan. WORLDMEM: Long-term consistent world simulation with memory. *arXiv preprint arXiv:2504.12369*, 2025.

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

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything V2. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. *arXiv preprint arXiv:2410.06940*, 2024.

Peiwen Zhang, Yufan Deng, Shangkun Sun, Juncheng Ma, Duomin Wang, Jonas Du, Zilin Pan, Ye Huang, Hao Liang, Songyan Huang, Ruihua Zhang, Enze Xie, Ming-Yu Liu, and Daquan Zhou. PhysisForcing: Physics reinforced world simulator for robotic manipulation. *arXiv preprint arXiv:2606.28128*, 2026.

Qihang Zhang, Shuangfei Zhai, Miguel Angel 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, 2025a.

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*, pp. 586–595, 2018.

Xiangdong Zhang, Jiaqi Liao, Shaofeng Zhang, Fanqing Meng, Xiangpeng Wan, Junchi Yan, and Yu Cheng. VideoREPA: Learning physics for video generation through relational alignment with foundation models. *arXiv preprint arXiv:2505.23656*, 2025b.

Haoyu Zhen, Qiao Sun, Hongxin Zhang, Junyan Li, Siyuan Zhou, Yilun Du, and Chuang Gan. TesserAct: Learning 4D embodied world models. *arXiv preprint arXiv:2504.20995*, 2025.

Boyang Zheng, Nanye Ma, Shengbang Tong, and Saining Xie. Diffusion transformers with representation autoencoders. *arXiv preprint arXiv:2510.11690*, 2025.

Kaichen Zhou, Yuhan Wang, Grace Chen, Xinhai Chang, Gaspard Beaudouin, Fangneng Zhan, Paul Pu Liang, and Mengyu Wang. PAGE-4D: Disentangled pose and geometry estimation for VGGT-4D perception. *arXiv preprint arXiv:2510.17568*, 2025.

Kaichen Zhou, Yuzhen Chen, Fangneng Zhan, Hang Hua, Grace Chen, Xinhai Chang, Ao Qu, Yilun Du, Zhuang Liu, Paul Pu Liang, and Mengyu Wang. GEM-4D: Geometry-enhanced video world models for robot manipulation. *arXiv preprint arXiv:2605.22882*, 2026.

Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *ACM Transactions on Graphics*, 37(4):65:1–65:12, 2018. doi: 10.1145/3197517.3201323.

APPENDIX FOR ICLR 2027 SUBMISSION *Learning Factorized World Transitions for Geometry-Aware Video World Models*

A Detailed Preliminaries **1**

A.1 Autoregressive Video Diffusion Models 1

A.2 Geometry and Visual Foundation Representations 2

A.3 An Interpretive View of Complementary Transitions 2

B Expanded Related Work **2**

B.1 Autoregressive Video World Models 2

B.2 Geometry-Aware World Modeling 3

B.3 Representation Alignment and Unified Latent Spaces 4

C Full Implementation Details **4**

D Complete Student Objective **6**

E Expanded Evaluation Protocol **7**

F Failure Case Analysis **9**

G Discussion and Limitations **10**

A DETAILED PRELIMINARIES

Our method builds upon an autoregressive video diffusion model and incorporates geometric and visual representations during training to guide learning. In this section, we briefly introduce these components as the foundation of our method.

A.1 AUTOREGRESSIVE VIDEO DIFFUSION MODELS

Training. We formulate world prediction using an autoregressive video diffusion model. Let $\mathbf{x} = \{x_1, \ldots, x_T\}$ denote a video sequence sampled from the data distribution, and let $\mathbf{c} = \{c_1, \ldots, c_T\}$ denote the corresponding frame-wise control signals. Following Diffusion Forcing (Chen et al., 2024), we independently sample a noise level $\mathbf{t} = \{t_1, \ldots, t_T\}$ for each frame and construct the noisy video as

$$x_i^{t_i} = (1 - t_i)x_i + t_i\epsilon_i, \qquad \epsilon_i \sim \mathcal{N}(0, I). \tag{11}$$

The target velocity field of this linear probability path is $\epsilon_i - x_i$. We train a video diffusion model v_θ by minimizing the following objective:

$$\mathcal{L}_{\mathrm{FM}} = \mathbb{E}_{\mathbf{x},\boldsymbol{\epsilon},\mathbf{t}} \left[\frac{1}{T} \sum_{i=1}^{T} \left\| v_\theta \left(\mathbf{x}^{\mathbf{t}}, \mathbf{t}, \mathbf{c} \right)_i - (\epsilon_i - x_i) \right\|_2^2 \right]. \tag{12}$$

This objective follows Flow Matching and DFoT (Lipman et al., 2023; Song et al., 2025).

Autoregressive Sampling. During inference, sampling follows the probability flow ordinary differential equation

$$\mathrm{d}\mathbf{x}^t = v_\theta \left(\mathbf{x}^t, t, \mathbf{c} \right) \mathrm{d}t. \tag{13}$$

After generating a fixed-length video clip, the model reuses a subset of the generated frames as context to recursively predict subsequent clips.

A.2 Geometry and Visual Foundation Representations

Geometry Foundation Representations. PAGE-4D (Zhou et al., 2025) extends feed-forward geometry reconstruction to dynamic scenes and supports camera-pose, depth, and point-map prediction. We use only its Transformer aggregator and point head. Given a clean video $\mathbf{x}$, the resulting geometry teacher E_g produces

$$\left(\mathbf{Z}_g, \widetilde{\mathbf{P}}, \mathbf{W}\right) = \mathrm{sg}\left(E_g(\mathbf{x})\right), \qquad \mathbf{P} = \mathrm{Norm}\left(\widetilde{\mathbf{P}}\right), \tag{14}$$

where $\mathbf{Z}_g$ denotes the PAGE-4D Transformer patch features, $\widetilde{\mathbf{P}}$ denotes the predicted world-coordinate point maps, and $\mathbf{W}$ denotes the corresponding confidence estimates. $\mathrm{Norm}(\cdot)$ centers and scale-normalizes the point maps using statistics from the first frame, while $\mathrm{sg}(\cdot)$ denotes the stop-gradient operation. We use $\mathbf{Z}_g$ as the geometric representation and $\mathbf{W}$ to weight the geometric supervision.

Visual Foundation Representations. We further use DINOv2 (Oquab et al., 2023) and V-JEPA (Bardes et al., 2024) encoders to extract complementary visual representations:

$$\mathbf{Z}_d = \mathrm{sg}\left(\{E_d(x_i)\}_{i=1}^{T}\right), \qquad \mathbf{Z}_j = \mathrm{sg}\left(E_j(\mathbf{x})\right). \tag{15}$$

Here, $\mathbf{Z}_d$ contains dense frame-wise patch representations, whereas $\mathbf{Z}_j$ captures spatiotemporal predictive information across the entire video clip.

A.3 An Interpretive View of Complementary Transitions

The factorization can be read through an idealized latent-variable view. A world transition shapes the observed frame pair, from which the training pathways obtain video, geometry, and semantic representations:

$$\mathcal{T}_{ij} \rightsquigarrow (x_i, x_j) \xrightarrow{\text{representation extraction}} (\mathbf{R}^v_{ij}, \mathbf{R}^g_{ij}, \mathbf{R}^s_{ij}). \tag{16}$$

The first arrow is conceptual; the second reflects the training data flow. DFoT, PAGE-4D, and DINOv2 provide different learned views of the same RGB frame pair, with noise, frame displacement, and controls supplied where appropriate. These views can share structural evidence without containing identical information: geometry describes correspondence and shape, while visual features retain appearance beyond what geometry determines.

In the implemented model, the video branch forms $(\mathbf{Q}_v, \mathbf{A}_v)$, the geometry-semantic branch forms $(\mathbf{Q}_g, \mathbf{A}_s)$, and fusion forms $(\mathbf{Q}_f, \mathbf{A}_f)$ (Eqs. 2–3). The shared codebook anchors $\mathbf{Q}_m$ across branches, while $\mathbf{A}_m$ remains continuous. For transition information, geometry reconstruction reads $\mathbf{Q}_m$, whereas visual reconstruction reads the coupled state $\mathbf{U}_m$; both also receive their current anchors and Δ (Eqs. 4–5). Thus "private" describes a predictive role complementary to the physical codes, not an independent sensor source. Equation 16 does not define a fitted likelihood or posterior. Camera poses specify the requested viewpoint, whereas actions in the Minecraft extension affect the transition itself.

This view yields three diagnostic expectations examined in the paper. Shared physical states should exhibit cross-view and temporal correspondence; private states should retain visual information useful for future reconstruction; and training-time fused evidence should improve the standalone backbone after consolidation. Table 5 probes the first two expectations, while Table 4 examines the third. Together, these measurements test the proposed functional roles and track their transfer into backbone prediction.

B Expanded Related Work

B.1 Autoregressive Video World Models

Latent world models and generative video models address related aspects of prediction. Dreamer learns behavior through trajectories imagined in a learned latent dynamics model (Hafner et al., 2020). Our focus is the organization of training-time transition representations for video prediction, rather than policy learning or planning in a persistent latent state. The transition encoders observe frame pairs during training, while the deployed DFoT predicts future RGB from history and controls.

Autoregressive video generation extends fixed-length video models to long-horizon world simulation. A model first generates a future video clip and then recursively conditions on its own outputs to predict subsequent states. Early latent autoregressive models such as VideoGPT model discrete spatiotemporal video tokens, while interactive world models such as Genie and GameNGen recursively predict environment observations under latent or explicit actions (Yan et al., 2021; Bruce et al., 2024; Valevski et al., 2024). Diffusion Forcing assigns independent noise levels to different positions in a sequence, thereby supporting future prediction, history-conditioned generation, and sequence interpolation within a unified training framework (Chen et al., 2024). DFoT instantiates this paradigm with a Transformer architecture and supports video generation conditioned on control signals such as per-frame camera poses (Song et al., 2025). Since these approaches do not require maintaining an explicit 3D scene during inference, they provide an efficient and scalable foundation for video world modeling.

A central challenge in autoregressive generation is the distribution gap between training and inference. Prior studies have sought to mitigate this problem through noisy history conditioning, training on self-generated samples, long-context modeling, historical frame retrieval, and explicit memory mechanisms (Chen et al., 2024; Huang et al., 2025; Po et al., 2025; Xiao et al., 2025; Wu et al., 2025b). Complementary to these efforts, we focus on the internal world-transition representations of video models, aiming to preserve more stable physical states during recursive generation.

B.2 Geometry-Aware World Modeling

Videos are inherently 2D projections of a dynamic 3D world. We organize existing geometry-aware world modeling approaches into three categories: explicit 3D modeling, implicit 3D supervision, and geometry foundation models.

Explicit 3D Modeling. One line of geometry-aware world models explicitly represents geometry through jointly generated RGB–XYZ or RGB–depth signals, point maps, point clouds, dynamic 3D scenes, or persistent 3D caches, allowing generative models to directly capture the spatial structure of a scene (Zhang et al., 2025a; Zhen et al., 2025; Jiang et al., 2025; Ren et al., 2025; Liu et al., 2025). Such structured representations support downstream applications including novel-view synthesis, camera control, 3D reconstruction, and robotic manipulation. Some approaches further combine video priors with 3D or 4D representations to improve the visual realism and cross-view consistency of dynamic scenes.

Implicit 3D Supervision. Rather than explicitly generating geometry, another line of work encourages video models to internalize 3D information through representation-level supervision. Geometry Forcing (Wu et al., 2025a) projects intermediate features of a video diffusion model into the feature space of a geometry foundation model. It introduces Angular Alignment and Scale Alignment to preserve the directional and magnitude-related information of geometric representations, respectively. More recent methods introduce denser transition-level supervision: GEM-4D distills 4D correspondence from a geometry foundation model, while PhysisForcing aligns point trajectories and semantic inter-region relations within interaction-focused regions (Zhou et al., 2026; Zhang et al., 2026). Such training-time alignment improves geometric or physical consistency with little additional inference overhead. Although our work shares the objective of internalizing 3D information, these methods supervise selected projections or relations of the video representation. Our method instead constructs a factorized transition space that explicitly preserves video-specific information alongside shared physical structure.

Geometry Foundation Models. Recent geometry foundation models provide a new means of obtaining geometric supervision from ordinary videos. Models such as VGGT (Wang et al., 2025) directly estimate camera parameters, depth, and dense point maps from multiple images, while encoding both local and global spatial relationships in their intermediate representations. PAGE-4D (Zhou et al., 2025) further models spatiotemporal geometry in dynamic scenes, providing richer geometric representations of camera motion and scene changes in videos. These models allow video world models to obtain teacher signals without requiring manually annotated 3D data for the target video dataset. We similarly employ a frozen geometry foundation model as a training-time teacher, but use its representations to incorporate geometric information into a unified physical space.

This interface carries forward quantities long studied in 3D reconstruction, including depth, camera pose, point correspondence, and surface structure (Schönberger & Frahm, 2016). The teacher uses this geometric knowledge to anchor transition learning; the deployed backbone predicts RGB without explicitly reconstructing every quantity estimated by the teacher.

B.3 Representation Alignment and Unified Latent Spaces

Guiding generative models with representations from frozen foundation models has shown promising results in both image and video generation. REPA (Yu et al., 2024) aligns intermediate diffusion features with those of a pretrained visual encoder, providing the generative model with more stable high-level visual semantics. VideoREPA (Zhang et al., 2025b) extends this idea to video generation, using spatiotemporal visual representations to improve the training and generation quality of video diffusion models. Representation autoencoders further show that semantically rich pretrained visual features can serve as high-capacity generative latents when paired with suitable decoders (Zheng et al., 2025). Dense patch features from DINOv2 (Oquab et al., 2023) encode object semantics and local spatial structure, whereas video encoders such as V-JEPA (Bardes et al., 2024) learn spatiotemporal predictive representations from consecutive frames. Together, these studies demonstrate that external visual representations can help generative models organize their internal features. Unlike these approaches, our method treats video and geometry as complementary observations of a shared world transition rather than aligning video features with a single target representation.

Shared and private subspaces have been studied as a way to retain common and source-specific information. Domain Separation Networks use this organization with reconstruction for domain adaptation (Bousmalis et al., 2016). In generative world-transition learning, the relevant distinction is predictive role: geometry anchors common structure, while a complementary pathway retains visual information used for future reconstruction. Our factorization assigns these roles through reconstruction objectives and permits information to interact in the coupled state.

Unified latent spaces also organize heterogeneous inputs into common token interfaces (Mizrahi et al., 2023; Chameleon Team, 2024). UniT (Chen et al., 2026) uses visual anchoring and cross-reconstruction to learn a physical language for human-to-humanoid action transfer. Its central problem is cross-embodiment alignment. Here, the central problem is organizing structural and complementary visual information for generative prediction and retaining their benefit after auxiliary modules are removed. DFoT participates in learning both transition pathways, and the acquired knowledge is subsequently consolidated into its original inference interface. These differences concern the predictive task, representation roles, and deployment requirement, rather than the use of a codebook alone.

These research directions address related questions at different levels. Latent-dynamics models such as Dreamer study prediction and planning within a learned state; JEPA-style work studies what visual information predictive representations retain (Assran et al., 2023; Bardes et al., 2024); multimodal latent spaces study interfaces across heterogeneous inputs. Our question is narrower: how learned visual and geometry-derived views of a video transition can share structure without discarding complementary information needed for generation, and how that training-time organization can benefit an auxiliary-free predictor.

C Full Implementation Details

Implementation Details. For camera-conditioned generation, we train DFoT (Song et al., 2025) on 16-frame, 256×256 clips using the first frame and per-frame camera poses. Table 6 relates the optimization phases to the conceptual stages in the main text. Acquisition uses 1,450 steps in total. The effective batch size is 40 except for the rollout-state adapter, which uses 8. Video-model learning rates for knowledge transfer and motion-aware refinement are 5×10^{-7} and 3×10^{-7}, respectively. PAGE-4D, DINOv2, and V-JEPA remain frozen. The final model retains the camera/action-conditioned DFoT interface, but not the learned physical conditioner or auxiliary representation modules.

Training Cost. The recorded optimization phases along the final checkpoint lineage take approximately 120 GPU-hours on A100/A800-class hardware.

Table 6: Optimization schedule for the RealEstate10K model. Conceptual stages organize several optimization phases; motion-aware refinement remains part of the evaluated final model.

Stage	Optimization phase	Steps	Effective batch
Acquire	Unified-space training	1,000	40
	Rollout-state adapter	200	8
	Selective joint refinement	250	40
Teach	Adapter warm-up	150	40
	Conditional teacher training	800	40
Consolidate	Student distillation	250	40
	Motion-aware rollout refinement	150	40

Checkpoint Roles. The complete representation checkpoint after acquisition is used for the factorized-space probes. The primary 100-video generation result uses the final backbone after the 150-step motion-aware refinement phase. These are distinct evaluation roles: probes examine the auxiliary space before consolidation, while generation tests what remains in the deployed backbone. Table 4 uses 25 matched clips along one checkpoint lineage to isolate successive training stages; its scores are not substituted for the 100-video main comparison.

Optimization Rationale. The geometry-first codebook schedule establishes a common physical reference before the video and fused branches learn its assignments. Conditioner training then makes the additional fused evidence useful for prediction, while selective distillation transfers beneficial corrections to the backbone. Warm-up and refinement phases implement this curriculum; the three conceptual stages describe their roles rather than three single optimization runs.

Codebook Usage and Optimization Diagnostic. The RVQ records assignment usage separately for its two 256-entry levels. For each logged distributed training batch, active ratio is the fraction of entries assigned at least once, and perplexity is the exponential entropy of the empirical assignment frequencies. Table 7 reports medians from the geometry branch in the acquisition runs and their subsequent selective-joint refinement along the checkpoint lineage. The acquisition trace has 100 logged batches; the first and last 10 show how assignments change during optimization. The refinement trace has 25 logged batches.

Table 7: Geometry-branch RVQ usage from training histories. Active is the per-batch active-code ratio, not the fraction ever used over the complete dataset; PPL is assignment perplexity. Values are medians over logged batches.

Phase	Logs	Active q_0 (%)	Active q_1 (%)	PPL q_0	PPL q_1
Acquisition, first 10	10	1.6	5.3	3.6	9.0
Acquisition, last 10	10	50.6	16.8	101.3	21.8
Selective joint refinement	25	48.4	16.0	101.8	18.3

Usage broadens during acquisition and remains near the final acquisition levels in joint refinement. The first residual level reaches roughly half of its available codes per batch, while the second reaches about one-sixth.

Data and Preprocessing. We use the RealEstate10K train/test partition and the 256×256 preprocessed RGB videos and aligned camera records distributed for DFoT (Song et al., 2025). Training samples contain 16 frames, with a base source-frame skip of 10; the training sampler can increase this skip by one per subepoch. Training-time horizontal flip, temporal reversal, and back-and-forth reordering have probabilities 0.5, 0.5, and 0.1, respectively, and are applied consistently to images and camera conditions. The 64-frame main test and the 256-frame stress test instead use source-frame skip 1, one conditioning frame, and no training augmentation. Clips shorter than the requested test horizon are filtered before selection, so the two test pools need not contain identical videos.

Key Configuration. The factorized representation uses 768-dimensional features on an 8×8 grid. Its physical RVQ has two stages of 256 entries, a 64-dimensional quantized space, commitment weight 0.25, and EMA decay 0.8. The frozen teachers are PAGE-4D for geometry, DINOv2 ViT-S/14 for image features, and V-JEPA ViT-L/16 for video features. Optimization uses mixed

precision, AdamW with $\beta = (0.9, 0.99)$ and weight decay 0.01, and gradient clipping at 1.0. In the 250-step student phase, the selective advantage-distillation and video-prior weights are 0.5 and 0.05. In the final 150-step refinement, these are 0.25 and 0.02; latent-motion and generated-context exposure weights are 0.01 and 0.05. The latter two terms are evaluated every 5 and 20 steps, respectively, with frequency compensation. These values describe the relevant active phases; inherited but inactive configuration fields are not treated as additional losses.

Additional Hyperparameters and Definitions. The conditioner receives $\widetilde{\mathbf{Q}}_f = \alpha \mathbf{Q}_f + \sigma\xi$, where $\alpha^2 = \mathrm{sigmoid}(\ell_j)$, $\sigma^2 = \mathrm{sigmoid}(-\ell_j)$, and ξ is Gaussian noise clipped to $[-4, 4]$. Its learned scalar read coefficient is $0.08\tanh(g)$, initialized with $\tanh(g) = 0.3$; a spatial pair-compatibility gate further modulates the update. Table 8 collects the configured transfer settings. The base and shuffled errors in Eq. 8 are detached comparison targets.

Table 8: Teacher and student transfer settings for the evaluated checkpoint lineage.

Component	Setting
Teacher error normalization	$\varepsilon = 10^{-4}$; $\lambda_b = \lambda_s = 1$; $\mu_b = \mu_s = 0.002$
Teacher auxiliary terms	Shuffled correction 0.25; pair-gate rank 0.25 (margin 0.1); pair-gate calibration 0.01
Physical-velocity head	$\lambda_p = 0$ in the evaluated teacher
Student selection	$a_{\max} = 1$; $\mathrm{sigmoid}(\ell_j) \in [0.1, 0.9]$; motion boost 1

The auxiliary term in Eq. 8 is

$$\mathcal{L}_{\text{aux}} = 0.25\mathcal{L}_{\text{id}} + 0.25\mathcal{L}_{\text{gate-rank}} + 0.01\mathcal{L}_{\text{gate-cal}}. \tag{17}$$

Let $d = \max(e_b, 10^{-4})$ and let δv_s be the shuffled-condition minus unconditioned future-frame FM velocity. With $E_s = \mathrm{sigmoid}(-1 - \ell_j)\,\mathrm{mean}[(\alpha\delta v_s)^2]$, the shuffled correction term averages $\sqrt{E_s/d + 10^{-6}} - 10^{-3}$ over valid pairs. For spatial compatibility gates r_c, r_s, the gate-ranking term is $[0.1 - \bar{r}_c + \bar{r}_s]_+$, where bars denote spatial means. Gate calibration averages $[-\log p_c - \log(1 - p_s)]/2$ over locations, with $p_k = \mathrm{clip}(0.8 r_k, 10^{-4}, 1 - 10^{-4})$. The physical-velocity head is instantiated but its loss is inactive in the evaluated teacher.

Baseline Checkpoints. DFoT and Geometry Forcing use released checkpoints (Song et al., 2025; Wu et al., 2025a). REPA and VideoREPA start from the same DFoT initialization and train for 500 steps (effective batch size 40, learning rate 8×10^{-6}, alignment weight 0.5); the Minecraft extension starts from released Action-DFoT weights.

D Complete Student Objective

Selective Transfer. In Eq. 10, $\varepsilon = 10^{-4}$ and $a_{\max} = 1$. The indicator $\mathbf{1}_\Omega$ retains valid transition pairs with $\mathrm{sigmoid}(\ell_j) \in [0.1, 0.9]$. Motion salience is the maximum of two bounded ratios: the current-to-future displacement divided by that displacement plus the mean adjacent-frame displacement, computed on pooled 8×8 video latents and on camera centers. With salience $s \in [0, 1]$, $w_{\text{motion}} = (1 + s)/\mathbb{E}_\Omega[1 + s]$, where the batch mean is detached. Thus motion weighting redistributes, rather than increases, the distillation weight.

Overall Training and Inference. The student is trained using

$$\mathcal{L}_{\text{student}} = \mathcal{L}_{\text{FM}} + \lambda_{\text{KT}}\mathcal{L}_{\text{KT}} + \lambda_{\text{prior}}\mathcal{L}_{\text{prior}} + \lambda_{\text{roll}}\mathcal{L}_{\text{roll}}, \tag{18}$$

where $\mathcal{L}_{\text{prior}}$ preserves the prediction of the frozen video backbone and $\mathcal{L}_{\text{roll}}$ combines generated-history exposure with multi-span latent-motion consistency. The complete training procedure therefore progresses from unified-space learning to conditional transfer and motion-aware rollout refinement. At inference time, the geometry and visual foundation models, RVQ, fusion modules, decoders, and conditioner are removed. Sampling follows the original autoregressive DFoT procedure in Sec. A.1 using only the consolidated video backbone, introducing no additional inference-time modules.

E EXPANDED EVALUATION PROTOCOL

Evaluation Metrics. We use FVD (Unterthiner et al., 2018), FID (Heusel et al., 2017), LPIPS (Zhang et al., 2018), SSIM (Wang et al., 2004), and PSNR to evaluate video distribution quality and visual fidelity. We use JEDi (Luo et al., 2024) to measure spatiotemporal predictive consistency in long videos. To further evaluate geometric consistency, we report Reprojection Error (RPE) (Wu et al., 2025a), where a lower value indicates better multi-view consistency. Within each protocol, methods use the same test indices, first-frame conditions, camera trajectories, sampling seeds, and autoregressive context. The 64-frame rollout experiment contains 100 videos, with RPE on its first 12; the 128- and 256-frame experiments use a separate 12-video stress cohort. FID, LPIPS, SSIM, PSNR, and JEDi omit the conditioning frame.

Evaluation Scope. The 64-frame test is the primary long-horizon comparison, extending the 16-frame training length by a factor of four. RPE is computed on its first 12 videos. The separate 12-video 256-frame stress cohort supplies the first 128 frames for the 128-frame scores. For each eligible RealEstate10K test video, the loader samples one valid start with random seed 0 and then shuffles the selected starts; evaluation uses indices 0–99 or 0–11 after horizon-specific length filtering. Generation uses seed $20260814 + o$ for shard offset o, shared across methods, 50 sampling steps, a four-frame sliding history, guidance scale 4.0, and stabilization level 0.02. ARKitScenes clips are selected before generation using synchronized sensor validity and a fixed camera-motion envelope; generated outputs are not used by the selection procedure. Its depth-derived metrics use one scale per clip, while sensor-anchored pose recall counts failed localizations as misses. These protocols are kept distinct from the small-scale Minecraft action-conditioning extension.

Primary-Table Uncertainty. We resample the 100 matched 64-frame clips with replacement 1,000 times and recompute each metric for both methods on the same sampled indices. Observed differences are computed from unrounded metric values. For UPCAST minus Geometry Forcing, the observed PSNR difference is $+1.50$ dB, with a 95% bootstrap percentile interval of $[+1.13, +1.89]$ dB; the FVD difference is -37.0, with interval $[-108.3, +11.4]$; and the JEDi difference is -0.53, with interval $[-1.42, +0.41]$. Thus the FVD and JEDi intervals include zero despite their favorable point estimates. These intervals quantify sensitivity to the fixed test-clip sample, not variability across training seeds.

Sensor-Referenced Geometry Protocol. The main ARKitScenes set uses 12 preselected Validation videos from distinct visits, with one 64-frame clip per video at 256×256 resolution and native-frame stride 1. RGB, depth, and confidence maps share timestamps; camera intrinsics must lie within 2 ms, and sensor poses are interpolated at the image timestamps. Candidate starts are checked every six synchronized frames; consecutive selected timestamps must differ by at most $2.1/60$ s. The fixed motion envelope, derived from 854 RealEstate10K 64-frame trajectories, requires a translation-path sum of 0.158503–2.112896 in the supplied pose coordinates and a summed rotation of $0.544140°$–$34.238981°$, with no translation step above 0.087918 pose units or rotation step above $2.481408°$. Translation thresholds are numerical pose-coordinate filters, not a claim of metric-scale equivalence between datasets. Among eligible windows, we take the lowest motion-score candidate with at least 55% valid depth pixels, allowing at most 30 candidates per video. Validity requires sensor confidence at least 1 and depth between 0.05 and 20 m. Selection uses sensor data only, before generation. The frozen metric-indoor Depth Anything V2 Small estimator is applied equally to real and generated RGB, with one median predicted-to-sensor scale per complete clip. For point-cloud metrics, every fourth target frame and every fourth pixel are back-projected with sensor intrinsics and poses, with at most 100,000 points per cloud. Chamfer-L_1 averages the two directed nearest-neighbor distances; F@10 cm is the harmonic mean of the fractions within 10 cm in both directions. Sensor-anchored pose evaluation uses SIFT correspondences, sensor depth for the real-view 3D points, and PnP-RANSAC with a 3-pixel reprojection threshold; failed PnP estimates remain in the denominator of Pose R@10.

V-JEPA Feature F-Score. The Ours-only diagnostic uses ℓ_2-normalized V-JEPA features and $k = 3$ nearest-neighbor manifolds (Kynkäänniemi et al., 2019). Precision and recall quantify coverage of the generated and reference feature sets; their harmonic mean is the reported F-Score. Because V-JEPA also supplies training-time regularization, this diagnostic characterizes agreement

in a training-related representation space, rather than independent geometric accuracy or pixel correctness. It is distinct from both JEDi and the metric-depth point-cloud F-scores.

Representation and Action Protocols. The factorized-space probe uses 64 RealEstate10K clips and three fixed temporal spans, $(0, 3)$, $(0, 7)$, and $(0, 15)$, giving 192 paired transitions. It evaluates the frozen complete selective-joint checkpoint with a fixed noise level of 0.1. Cross-modal retrieval uses other transitions as negatives; same-scene span controls restrict candidates to the three spans of one clip, with chance level $1/3$. Decoder interventions change physical or appearance inputs while leaving the decoder frozen. For the Minecraft quickstart extension, the official 10 training and 10 validation trajectories are kept disjoint; the Action-DFoT adaptation uses 500 steps, including 100 bootstrap steps, an effective batch size of 40, and the same 16-frame, one-context-frame interface for paired evaluation.

F Failure Case Analysis

Additional ARKitScenes Failure Examples. Table 9 reports only per-clip metrics on two additional zero-shot Validation clips outside the main 12. These clips were chosen after evaluation because UPCAST underperforms Geometry Forcing on measured depth, point-cloud, and pose criteria. They illustrate limitations, not their frequency in the validation split; no aggregate from the additional clips is used in the main comparison.

Table 9: Metric-only failure examples from additional ARKitScenes clips. Both methods use the main zero-shot protocol. Values are per clip, not estimates of dataset-wide performance.

Additional clip	Method	AbsRel ↓	F@10 cm ↑	Pose R@10 ↑
37	Geometry Forcing	0.129	0.717	0.302
37	UPCAST	0.243	0.316	0.238
72	Geometry Forcing	0.056	0.917	0.984
72	UPCAST	0.080	0.794	0.222

RealEstate10K Failure Example. Figure 4 shows a RealEstate10K failure under a moving camera. The conditioning frame shows the open closet but not the bathroom interior to its left. By frame 21, the reference reveals a white toilet in that room, whereas UPCAST generates an empty, brightly lit wall and floor; the mismatch remains at frame 41. In the initially visible closet, dark hanging garments and the recessed interior give way to a lighter, flatter arrangement. The case thus combines an incorrect completion of unseen content with a change to previously observed appearance. The unseen bathroom is underdetermined by the single input view, while recursive conditioning on generated frames may allow errors in the closet to persist.



Figure 4: An illustrative failure from the 64-frame RealEstate10K main test (index 28). The input shows the closet but not the adjoining bathroom. Paired rows show zero-based frames 21 and 41: the reference reveals a toilet on the left, while the generated room remains empty and the closet changes. All images are unmodified frames from the same lossless evaluation record.

G Discussion and Limitations

Training-Compute Comparability. The training budgets are not strictly matched: DFoT and Geometry Forcing use released checkpoints, whereas UPCAST follows the multi-stage schedule in Table 6. The comparisons measure generation quality under the stated protocols, not performance per unit of additional training compute. A matched-compute study would separate these effects more fully.

Scope of the World-Transition View. Our world-transition view concerns information useful for predicting future observations. The shared and private states are defined by their training objectives and predictive roles; the probes measure those roles, while the stage-wise comparison follows one checkpoint lineage. Training uses RGB video and frozen features computed from it, with no independent 3D sensor measurements.

Teacher and Evaluation Dependence. Acquisition relies on frozen PAGE-4D, DINOv2, and V-JEPA representations, which can carry domain-specific errors or omit information needed for a new scene. At evaluation, the ARKitScenes depth metrics depend on a frozen RGB depth estimator and sensor-based alignment, while pose recall also depends on feature matching and PnP success. These are useful but indirect measurements of generated-view geometry. The 12 motion-filtered clips cover a controlled zero-shot setting; additional examples in Section F show that geometric outcomes can reverse on other scenes and motions.

Horizon, Scale, and Future Inputs. The main model trains on 16-frame clips, and its 128/256-frame evaluations show that substantial appearance degradation remains under recursive generation. Camera poses are supplied throughout rollout, and the Minecraft action extension uses a small quick-start split; broader interactive dynamics and larger training scales remain untested. The formulation could also assign roles to other information sources: a language description may be treated as an observation, whereas an instruction is more naturally a goal or control. Testing these roles, or extending the model to proprioception or touch, would require suitable encoders, conditioning, and dedicated evaluations.