GAD: GEOMETRY ATTENTION DISTILLATION FOR CONSISTENT VIDEO GENERATION

Anonymous authors
Paper under double-blind review

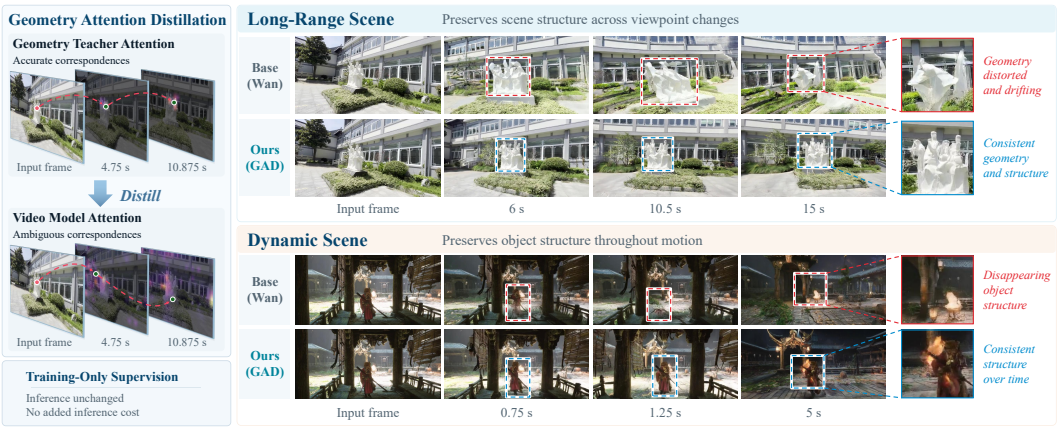


Figure 1: **Geometry attention distillation for consistent video generation.** Correspondences from a geometry teacher guide video attention during training (left), improving scene geometry under viewpoint changes and object structure during motion (right), without added inference cost.

ABSTRACT

Video diffusion models generate visually compelling short-term sequences but often struggle to preserve scene geometry and object structure over time. In this paper, we first demonstrate that current video attention cannot identify reliable correspondences across frames, with responses dispersed across multiple regions even after recognizable scene structure has emerged. We argue that reliable cross-frame associations are essential for information exchange as scene structure forms in video generation. Thus, we introduce Geometry Attention Distillation (**GAD**), which transfers correspondences from a frozen geometry model's attention to video attention. Extracted from clean videos, these correspondences link the same static surface or moving object across frames. Since structure formation begins in deeper layers at higher noise and shallower layers at lower noise, we apply attention distillation at a student layer selected according to the noise timestep. After training, video attention focuses more consistently on corresponding locations. Experiments across two video backbones demonstrate improved long-range scene consistency and dynamic temporal consistency, with geometric gains across generation lengths. On Wan2.2 5B, **GAD** improves cross-frame matching accuracy by at least 30% over the Wan baseline on both long-range benchmarks. Notably, **GAD** achieves these gains without additional inference modules, offering a promising approach to efficient and consistent video-based world simulation.

1 INTRODUCTION

Recent video diffusion models (Blattmann et al., 2023; Kong et al., 2024; Yang et al., 2025b; Wan et al., 2025) generate visually compelling sequences from text and images, yet struggle to preserve scene geometry over long sequences and maintain coherent object structure during motion. Such consistency is essential for interactive world modeling (Yang et al., 2023; Parker-Holder et al., 2025), robotic planning (Du et al., 2023; Ko et al., 2024), and autonomous driving simulation (Hu et al., 2023; Wang et al., 2024b), where generated videos must represent a coherent environment over time.

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.

To improve consistency, existing approaches employ camera-aware positional encoding (Li et al., 2025a; Yin et al., 2026), explicit geometric conditioning (Ren et al., 2025; Cao et al., 2025; Huang et al., 2025a), history-conditioned generation (Li et al., 2025b; Zhang et al., 2025a), and geometry-aware feature alignment (Wu et al., 2026; Huang et al., 2026a). However, such guidance does not necessarily ensure that video models reliably associate observations of the same scene element across frames, a capability essential for consistent video generation.

To understand how cross-frame associations develop, we examine attention patterns in video models and the emergence of scene structure across noise levels and network depths. Our analysis shows that video attention cannot reliably identify the same scene element across frames: its responses are often dispersed across multiple regions, even after recognizable scene structure has emerged. Such ambiguous associations can mix information from unrelated regions, undermining the preservation of scene geometry and object structure across frames. Ideally, cross-frame attention should connect observations of the same scene element, enabling information exchange that supports consistent structure formation. We further observe that recognizable scene structure begins to emerge in deeper layers at higher noise and shallower layers at lower noise. This suggests guiding video attention where structure and cross-frame associations are beginning to form, with the guidance depth adapted to the noise level. For a source of this guidance, we turn to geometry foundation models (Leroy et al., 2024; Wang et al., 2025a) with promising multi-view matching capabilities. Probing VGGT-Ω (Wang et al., 2026a) reveals attention responses that associate the same static surface or moving object across frames, providing geometric correspondence targets for guiding video attention.

Building on these insights, we introduce Geometry Attention Distillation (**GAD**), an attention distillation framework that guides cross-frame associations during structure formation. During training, we extract correspondence targets from the attention of a frozen geometry model on clean videos. The total attention each source token assigns to a target frame serves as a reliability cue for weighting its supervision. Furthermore, we select the student layer for attention distillation according to the noise timestep, guiding cross-frame associations as scene structure forms. At the selected layer, pairwise supervision encourages spatial discrimination within frame pairs, while trackwise supervision extends temporal coverage by matching a common source to multiple target frames. These complementary distillation objectives are optimized jointly with the original video generation loss, transferring cross-frame relations while preserving the flexibility of video representations. After training, video attention focuses more consistently on corresponding regions across frames, while generated videos better preserve scene geometry and moving-object structure.

We evaluate **GAD** on camera-controlled video generation using RealEstate10K and DL3DV-10K for long-range scene consistency and OmniWorld-Game for dynamic consistency. Across two video backbones, **GAD** improves geometric consistency in long-range generation and temporal consistency in dynamic generation. Geometric improvements across generation lengths further support the applicability of **GAD** across backbone configurations and video durations. Notably, with Wan2.2 5B, **GAD** delivers at least a 30% relative improvement in cross-frame matching accuracy on both long-range benchmarks, without additional inference modules or computations.

Our main contributions are summarized as follows:

- We analyze cross-frame associations in video models, revealing persistent geometric ambiguities in attention and a noise-dependent shift in the depth at which scene structure begins to form. These observations motivate geometry attention distillation at layers where structure is emerging.
- We propose **GAD**, an attention distillation framework that transfers geometric correspondence knowledge to video attention for improved cross-frame consistency.
- Extensive experiments demonstrate that **GAD** improves long-range scene consistency and dynamic temporal consistency across two video backbones, with consistent improvements across generation lengths and no additional inference modules.

2 Related Work

2.1 Consistent Video Generation

Video diffusion models (Ho et al., 2022; Blattmann et al., 2023; Kong et al., 2024; Yang et al., 2025b; Wan et al., 2025) generate visually compelling videos but still struggle to maintain consis-

tent scene geometry and object structure across frames. Existing efforts (Yang et al., 2025a; Huang et al., 2025b; Zhu et al., 2026) address these limitations through temporal coordination, historical context, and spatial guidance. Sampling strategies improve temporal continuity (Qiu et al., 2024; Lu et al., 2024; Yang et al., 2026a; Chen et al., 2026a) while attention design regulates information exchange (Zhou et al., 2024; Tang et al., 2026). Long-range generation further benefits from compressed or hierarchical context (Zhang et al., 2025a; Zhu et al., 2025; Zhao et al., 2026; Chandratreya et al., 2026), selective memory access (Wu et al., 2025; Ji et al., 2025; Ding et al., 2026; Yang et al., 2026b), and training that addresses error accumulation and long-context dependencies (Yin et al., 2025; Chang et al., 2026; Chen et al., 2026b). Camera-aware positional encodings complement this temporal context with spatial guidance (Li et al., 2025a; Yin et al., 2026; Jang et al., 2026). These approaches improve information retention and exchange across frames, but do not necessarily ensure that attention connects observations of the same scene element. Complementing these efforts, **GAD** uses geometric correspondences to directly supervise video attention, guiding associations between observations of the same scene element across viewpoints and through motion.

2.2 Geometry Foundation Models for Video Consistency

Geometry foundation models (Wang et al., 2024a; 2025b;a) make learned structural priors accessible to video generation. Together with explicit representations, these priors provide spatial references for view-conditioned synthesis (Li et al., 2025b; Yu et al., 2025; 2024b; Liu et al., 2026) and persistent scene memory (Huang et al., 2025a; Wang et al., 2026c; Lv et al., 2026). Geometry also guides learning through feature alignment and positional encoding (Wu et al., 2026; Jin et al., 2026), joint appearance–geometry modeling (Lu et al., 2025; Dai et al., 2026; Chen et al., 2025), and depth or motion constraints (Bai et al., 2025; Fang et al., 2025). Preference optimization and reinforcement learning further use geometric feedback to improve generated content (Du et al., 2026; An et al., 2026; Wang et al., 2026b; Ackermann et al., 2026). However, methods using explicit geometric conditions (Ren et al., 2025; Cao et al., 2025) can inherit reconstruction and reprojection errors and face additional challenges in dynamic scenes. While implicit geometric learning (Huang et al., 2026a; Hu et al., 2026) offers greater flexibility for dynamic scenes, direct feature alignment may still impose representational constraints beyond those needed for geometric consistency. To this end, **GAD** draws correspondence targets from geometry model attention to supervise cross-frame associations in video attention, transferring relations between observations of the same scene element without matching feature representations or requiring reconstructed geometry at inference.

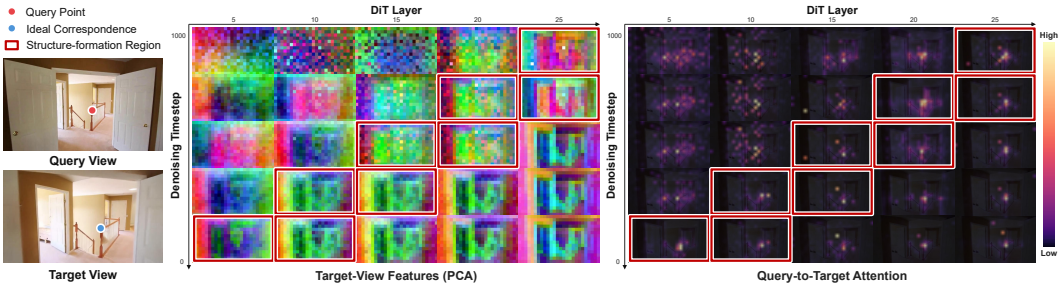
2.3 Attention Distillation and Correspondence Supervision

Knowledge distillation can transfer attention patterns and relational structure in addition to individual features (Zagoruyko & Komodakis, 2016; Tung & Mori, 2019; Qiu et al., 2023), with cross-stage connections shaping how teacher knowledge reaches the student (Chen et al., 2021). In diffusion models, feature alignment (Yu et al., 2024a; Hwang et al., 2025; Bhowmik et al., 2026) and token-relation distillation (Zhang et al., 2025b; Lian et al., 2026) provide complementary supervision, refined through token selection or noise-dependent prior granularity (Pang et al., 2026; Min et al., 2026). Diffusion representations also support correspondence estimation (Tang et al., 2023), consistent video editing (Geyer et al., 2024; Qi et al., 2023), and sampling guidance (Zu et al., 2026). More directly, CAMEO (Kwon et al., 2026) improves novel-view synthesis by supervising a fixed attention layer with correspondences derived from static 3D pointmaps. For consistent video generation, **GAD** instead extracts reliability-weighted correspondences from geometry model attention across static surfaces and moving object parts, and routes attention distillation to student layers according to the noise timestep to guide cross-frame associations during structure formation.

3 Analysis and Motivation

3.1 Attention Patterns in Video Models

Cross-frame consistency requires associating the same scene element across viewpoint changes or object motion. We examine how video attention establishes these associations across noise levels and network depths. In Fig. 2a, we fix a source query and a target frame within a single video generation trajectory, and visualize the target frame's intermediate token features using PCA alongside



(a) Features and attention from video generative models across noise levels and layers.



(b) Attention from VGGT-Ω on static surfaces and moving objects.

Figure 2: **Correspondence patterns in video generation and geometry foundation models.** (a) In the *structure-formation region* (red boxes), recognizable scene structure and cross-frame associations begin to emerge, at deeper layers under higher noise and shallower layers under lower noise. However, attention remains ambiguous even after structure becomes clear. (b) Selected VGGT-Ω heads attend to the same static surface or moving object across frames.

query-to-target attention responses. The PCA visualizations reveal recognizable scene structure, while the attention maps show whether responses concentrate near the reference correspondence or favor other regions. Further visualization details are provided in Appendix E.1.

This joint analysis reveals two key findings. First, video attention cannot reliably identify the same scene element across frames. Its responses can remain dispersed across multiple regions, including those with similar appearance or semantics, even after recognizable scene structure has emerged in the PCA features. Second, the emergence of scene structure and cross-frame associations depends on both noise level and network depth. We distinguish three qualitative regimes. In the *noise-dominated region*, PCA features show little recognizable scene structure, while attention responses remain scattered rather than concentrated around the corresponding region. In the *structure-formation region* (red boxes in Fig. 2a), recognizable scene structure begins to emerge, and attention starts to respond to corresponding regions across frames, although competing responses remain. This region occurs in deeper layers at higher noise and shallower layers at lower noise. In the *structure-established region*, scene layout is more clearly defined in the PCA features and cross-frame associations are more apparent, but they are not necessarily geometrically accurate.

These observations motivate explicitly guiding video attention toward geometrically corresponding regions to help preserve scene and object structure across frames. Furthermore, the three observed regimes suggest that the placement of this guidance matters. In the noise-dominated region, weak structural evidence may make correspondence supervision difficult to exploit; in the structure-established region, additional guidance may have less opportunity to shape associations that have already formed. The structure-formation region offers an intermediate stage where correspondence learning has a structural basis while the associations are still developing. We therefore hypothesize that geometric guidance is most effective in this region. Since this region emerges at different depths depending on noise, we adapt the supervision layer to the noise timestep.

3.2 Geometric Correspondence in Geometry Foundation Models

Recent advances in geometry foundation models have demonstrated strong multi-view matching capabilities, motivating their use as a source of correspondence supervision. Probing VGGT-Ω (Wang et al., 2026a), we find that attention from selected layers and heads captures correspondences on both static surfaces and moving objects. As shown in Fig. 2b, attention responses remain concentrated around the same railing location across viewpoint changes in the static example, and follow the per-

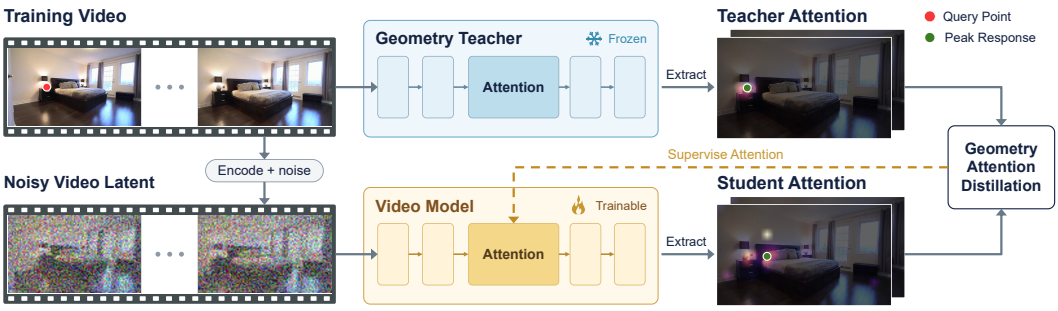


Figure 3: **Overview of GAD.** Correspondences extracted by a frozen geometry teacher from clean videos guide the video model's attention during denoising. The correspondence supervision complements the generation objective during training, without changing the inference procedure.

son's hat as the person moves in the dynamic example. These observations suggest that geometry model attention can guide associations with the same scene element across both camera and object motion. Visualization details and layer–head selection are described in Appendices B.1 and E.2.

However, VGGT-Ω can still produce an attention peak when the scene element represented by a source token is occluded or absent from the target view, making the peak alone an unreliable correspondence cue. As illustrated in Fig. 6, ambiguous or incorrect responses can exhibit visible peaks while assigning little total attention to the target frame's spatial tokens. We call this total the target-frame attention mass and use it as a reliability cue to downweight potentially unreliable matches.

4 Geometry Attention Distillation

Building on the analysis in Section 3, we distill geometric correspondences from VGGT-Ω attention into video model attention to improve cross-frame consistency, as illustrated in Fig. 3.

4.1 Confidence-Aware Teacher Correspondences

Given a clean training video, we extract attention from the frozen VGGT-Ω teacher and align its responses with the student's spatial and temporal token grid. Teacher attention is normalized jointly across all input frames, allowing the total attention assigned to each target frame to serve as a reliability cue, as discussed in Section 3.2. Let $a = (i, u)$ denote source token u in frame i, and let $\bar{A}_{a,j}(v)$ denote its aligned attention weight at spatial token v in target frame $j \neq i$, retaining this original frame mass. We extract the candidate correspondence and target-frame attention mass as

$$c_{a,j} = \arg\max_{v} \bar{A}_{a,j}(v), \qquad m_{a,j} = \sum_{v} \bar{A}_{a,j}(v). \tag{1}$$

We convert the attention mass into a reliability weight $w_{a,j} = f(m_{a,j})$, where f is a nondecreasing mapping to $[0, 1]$ and invalid relations receive zero weight.

To tolerate small localization errors, we construct a soft target $y_{a,j}(v)$ by Gaussian smoothing around $c_{a,j}$. It is normalized over the valid spatial neighborhood $\mathcal{N}(c_{a,j})$ and zero elsewhere in the target frame. Thus, $y_{a,j}$ specifies where the student should match, while $w_{a,j}$ controls its relative contribution to the supervision. The target preserves the teacher-derived correspondence rather than reproducing its full attention distribution. Teacher attention alignment, confidence weighting, and target smoothing are detailed in Appendix B.1.

4.2 Student Attention Distillation with Noise-Timestep Routing

Noise-timestep routing. To guide cross-frame associations during structure formation, we adapt the supervision depth to the noise level, following the analysis in Section 3.1. Let $t \in [0, 1000]$ denote the model noise timestep, with larger t indicating higher noise. We select one student layer per training sample using interval boundaries $0 = b_0 < \cdots < b_K = 1000$:

$$\ell(t) = l_k \quad \text{for} \quad b_{k-1} \leq t < b_k, \qquad k = 1, \ldots, K, \tag{2}$$

where K is the number of noise intervals, l_k is the layer assigned to interval k, and the final interval includes $t = 1000$. Specifically, we assign shallower layers to lower noise and deeper layers to higher noise to follow the noise-dependent onset of structure formation.

Correspondence supervision. At the selected layer $\ell(t)$, we aggregate query–key scores across attention heads to form a student matching distribution. Inspired by CAMEO (Kwon et al., 2026), a lightweight head-aggregation MLP weights and combines these scores, followed by a softmax over candidate tokens $\mathcal{K}$ in target frame j. For source token a, we denote the resulting distribution by $P_{a,j}(v \mid \mathcal{K})$, omitting the layer index for brevity. This aggregation is used only during training.

For each source–target-frame relation (a, j), the candidate set contains the full positive neighborhood, $\mathcal{N}(c_{a,j}) \subseteq \mathcal{K}$. We compare the student distribution with the teacher-derived target through a per-correspondence cross-entropy loss:

$$\mathcal{L}_{\text{corr}}(a, j; \mathcal{K}) = - \sum_{v \in \mathcal{N}(c_{a,j})} y_{a,j}(v) \log P_{a,j}(v \mid \mathcal{K}). \tag{3}$$

When aggregating these per-correspondence losses, we use reliability weight $w_{a,j}$ to control the relative contribution of each correspondence.

To select the correspondences used for supervision, we employ two complementary sampling schemes that balance spatial discrimination and temporal coverage. *Pairwise supervision* samples source tokens from frame pairs and compares each with all spatial tokens in its target frame, providing a full set of spatial alternatives for matching. *Trackwise supervision* uses correspondences from a common source token to multiple target frames, using the positive neighborhood and sampled negative tokens in each frame to extend temporal coverage at lower cost than evaluating all spatial candidates. The corresponding objectives, $\mathcal{L}_{\text{pair}}$ and $\mathcal{L}_{\text{track}}$, aggregate the per-correspondence losses $\mathcal{L}_{\text{corr}}$ using their reliability weights. Attention aggregation and routing configurations are detailed in Appendix B.2, while sampling procedures and loss reductions are specified in Appendix A.

4.3 Overall Training Procedure

We retain the flow-matching objective (Lipman et al., 2022) and add correspondence losses:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{FM}} + \lambda_1 \mathcal{L}_{\text{pair}} + \lambda_2 \mathcal{L}_{\text{track}}, \tag{4}$$

where λ_1 and λ_2 control the pairwise and trackwise supervision strengths, respectively. Both correspondence losses are evaluated at the routed layer $\ell(t)$, with averaging over training samples and noise levels implicit. The flow-matching objective continues to train the backbone's denoising velocity prediction, while the correspondence losses guide cross-frame associations.

In practice, we first cache teacher correspondences and confidence information, preserving alignment with the student's temporal sampling and spatial transformations. Each iteration constructs the noisy video latent z_t at noise timestep t, selects layer $\ell(t)$, and computes the pairwise and trackwise losses from its query–key responses. We update the video transformer and aggregation MLPs while keeping the geometry teacher and latent encoder frozen. At inference, only the trained video model is used, with no additional inference modules or correspondence caches; its attention computation and sampling procedure remain unchanged. The flow-matching formulation and training and inference configurations are detailed in Appendix B.3.

5 Experiments

We evaluate **GAD** on camera-controlled video generation, assessing long-range scene consistency and dynamic temporal consistency alongside generation quality and inference efficiency. Implementation and evaluation details are provided in Appendices B and C, while limitations are discussed in Appendix F. Video comparisons are included in a separately submitted supplementary archive.

5.1 Setting

Training. Our training corpus comprises 35K clips from RealEstate10K (Zhou et al., 2018), DL3DV-10K (Ling et al., 2024), and OmniWorld-Game (*i.e.*, OmniWorld's game subset (Zhou et al.,

Table 1: **Long-range camera-controlled video generation on RealEstate10K and DL3DV-10K. GAD** improves all seven metrics over the Wan baselines on both backbones. Top-3 results are highlighted as the first, second and third.

Method	RealEstate10K							DL3DV-10K						
	FID↓	FVD↓	CLIP↑	R_{dist} ↓	T_{dist} ↓	M_{rate} ↑	M_{dist} ↓	FID↓	FVD↓	CLIP↑	R_{dist} ↓	T_{dist} ↓	M_{rate} ↑	M_{dist} ↓
GF	61.46	464.8	30.03	6.229	0.134	0.188	55.66	104.1	813.7	29.42	8.145	0.126	0.209	54.50
Gen3R	33.35	305.3	30.88	5.167	0.103	0.321	36.52	45.23	373.7	32.66	6.457	0.088	0.310	39.27
Voyager	58.38	410.3	30.47	4.779	0.099	0.334	33.01	65.77	608.4	32.16	4.740	0.073	0.313	41.02
Gen3C	31.07	264.7	31.11	4.288	0.076	0.315	35.07	38.63	325.6	33.46	4.732	0.056	0.289	42.19
Uni3C	34.46	289.6	31.32	5.146	0.096	0.384	32.11	43.28	351.1	33.30	5.589	0.078	0.331	38.42
Wan2.1 1.3B	30.83	267.4	31.11	7.236	0.162	0.336	38.19	40.06	336.9	33.93	6.395	0.112	0.318	42.41
+ **GAD**	30.53	259.0	31.69	4.734	0.093	0.387	32.16	38.83	328.8	34.13	5.029	0.071	0.358	37.05
Wan2.2 5B	31.32	258.8	31.77	6.705	0.175	0.263	45.53	38.60	322.1	33.85	9.573	0.108	0.257	48.90
+ **GAD**	29.20	242.1	32.03	4.335	0.088	0.350	31.36	37.93	318.0	34.24	4.890	0.065	0.335	38.85

Table 2: **Dynamic camera-controlled video generation on OmniWorld-Game. GAD** improves subject and background consistency and motion smoothness, without added inference cost.

Method	Rendering			VBench Metrics					Inference Cost		
	FID↓	FVD↓	CLIP↑	Subj. Cons.↑	Bg. Cons.↑	Motion Sm.↑	Aesth. Q.↑	Image Q.↑	Time (s)↓	Peak Mem. (GB)↓	Avg. Mem. (GB)↓
GF	162.0	1543	20.48	0.698	0.851	0.961	0.397	0.380	194.6	48.96	33.80
Gen3R	81.91	870.1	26.84	0.817	0.907	0.957	0.482	0.505	209.3	50.66	37.05
Voyager	103.6	987.7	26.09	0.805	0.906	0.963	0.492	0.521	2583.2	67.45	62.27
Gen3C	70.41	774.8	27.55	0.810	0.908	0.970	0.493	0.527	825.9	63.71	57.20
Uni3C	80.63	856.6	27.56	0.804	0.908	0.968	0.504	0.553	444.0	49.77	37.68
GT	-	-	-	0.867	0.919	0.981	0.553	0.662	-	-	-
Wan2.1 1.3B	65.16	665.9	27.39	0.813	0.889	0.968	0.533	0.591	172.9	25.44	22.23
+ **GAD**	60.35	625.7	27.67	0.854	0.910	0.977	0.534	0.586	171.2	25.44	22.76
Wan2.2 5B	61.29	657.6	27.46	0.825	0.900	0.964	0.542	0.636	69.60	34.85	26.11
+ **GAD**	59.52	622.3	28.23	0.853	0.913	0.976	0.546	0.641	70.44	34.85	26.08

2025)), covering real-world scenes and game environments. We train at a resolution of 832×480 for 10 epochs on 32 NVIDIA A800 GPUs. We fine-tune the video DiT and aggregation MLPs with AdamW at a learning rate of 10^{-5} while keeping the text encoder and VAE frozen.

Datasets. All three benchmarks evaluate camera-controlled video generation conditioned on a first frame, a text prompt, and a target camera trajectory. We assess long-range scene consistency with 15-second videos on RealEstate10K and DL3DV-10K, and dynamic temporal consistency with 5-second videos on OmniWorld-Game. For all three benchmarks, we use scene-disjoint training and test splits to prevent data leakage between training and evaluation.

Baselines. We compare with Geometry Forcing (GF) (Wu et al., 2026) and Gen3R (Huang et al., 2026a), which incorporate geometry through learned representations, and Voyager (Huang et al., 2025a), Gen3C (Ren et al., 2025), and Uni3C (Cao et al., 2025), which use explicit geometric conditioning. We apply **GAD** to Wan2.1 1.3B and Wan2.2 5B (Wan et al., 2025). For each backbone, the Wan baseline is fine-tuned on the same training data using only the flow-matching objective, under training settings matched to **GAD**. Throughout our experimental comparisons, Wan denotes this fine-tuned baseline, and + **GAD** denotes training with the additional distillation objectives. Baseline adaptation procedures are detailed in Appendix C.2.

Metrics. We report FID, FVD, and CLIP for generation quality and text alignment. To evaluate geometric consistency on RealEstate10K and DL3DV-10K, we measure camera accuracy using rotation distance R_{dist} (degrees) and normalized translation error T_{dist}, following Li et al. (2025b), and cross-view consistency using matching hit rate M_{rate} and mean target-position error M_{dist} (pixels), following Lu et al. (2025). For dynamic scenes on OmniWorld-Game, we report five VBench metrics (Huang et al., 2024): subject consistency, background consistency, motion smoothness, aesthetic

Table 3: **Generation duration.**

Duration	$R_{\text{dist}} \downarrow$	$T_{\text{dist}} \downarrow$	$M_{\text{rate}} \uparrow$	$M_{\text{dist}} \downarrow$
5s	3.020	0.064	0.529	21.84
5s + **GAD**	**1.684**	**0.053**	**0.571**	**19.31**
10s	5.948	0.072	0.371	38.34
10s + **GAD**	**3.376**	**0.056**	**0.453**	**33.33**
15s	9.573	0.108	0.257	48.90
15s + **GAD**	**4.890**	**0.065**	**0.335**	**38.85**

Table 4: **Noise-timestep routing ablation.**

Exp. ID	Loss Assignment	$R_{\text{dist}} \downarrow$	$T_{\text{dist}} \downarrow$	$M_{\text{rate}} \uparrow$	$M_{\text{dist}} \downarrow$
R1 (Wan)	None	9.573	0.108	0.257	48.90
R2	L5 at All noise	7.440	0.086	0.295	43.89
R3	L15 at All noise	7.346	0.090	0.286	42.71
R4	L25 at All noise	6.540	0.075	0.289	43.55
R5	L5 only at High noise	8.932	0.098	0.262	46.36
R6	L25 only at Low noise	9.290	0.106	0.258	47.17
R7	L25 / L5 at Low / High noise	8.755	0.104	0.268	45.77
R8 (Ours)	L5 / L15 / L25 at Low / Medium / High noise	**4.890**	**0.065**	**0.335**	**38.85**

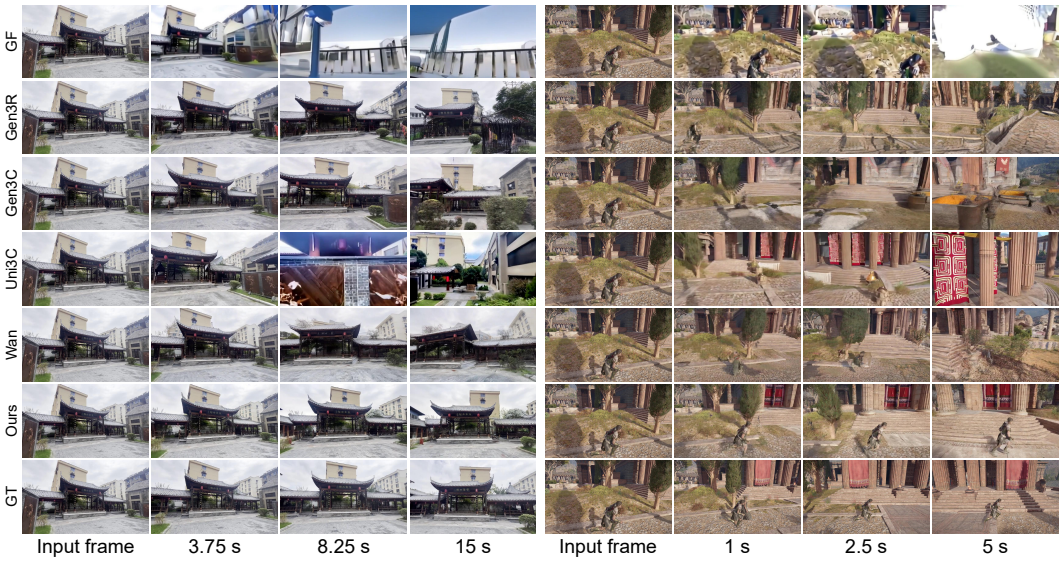


Figure 4: **Qualitative comparison of camera-controlled video generation. GAD** better preserves scene geometry over 15 seconds (left) and moving-object structure over 5 seconds (right).

quality, and imaging quality, with ground-truth scores provided for reference. Finally, we assess inference efficiency through runtime and GPU memory usage measured on an NVIDIA A800 GPU.

5.2 Results

Long-range scene consistency. Compared with the Wan baselines, **GAD** improves cross-view matching and camera accuracy while achieving better FID, FVD, and CLIP scores across both datasets and backbones Table 1. Notably, on Wan2.2 5B, matching hit rate increases by at least 30% on each dataset. Together, these gains demonstrate improved geometric consistency alongside better generation quality. Across both datasets, **GAD** variants achieve the best results among the baselines in both cross-view matching and rendering quality. Although Gen3C achieves lower camera-pose errors, both **GAD** variants outperform it in cross-view matching, indicating better preservation of scene correspondences. Qualitative comparisons in Figs. 4 and 7 further show that **GAD** better preserves scene layout and geometric structure throughout long-range camera-controlled sequences.

Dynamic video consistency. The benefits extend to dynamic videos: both **GAD** variants outperform the Wan baselines and other evaluated methods in subject consistency, background consistency, and motion smoothness (Table 2). Improvements in both subject and background scores suggest that correspondence supervision is useful beyond static scene content, consistent with transferring relations between moving objects rather than enforcing fixed spatial locations. Lower FID and FVD and higher CLIP scores further indicate that these gains coexist with improved generation quality and text alignment. Qualitative comparisons in Figs. 4 and 8 complement these scores: **GAD** better preserves the structure of moving characters and foreground objects, while maintaining a more stable background layout as the viewpoint changes. Attention visualizations in Fig. 5 further show that,

Table 5: **Distillation component ablation.** Confidence-aware teacher weighting (CAT) combined with trackwise (TWS) and pairwise sampling (PWS) gives the best results across all seven metrics.

CAT	TWS	PWS	FID↓	FVD↓	CLIP↑	R_{dist} ↓	T_{dist} ↓	M_{rate} ↑	M_{dist} ↓
×	×	×	38.60	322.1	33.85	9.573	0.108	0.257	48.90
×	✓	✓	39.04	329.4	33.92	8.753	0.092	0.230	42.94
✓	✓	×	38.61	321.4	34.01	6.028	0.082	0.274	40.20
✓	×	✓	38.35	320.3	34.05	5.312	0.076	0.291	39.67
✓	✓	✓	**37.93**	**318.0**	**34.24**	**4.890**	**0.065**	**0.335**	**38.85**

compared with the Wan baseline, **GAD** focuses video attention more consistently on corresponding scene regions in both static and dynamic examples. These more localized responses accompany better preservation of scene geometry and moving-object structure in the generated videos.

Applicability and inference efficiency. On DL3DV-10K with Wan2.2 5B, we further evaluate generation at 5, 10, and 15 seconds, separately training **GAD** and its Wan baseline under matched settings for each duration. As shown in Table 3, **GAD** improves all four geometric metrics at every duration, and complete results in Table 7 show broadly comparable or improved generation quality. Together, these results demonstrate the applicability of our method across backbone configurations and generation lengths. Notably, **GAD** achieves these gains without additional inference modules or computations, preserving the backbone's inference procedure. Table 2 reports unchanged peak GPU memory and minor runtime differences consistent with run-to-run variability. Time-weighted average memory is also sensitive to variations in the duration of individual execution stages.

5.3 Ablations

We study noise-dependent supervision and **GAD**'s components on DL3DV-10K with Wan2.2 5B.

Noise-dependent supervision. As shown in Table 4, fixed-layer supervision (R2–R4) already improves geometric metrics over the Wan baseline (R1), while the proposed noise-timestep routing (R8) outperforms all other configurations (R1–R7); complete metrics appear in Table 6. Thus, correspondence targets are useful on their own, while their placement determines how effectively they guide generation. The lower performance of R5–R7 compared with R2–R4 further suggests that varying supervision with noise is insufficient without an appropriate layer choice. These results support the motivation in Section 3: supervision should adapt to the noise timestep to target the structure-formation region rather than constrain a fixed layer throughout denoising.

Contributions of distillation components. Table 5 shows that transferring teacher correspondences indiscriminately is insufficient: without confidence weighting, matching hit rate falls below the Wan baseline despite modest improvements in camera error. This supports accounting for uncertainty in teacher matches rather than treating every target equally. With confidence weighting, either pairwise or trackwise supervision improves all geometric metrics, while their combination performs best across all seven metrics. The joint gains are consistent with their complementary roles: pairwise supervision discriminates spatial matches within a target frame, whereas trackwise supervision extends correspondence coverage across time.

6 Conclusion

We have presented **GAD**, an attention distillation framework that transfers geometric correspondence knowledge from a frozen geometry teacher to video attention. Motivated by geometric ambiguities in video attention and the noise-dependent emergence of scene structure, **GAD** guides cross-frame associations during structure formation while preserving the video model's representational flexibility. Experiments show that **GAD** consistently focuses attention on corresponding regions and improves long-range scene consistency and dynamic temporal consistency across two video backbones, with geometric improvements across generation lengths. These gains require no additional inference modules. Our findings highlight geometric correspondence transfer through attention as a promising approach to consistent video generation and video-based world simulation.

AI USE STATEMENT

AI tools assisted with revising prose and theoretical exposition, LaTeX formatting, and scripts used to assemble and visualize analysis outputs. The authors designed the method and experiments, ran the experiments, verified the reported results and mathematical arguments, and take full responsibility for the paper.

REPRODUCIBILITY STATEMENT

Section 4 defines the teacher targets, student responses, routing rule, and training objective. Sampling procedures and loss reductions are detailed in Appendix A, while implementation and evaluation protocols are provided in Appendices B and C. Complete routing and duration results appear in Appendix D, and the PCA and attention analysis protocols for Fig. 2 are described in Appendix E.

REFERENCES

Jan Ackermann, Shengqu Cai, Boyang Deng, Zhengfei Kuang, Songyou Peng, and Gordon Wetzstein. Geoflow: Enforcing implicit geometric consistency in video generation. *arXiv preprint arXiv:2605.18365*, 2026.

Zhaochong An, Orest Kupyn, Théo Uscidda, Andrea Colaco, Karan Ahuja, Serge Belongie, Mar Gonzalez-Franco, and Marta Tintore Gazulla. Vggrpo: Towards world-consistent video generation with 4d latent reward. *arXiv preprint arXiv:2603.26599*, 2026.

Yunpeng Bai, Shaoheng Fang, Chaohui Yu, Fan Wang, and Qixing Huang. Geovideo: Introducing geometric regularization into video generation model. *arXiv preprint arXiv:2512.03453*, 2025.

Aritra Bhowmik, Denis Korzhenkov, Cees G Snoek, Amirhossein Habibian, and Mohsen Ghafoorian. Moalign: Motion-centric representation alignment for video diffusion models. In *International Conference on Learning Representations*, volume 2026, pp. 122379–122397, 2026.

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

Chenjie Cao, Jingkai Zhou, Shikai Li, Jingyun Liang, Chaohui Yu, Fan Wang, Xiangyang Xue, and Yanwei Fu. Uni3c: Unifying precisely 3d-enhanced camera and human motion controls for video generation. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, pp. 1–12, 2025.

Ishaan Preetam Chandratreya, David Charatan, Basile Van Hoorick, Sergey Zakharov, Vitor Guizilini, Phillip Isola, and Vincent Sitzmann. Millivid: Hierarchical latents for long-range consistency in video generation. *arXiv preprint arXiv:2606.09056*, 2026.

Shuning Chang, Weihua Chen, Jiasheng Tang, Hao Xu, Zeyu Zhang, Hangjie Yuan, Yu Lu, Ruigang Niu, Fan Wang, Bohan Zhuang, et al. Towards error-free long video generation. *arXiv preprint arXiv:2606.22370*, 2026.

Fangda Chen, Shanshan Zhao, Longrong Yang, Chuanfu Xu, Zhigang Luo, and Long Lan. Freespec: Training-free long video generation via singular-spectrum reconstruction. *arXiv preprint arXiv:2605.06509*, 2026a.

Junyi Chen, Haoyi Zhu, Xianglong He, Yifan Wang, Jianjun Zhou, Wenzheng Chang, Yang Zhou, Zizun Li, Zhoujie Fu, Jiangmiao Pang, et al. Deepverse: 4d autoregressive video generation as a world model. *arXiv preprint arXiv:2506.01103*, 2025.

Pengguang Chen, Shu Liu, Hengshuang Zhao, and Jiaya Jia. Distilling knowledge via knowledge review. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5006–5015. IEEE, 2021.

Shuo Chen, Cong Wei, Sun Sun, Ping Nie, Kai Zhou, Ge Zhang, Ming-Hsuan Yang, and Wenhu Chen. Context forcing: Consistent autoregressive video generation with long context. *arXiv preprint arXiv:2602.06028*, 2026b.

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.

Yixuan Ding, Jiahao Kong, Wei Huang, Ruijie Quan, and Yi Yang. Layerrecall: A state-conditioned memory router for long-horizon consistency in video generation. *arXiv preprint arXiv:2608.28460*, 2026.

Carl Doersch, Ankush Gupta, Larisa Markeeva, Adria Recasens, Lucas Smaira, Yusuf Aytar, Joao Carreira, Andrew Zisserman, and Yi Yang. Tap-vid: A benchmark for tracking any point in a video. *Advances in Neural Information Processing Systems*, 35:13610–13626, 2022.

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. *arXiv preprint arXiv:2601.23286*, 2026.

Yilun Du, Sherry Yang, Bo Dai, Hanjun Dai, Ofir Nachum, Josh Tenenbaum, Dale Schuurmans, and Pieter Abbeel. Learning universal policies via text-guided video generation. *Advances in neural information processing systems*, 36:9156–9172, 2023.

Shaoheng Fang, Hanwen Jiang, Yunpeng Bai, Niloy J Mitra, and Qixing Huang. Worldreel: 4d video generation with consistent geometry and motion modeling. *arXiv preprint arXiv:2512.07821*, 2025.

Michal Geyer, Omer Bar Tal, Shai Bagon, and Tali Dekel. Tokenflow: Consistent diffusion features for consistent video editing. In *International Conference on Learning Representations*, volume 2024, pp. 1608–1620, 2024.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in neural information processing systems*, 35:8633–8646, 2022.

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

JiaKui Hu, Jialun Liu, Liying Yang, Xinliang Zhang, Kaiwen Li, Shuang Zeng, Yuanwei Li, Haibin Huang, Chi Zhang, and Yanye Lu. Geometry-as-context: Modulating explicit 3d in scene-consistent video generation to geometry context. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4258–4268, 2026.

Jiaxin Huang, Yuanbo Yang, Bangbang Yang, Lin Ma, Yuewen Ma, and Yiyi Liao. Gen3r: 3d scene generation meets feed-forward reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25358–25369, 2026a.

Junchao Huang, Guian Fang, Shengju Qian, Xianghao Kong, Zhuoran Zhao, Wei Huang, Yihua Du, Zixin Zhang, Justin Cui, Yuchao Gu, et al. Solarwm: Open data and scalable training for long-horizon video world models. *arXiv preprint arXiv:2609.02886*, 2026b.

Tianyu Huang, Wangguandong Zheng, Tengfei Wang, Yuhao Liu, Zhenwei Wang, Junta Wu, Jie Jiang, Hui Li, Rynson Lau, Wangmeng Zuo, et al. Voyager: Long-range and world-consistent video diffusion for explorable 3d scene generation. *ACM Transactions on Graphics (TOG)*, 44 (6):1–15, 2025a.

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*, 2025b.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21807–21818. IEEE, 2024.

Sungwon Hwang, Hyojin Jang, Kinam Kim, Minho Park, and Jaegul Choo. Cross-frame representation alignment for fine-tuning video diffusion models. *arXiv preprint arXiv:2506.09229*, 2025.

Wonbong Jang, Shikun Liu, Soubhik Sanyal, Juan Camilo Perez, Kam Woh Ng, Sanskar Agrawal, Juan-Manuel Perez-Rua, Yiannis Douratsos, and Tao Xiang. Rays as pixels: Learning a joint distribution of video and camera trajectories. In *Forty-third International Conference on Machine Learning*, 2026.

Sihui Ji, Xi Chen, Shuai Yang, Xin Tao, Pengfei Wan, and Hengshuang Zhao. Memflow: Flowing adaptive memory for consistent and efficient long video narratives. *arXiv preprint arXiv:2512.14699*, 2025.

Seonghyun Jin, Youngmin Kim, Sunwoo Park, and Jong Chul Ye. Crepe: Curved ray expectation positional encoding for unified-camera-controlled video generation. *arXiv preprint arXiv:2605.12938*, 2026.

Po-Chen Ko, Jiayuan Mao, Yilun Du, Shao-Hua Sun, and Joshua B Tenenbaum. Learning to act from actionless videos through dense correspondences. In *International Conference on Learning Representations*, volume 2024, pp. 40938–40958, 2024.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

Minkyung Kwon, Jinhyeok Choi, Jiho Park, Seonghu Jeon, Jinhyuk Jang, Junyoung Seo, Minseop Kwak, Jin-Hwa Kim, and Seungryong Kim. Correspondence-attention alignment for multi-view diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2316–2326, 2026.

Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. In *European conference on computer vision*, pp. 71–91. Springer, 2024.

Ruilong Li, Brent Yi, Junchen Liu, Hang Gao, Yi Ma, and Angjoo Kanazawa. Cameras as relative positional encoding. *Advances in Neural Information Processing Systems*, 2025a.

Runjia Li, Philip Torr, Andrea Vedaldi, and Tomas Jakab. Vmem: Consistent interactive video scene generation with surfel-indexed view memory. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 25690–25699. IEEE, 2025b.

Jiesong Lian, Zixiang Zhou, Ruizhe Zhong, Yuan Zhou, Qinglin Lu, Rui Wang, Long Hu, Yixue Hao, and Baoru Huang. Sara: Semantically adaptive relational alignment for video diffusion models. *arXiv preprint arXiv:2605.07800*, 2026.

Haotong Lin, Sili Chen, Junhao Liew, Donny Y Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. *arXiv preprint arXiv:2511.10647*, 2025.

Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, et al. 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.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Fangfu Liu, Wenqiang Sun, Hanyang Wang, Yikai Wang, Haowen Sun, Junliang Ye, Jun Zhang, and Yueqi Duan. Reconx: Reconstruct any scene from sparse views with video diffusion model. *IEEE Transactions on Image Processing*, 2026.

Ruijie Lu, Yixin Chen, Junfeng Ni, Baoxiong Jia, Yu Liu, Diwen Wan, Gang Zeng, and Siyuan Huang. Movis: Enhancing multi-object novel view synthesis for indoor scenes. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26767–26778. IEEE, 2025.

Yu Lu, Yuanzhi Liang, Linchao Zhu, and Yi Yang. Freelong: Training-free long video generation with spectralblend temporal attention. *Advances in Neural Information Processing Systems*, 37: 131434–131455, 2024.

Zelong Lv, Sicheng Xu, Jianfeng Xiang, Ruicheng Wang, Yue Dong, Yu Deng, Guangzhong Sun, and Jiaolong Yang. Octworld: Long-range world-consistent video generation with octree-based 3d mapping. *arXiv preprint arXiv:2609.03919*, 2026.

Ruibin Min, Yexin Liu, Aimin Pan, Changsheng Lu, Jiafei Wu, Kelu Yao, Xiaogang Xu, and Harry Yang. Ahpa: Adaptive hierarchical prior alignment for diffusion transformers. *arXiv preprint arXiv:2605.03317*, 2026.

David Nordström, Johan Edstedt, Georg Bökman, Jonathan Astermark, Anders Heyden, Viktor Larsson, Mårten Wadenbäck, Michael Felsberg, and Fredrik Kahl. Loma: Local feature matching revisited. *arXiv preprint arXiv:2604.04931*, 2026.

Lianyu Pang, Tianlin Pan, Cheng Da, Changqian Yu, Huan Yang, Kun Gai, Song Guo, and Wenhan Luo. Maskalign: Token-subset representation alignment for efficient diffusion training. *arXiv preprint arXiv:2606.08788*, 2026.

Jack Parker-Holder, Shlomi Fruchter, et al. Genie 3: A new frontier for world models. *Google DeepMind Blog*, 4, 2025.

Chenyang Qi, Xiaodong Cun, Yong Zhang, Chenyang Lei, Xintao Wang, Ying Shan, and Qifeng Chen. Fatezero: Fusing attentions for zero-shot text-based video editing. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15886–15896. IEEE, 2023.

Haonan Qiu, Menghan Xia, Yong Zhang, Yingqing He, Xintao Wang, Ying Shan, and Ziwei Liu. Freenoise: Tuning-free longer video diffusion via noise rescheduling. In *International Conference on Learning Representations*, volume 2024, pp. 5260–5274, 2024.

Yiqiao Qiu, Yixing Shen, Zhuohao Sun, Yanchong Zheng, Xiaobin Chang, Weishi Zheng, and Ruixuan Wang. Sats: Self-attention transfer for continual semantic segmentation. *Pattern Recognition*, 138:109383, 2023.

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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6121–6132. IEEE, 2025.

Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion. *Advances in neural information processing systems*, 36: 1363–1389, 2023.

Luyao Tang, Bingjun Luo, Dong Yi, Jialin Guo, Haoning Xi, Cheng Chen, Yizhou Yu, and Chaoqi Chen. Temporal state transport in video generation: Diagnosing and correcting spectral imbalance. *arXiv preprint arXiv:2609.08505*, 2026.

Frederick Tung and Greg Mori. Similarity-preserving knowledge distillation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1365–1374, 2019.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5294–5306. IEEE, 2025a.

Jianyuan Wang, Minghao Chen, Shangzhan Zhang, Nikita Karaev, Johannes Schönberger, Patrick Labatut, Piotr Bojanowski, David Novotny, Andrea Vedaldi, and Christian Rupprecht. Vggt-ω. *arXiv preprint arXiv:2605.15195*, 2026a.

Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10510–10522. IEEE, 2025b.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20697–20709. IEEE, 2024a.

Weijie Wang, Xiaoxuan He, Youping Gu, Yifan Yang, Zeyu Zhang, Yefei He, Yanbo Ding, Xirui Hu, Donny Y Chen, Zhiyuan He, et al. World-r1: Reinforcing 3d constraints for text-to-video generation. *arXiv preprint arXiv:2604.24764*, 2026b.

Weijie Wang, Haoyu Zhao, Yifan Yang, Feng Chen, Zeyu Zhang, Yefei He, Zicheng Duan, Donny Y Chen, Yuqing Yang, and Bohan Zhuang. Latent spatial memory for video world models. *arXiv preprint arXiv:2606.09828*, 2026c.

Xiaofeng Wang, Zheng Zhu, Guan Huang, Xinze Chen, Jiagang Zhu, and Jiwen Lu. Drivedreamer: Towards real-world-drive world models for autonomous driving. In *European conference on computer vision*, pp. 55–72. Springer, 2024b.

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.

Xiaofei Wu, Guozhen Zhang, Zhiyong Xu, Yuan Zhou, Qinglin Lu, and Xuming He. Pack and force your memory: Long-form and consistent video generation. *arXiv preprint arXiv:2510.01784*, 2025.

Haoning Yang, Xinyuan Chen, Yaohui Wang, and Guo Lu. Diff-vf: Training-free high-quality long video generation via diffusion model. *ACM Transactions on Multimedia Computing, Communications and Applications*, 2026a.

Sherry Yang, Yilun Du, Kamyar Ghasemipour, Jonathan Tompson, Leslie Kaelbling, Dale Schuurmans, and Pieter Abbeel. Learning interactive real-world simulators. *arXiv preprint arXiv:2310.06114*, 2023.

Shuai Yang, Wei Huang, Ruihang Chu, Yicheng Xiao, Yuyang Zhao, Xianbang Wang, Muyang Li, Enze Xie, Yingcong Chen, Yao Lu, et al. Longlive: Real-time interactive long video generation. *arXiv preprint arXiv:2509.22622*, 2025a.

Zhenhao Yang, Xiaoshi Wu, Zhengyao Lv, Xiaoyu Shi, Xintao Wang, Pengfei Wan, Kun Gai, and Kwan-Yee K Wong. Decmem: Towards minute-long consistent world generation with decoupled memory. *arXiv preprint arXiv:2605.31336*, 2026b.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. In *International Conference on Learning Representations*, volume 2025, pp. 83048–83077, 2025b.

Minghao Yin, Jiahao Lu, Wenbo Hu, Wang Zhao, Ying Shan, and Kai Han. Scope: Sightline-coordinate positional encoding for video diffusion transformers. *arXiv preprint arXiv:2606.27345*, 2026.

Tianwei Yin, Qiang Zhang, Richard Zhang, William T Freeman, Fredo Durand, Eli Shechtman, and Xun Huang. From slow bidirectional to fast autoregressive video diffusion models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22963–22974. IEEE, 2025.

Hong-Xing Yu, Haoyi Duan, Charles Herrmann, William T Freeman, and Jiajun Wu. Wonderworld: Interactive 3d scene generation from a single image. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5916–5926. IEEE, 2025.

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*, 2024a.

Wangbo Yu, Jinbo Xing, Li Yuan, Wenbo Hu, Xiaoyu Li, Zhipeng Huang, Xiangjun Gao, Tien-Tsin Wong, Ying Shan, and Yonghong Tian. Viewcrafter: Taming video diffusion models for high-fidelity novel view synthesis. *arXiv preprint arXiv:2409.02048*, 2024b.

Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv preprint arXiv:1612.03928*, 2016.

Lvmin Zhang, Shengqu Cai, Muyang Li, Gordon Wetzstein, and Maneesh Agrawala. Frame context packing and drift prevention in next-frame-prediction video diffusion models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025a.

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

Wenqu Zhao, Xuemin Chi, Xin Zhang, Guoqing Ma, Baorun Li, Jianjie Fang, Peizhi Tang, Chen Gao, and Wei Wu. Densitykv: Density-guided kv cache compression for long video generation. *arXiv preprint arXiv:2608.27922*, 2026.

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.

Yang Zhou, Yifan Wang, Jianjun Zhou, Wenzheng Chang, Haoyu Guo, Zizun Li, Kaijing Ma, Xinyue Li, Yating Wang, Haoyi Zhu, et al. Omniworld: A multi-domain and multi-modal dataset for 4d world modeling. *arXiv preprint arXiv:2509.12201*, 2025.

Yupeng Zhou, Daquan Zhou, Ming-Ming Cheng, Jiashi Feng, and Qibin Hou. Storydiffusion: Consistent self-attention for long-range image and video generation. *Advances in Neural Information Processing Systems*, 37:110315–110340, 2024.

Hongzhou Zhu, Min Zhao, Guande He, Hang Su, Chongxuan Li, and Jun Zhu. Causal forcing: Autoregressive diffusion distillation done right for high-quality real-time interactive video generation. *arXiv preprint arXiv:2602.02214*, 2026.

Tianrui Zhu, Shiyi Zhang, Zhirui Sun, Jingqi Tian, and Yansong Tang. Memorize-and-generate: Towards long-term consistency in real-time video generation. *arXiv preprint arXiv:2512.18741*, 2025.

Wenqiang Zu, Shenghao Xie, Bo Lei, and Lei Ma. Training-free representation guidance for diffusion models with a representation alignment projector. *arXiv preprint arXiv:2601.22468*, 2026.

A Loss and Sampling Details

Sampling source–target relations. Pairwise supervision samples valid directed frame pairs without replacement, excluding pairs within the same frame. Within each selected pair, source tokens are sampled uniformly without replacement from those with valid cached correspondences. Each selected query is compared with all spatial keys in its target frame. If fewer valid pairs or queries are available than the sampling budget, the remaining entries are masked out.

Trackwise supervision first samples one source frame containing eligible source tokens, then samples tokens that have valid correspondences in at least two other frames. Each source is matched to all of its valid target frames, excluding its own frame. For each target frame, the candidate set contains the Gaussian positive neighborhood and uniformly sampled spatial negatives without replacement from outside that neighborhood. Both schemes resample their relations during training; confidence weights affect the loss reduction rather than the probability of selecting a valid relation.

Pairwise reduction. Let $\mathcal{P}$ denote the sampled valid source–target-frame relations, and Ω_j all spatial tokens in target frame j. Using $\mathcal{L}_{\mathrm{corr}}$ from Eq. (3), the pairwise loss is

$$\mathcal{L}_{\mathrm{pair}} = \frac{\sum_{(a,j)\in\mathcal{P}} w_{a,j}\left[\mathcal{L}_{\mathrm{corr}}(a,j;\Omega_j) + \eta\mathcal{L}_{\mathrm{hard}}(a,j)\right]}{\sum_{(a,j)\in\mathcal{P}} w_{a,j}}. \tag{5}$$

The hard-negative term contrasts the weighted positive partition with the highest-scoring negatives outside the positive neighborhood:

$$\begin{aligned} Z^{+}_{a,j} &= \sum_{v\in\mathcal{N}(c_{a,j})} y_{a,j}(v)\exp s_{a,j}(v), \\ Z^{-}_{a,j} &= \sum_{v\in\mathrm{Top}_{K_{\mathrm{neg}}}(\Omega_j\setminus\mathcal{N}(c_{a,j}))} \exp s_{a,j}(v), \\ \mathcal{L}_{\mathrm{hard}}(a,j) &= -\log\frac{Z^{+}_{a,j}}{Z^{+}_{a,j}+Z^{-}_{a,j}}. \end{aligned} \tag{6}$$

Top-K_{neg} selection uses the student logits, with $K_{\mathrm{neg}} = 64$. The positive target is normalized over valid in-grid neighbors; invalid border positions are excluded. The hard-negative coefficient η is fixed within each controlled comparison.

Trackwise reduction. Let $\mathcal{S}$ be sampled source tokens and $\mathcal{J}_a$ the valid target frames for source a, excluding its own frame. Each $\mathcal{K}_{a,j}$ contains the positive neighborhood and randomly sampled negatives from outside it. We define

$$\begin{aligned} \overline{\mathcal{L}}_{\mathrm{corr},a} &= \frac{\sum_{j\in\mathcal{J}_a} w_{a,j}\mathcal{L}_{\mathrm{corr}}(a,j;\mathcal{K}_{a,j})}{\sum_{j\in\mathcal{J}_a} w_{a,j}}, \\ \overline{w}_a &= \frac{1}{|\mathcal{J}_a|}\sum_{j\in\mathcal{J}_a} w_{a,j}, \\ \mathcal{L}_{\mathrm{track}} &= \frac{\sum_{a\in\mathcal{S}} \overline{w}_a\overline{\mathcal{L}}_{\mathrm{corr},a}}{\sum_{a\in\mathcal{S}}\overline{w}_a}. \end{aligned} \tag{7}$$

With uniform confidence, each source contributes its average target-frame loss, irrespective of track length. With nonuniform confidence, uncertain target frames contribute less within a source, and sources with low average confidence contribute less overall. Empty valid sets produce zero auxiliary loss; implementations clamp denominators for numerical stability.

Combining the two objectives. The implementation uses an overall auxiliary coefficient λ_{aux} and relative branch weights β_{pair} and β_{track}. With both branches active, the coefficients in Eq. (4) are

$$\begin{aligned} \lambda_1 &= \lambda_{\mathrm{aux}}\frac{\beta_{\mathrm{pair}}}{\beta_{\mathrm{pair}}+\beta_{\mathrm{track}}}, \\ \lambda_2 &= \lambda_{\mathrm{aux}}\frac{\beta_{\mathrm{track}}}{\beta_{\mathrm{pair}}+\beta_{\mathrm{track}}}. \end{aligned} \tag{8}$$

When only one branch is enabled or has valid supervision, the branch weights are renormalized over the active branch. If neither branch has valid supervision, the auxiliary loss is zero. This convention keeps the overall auxiliary coefficient separate from the allocation between the two objectives.

B Implementation Details

B.1 Teacher Targets and Alignment

Reference correspondences for head selection. We select the teacher layer–head pair using reference correspondences from both static and dynamic scenes. For the static set, we randomly sample 50 DL3DV scenes and estimate per-view depth and camera parameters with VGGT-Ω, retaining only high-confidence depth estimates. Inspired by CAMEO (Kwon et al., 2026), we construct correspondence labels through cycle projection. A sampled source pixel is unprojected into 3D using its depth and camera parameters, then projected into a target view and assigned to a target token. We unproject the center of that target token using the target-view depth and project it back into the source view. A correspondence is retained only if the projections remain in bounds, the projected and estimated depths agree in both directions, and the return location is sufficiently close to the source sample. When multiple samples within a source token pass these checks, we retain the correspondence with the smallest cycle error. These are geometry-derived reference labels, rather than manually annotated ground truth. For the dynamic set, we use the provided ground-truth point correspondences from 30 dynamic scenes in TAPVid-DAVIS (Doersch et al., 2022).

Layer–head evaluation and selection. We randomly sample query tokens from the input frame of each video and examine every global-attention layer–head pair. For each query, we extract the maximum-response location in each subsequent view and its target-frame attention mass, following Eq. (1). We compare the predicted locations with the reference correspondences to measure correspondence error and examine the associated attention masses as reliability cues. Considering correspondence consistency across both static and dynamic tokens, we select layer 13, head 10, which provides the best overall performance in this evaluation. All layer and head indices here are zero-based, and this single pair is used for training-label extraction in all experiments.

The evaluation also reveals substantial specialization across heads. For example, layer 13, head 0 captures static-token correspondences well but fails to reliably associate dynamic tokens, whereas layer 12, head 12 captures dynamic-token correspondences well but performs poorly on static tokens. Other pairs exhibit little correspondence consistency for either type of token. These observations motivate selecting a pair using both static and dynamic scenes, rather than relying on either set alone.

Limitations of projection-based labels. The depth-and-camera cycle projection above assumes a static scene and therefore cannot directly recover correspondences for moving scene elements. We also observe substantial errors around thin structures and closely spaced surfaces, where depth errors and token discretization can associate different surfaces. These limitations can affect methods that rely on such projection-derived supervision, including CAMEO. Extracting correspondences from VGGT-Ω attention instead exploits the model's learned cross-view consistency without imposing a static-scene projection constraint. In our observations, this alleviates these failure cases and also supports correspondences for dynamic scene elements.

Teacher extraction. We precompute correspondence targets with the frozen VGGT-Ω 1B teacher, using its 512-resolution checkpoint and global-attention layer 13, head 10 (zero-based indices). The teacher processes clean video frames. Attention is normalized over all tokens in all teacher input frames, including special tokens, whereas the target-frame mass sums only the patch-token responses in that frame.

Source patch queries are resampled to the student spatial grid before query normalization. For each target frame, the conditional patch-attention map is spatially resampled and renormalized before taking its maximum; its frame attention mass is retained separately. In Eq. (1), $\bar{A}$ denotes this aligned conditional map multiplied by the retained mass. Its sum and argmax therefore recover the cached mass and correspondence, respectively. The cache stores these quantities rather than the full attention matrix.

Temporal and spatial alignment. Teacher inputs follow the temporal representative indices used by the training label pipeline. For a $4n+1$-frame clip, these are $[0, 2, 6, \ldots, 4n-2]$, with the clip's source offset applied when reading the original video. The resulting $n+1$ teacher views align with the student's latent time axis. The spatial grids are 30×52 for Wan2.1 1.3B and 15×26 for Wan2.2 5B at 832×480 resolution. RGB frames, camera intrinsics, and teacher targets use consistent crop and resize geometry. Cache metadata records the source frames, clip offset, grid, and teacher identity, and is checked when loading labels. When a shorter prefix is read from a longer cache, the retained masses remain those computed in the original teacher context; they are not recomputed by renormalizing over the prefix.

Confidence weighting. We map the target-frame mass to a continuous weight:

$$w_{a,j} = \chi_{a,j} \left[\mathrm{sigmoid}\left(\frac{\log_2(\max(m_{a,j}, \varepsilon)) - \mu}{\kappa} \right) \right]^{\gamma}, \tag{9}$$

where $\chi_{a,j} \in \{0, 1\}$ indicates a valid cached relation. We use $\varepsilon = 2^{-20}$ and set $\gamma = 2$ empirically to further downweight low-confidence relations. The grid-specific (μ, κ) values are approximately $(-9.86777, 1.81857)$ for the 30×52 grid and $(-9.75580, 1.59004)$ for the 15×26 grid. The logarithm accommodates attention masses spanning multiple orders of magnitude, and the lower bound avoids taking the logarithm of zero. These weights set the relative contributions of correspondences in Eqs. (5) and (7).

Human visibility annotations. We determine the sigmoid mapping from human visibility annotations on DL3DV, using attention labels computed in an 81-frame teacher context. The annotated subset contains 55 scenes: 48 randomly sampled scenes and seven manually selected challenging scenes involving thin structures, occlusions, or out-of-view content. Each annotation page shows one source token and seven target frames sampled from seven temporal intervals, independently of attention mass. Annotators judge whether the source-token content is visible, not visible, or uncertain in each target frame; attention statistics are hidden during annotation. Of 803 judgments, we exclude 230 uncertain cases and retain 573 source-token/target-frame pairs, comprising 401 visible and 172 not-visible cases. These labels assess visibility rather than the localization accuracy of the attention maximum.

Selecting and fitting the mass mapping. We compare three attention statistics as visibility cues for each source-token/target-frame pair (a, j). The *mass*, $m_{a,j} = \sum_{v \in \Omega_j} \bar{A}_{a,j}(v)$, is the total attention assigned to the target frame's spatial patch tokens, with the original normalization over all tokens in all teacher input frames preserved, as in Eq. (1). To describe how this attention is distributed within the target frame, let $p_{a,j}(v) = \bar{A}_{a,j}(v)/m_{a,j}$ denote the conditional distribution on the aligned spatial grid Ω_j for $m_{a,j} > 0$. The *normalized attention entropy* is $H_{a,j} = -\sum_{v \in \Omega_j} p_{a,j}(v) \log p_{a,j}(v) / \log |\Omega_j|$, with $0 \log 0 = 0$; it ranges from zero for a single-token peak to one for a uniform distribution. The *top-1 attention response* is $\max_{v \in \Omega_j} p_{a,j}(v)$, the largest conditional attention probability in that frame. Thus, mass measures how much attention reaches a frame, whereas entropy and top-1 describe its spatial concentration within that frame. We use $\log_2(\max(m_{a,j}, \varepsilon))$, $1 - H_{a,j}$, and the top-1 response as regression features.

In five-fold cross-validation grouped by scene, logistic regression using only log-mass achieves ROC-AUCs of 0.903 and 0.932 for the 30×52 and 15×26 layouts, respectively. Adding entropy and the top-1 response increases these values only to 0.910 and 0.933, so we retain the simpler mass-only mapping. Grouping by scene keeps all pairs from a scene in the same fold.

For each layout, we fit a class-balanced logistic regression on all 573 definite annotations, with $x = \log_2(\max(m, \varepsilon))$ as its sole input. Its output $\mathrm{sigmoid}(ax + b)$ is equivalently written as $\mathrm{sigmoid}((x - \mu)/\kappa)$, with $\mu = -b/a$ and $\kappa = 1/a$, yielding the parameters reported above. Class balancing reduces the influence of the larger visible class, yielding a balanced visibility confidence for weighting correspondence supervision. We use this continuous confidence instead of a hard visibility threshold, then apply the empirical exponent γ in Eq. (9); the exponent is not part of the logistic fit. The mapping fitted on static DL3DV scenes is shared across static and dynamic supervision.

Local correspondence targets. We use a 3×3 Gaussian positive neighborhood with $\sigma = 1$ in token units. Weights are computed from token-grid distances to the cached correspondence, with

out-of-grid positions excluded and the remaining weights normalized to sum to one. This local target tolerates small localization errors without requiring the student to reproduce the teacher's full attention distribution.

B.2 Student Responses and Noise-Timestep Routing

Multi-head response aggregation. Let $\mathbf{q}_a^{\ell,h}$ and $\mathbf{k}_{j,v}^{\ell,h}$ be the query and key after normalization and rotary positional encoding at layer ℓ and head h. We aggregate their relation scores with a learnable gate:

$$\begin{aligned} r_{a,j}^h(v) &= \frac{\langle \mathbf{q}_a^{\ell,h}, \mathbf{k}_{j,v}^{\ell,h} \rangle}{\sqrt{d}}, \\ \boldsymbol{\alpha}_{a,j}(v) &= \mathrm{softmax}(g_\phi(\mathbf{r}_{a,j}(v))), \\ s_{a,j}(v) &= \frac{1}{\tau} \sum_{h=1}^{H} \alpha_{a,j}^h(v) r_{a,j}^h(v), \\ P_{a,j}(v \mid \mathcal{K}) &= \frac{\exp s_{a,j}(v)}{\sum_{v' \in \mathcal{K}} \exp s_{a,j}(v')}. \end{aligned} \tag{10}$$

Here, d is the head dimension, H is the number of heads, and $\mathcal{K}$ contains candidate spatial keys in the target frame. The head-aggregation MLP g_ϕ serves as the gate and has an $H \to 4H \to H$ architecture with a GELU activation. Its final linear layer is initialized to zero, producing uniform head weights at initialization. Each supervised layer has its own gate. Auxiliary logits and aggregation are computed in FP32, with temperature $\tau = 1$. The gate acts on each query–key pair and does not replace the attention operation used for generation. The pairwise hard-negative term uses $K_{\mathrm{neg}} = 64$ negatives and coefficient $\eta = 0.1$.

Noise coordinate and layer selection. Throughout the paper, $t \in [0, 1000]$ denotes the model noise timestep presented to the network by the Wan scheduler after applying its noise schedule. The corresponding normalized interpolation coefficient is $s = t/1000 \in [0, 1]$. Noise-timestep routing compares this model noise timestep with the configured thresholds, rather than using the integer index of a training or inference step. We use the same routing schedule for Wan2.1 1.3B and Wan2.2 5B in both the static and dynamic generation settings: L5 for $t < 850$, L15 for $850 \le t < 950$, and L25 for $t \ge 950$, as specified in Table 4. All layer indices are zero-based. Boundary values belong to the higher-noise interval, and only one layer contributes auxiliary supervision for a sample. The noise-timestep values in Fig. 2a and the routing tables use this same coordinate.

B.3 Training and Inference

Training configuration. As described in Section 5.1, the training corpus contains 35K clips, and training runs for 10 epochs on 32 NVIDIA A800 GPUs. We fine-tune the video DiT and aggregation MLPs using AdamW with a learning rate of 10^{-5}, keeping the text encoder, VAE, and geometry teacher frozen. Clips have a resolution of 832×480, with 241 frames for long-range generation and 81 frames for dynamic generation. The duration comparison additionally uses 161-frame clips. At 16 fps, the first-to-last-frame spans are 5, 10, and 15 seconds for 81, 161, and 241 frames. Each task and duration uses a separate Wan/**GAD** training pair, sharing initialization, data, optimizer, learning-rate schedule, effective batch size, and training duration.

For joint pairwise and trackwise supervision, we sample up to 32 directed frame pairs with 256 source queries per pair, together with 256 trackwise source tokens and 128 random negatives per source–target-frame relation. In the single-branch ablations in Table 5, pairwise-only supervision uses up to 128 frame pairs with 320 queries per pair, while trackwise-only supervision uses up to 384 source tokens and 192 random negatives per source–target-frame relation. These budgets apply to each training sample at its selected supervision layer. Pairwise queries are compared with all spatial tokens in the target frame; trackwise sources are sampled from one source frame and matched to all valid target frames, with the positive neighborhood included in addition to the sampled negatives. Sampling and masking follow Appendix A.

Across all **GAD** experiments, including both backbones and the static and dynamic generation settings, we use $\lambda_{\mathrm{aux}} = 0.05$ and equal relative branch weights $\beta_{\mathrm{pair}} = \beta_{\mathrm{track}} = 1$. Following the

normalization in Appendix A, this gives $\lambda_1 = \lambda_2 = 0.025$ in Eq. (4) when both branches are active. If only one branch is active, its coefficient is 0.05.

Flow-matching objective. For model noise timestep $t \in [0, 1000]$, let $s = t/1000$ be the normalized interpolation coefficient. We sample $\epsilon \sim \mathcal{N}(0, I)$ and construct $z_t = (1 - s)z_0 + s\epsilon$. Under conditions $\mathbf{c}$, the video model v_θ receives t and predicts the velocity $\epsilon - z_0$ with respect to s. We retain the backbone's noise-timestep sampling and loss weighting $\rho(t)$, expressed in the model-timestep coordinate:

$$\mathcal{L}_{\mathrm{FM}} = \mathbb{E}_{z_0,\epsilon,t}\left[\rho(t)\|v_\theta(z_t, t, \mathbf{c}) - (\epsilon - z_0)\|_2^2\right]. \tag{11}$$

For all three benchmarks, $\mathbf{c}$ includes the first frame, a text prompt, and a target camera trajectory. The backbone's first-frame conditioning mechanism is retained during training and inference.

Auxiliary computation and inference. Pairwise supervision evaluates sampled queries against all spatial keys in their selected target frames. Trackwise supervision evaluates only the positive and sampled-negative candidates for its selected sources. Neither branch constructs a full all-frame attention matrix. These computations add training cost but are absent at inference, together with the teacher and correspondence caches. Inference uses only the trained video model and its original sampler; all requested frames are jointly denoised over the sampling steps. Paired Wan and **GAD** evaluations use the same input conditions and generation seeds.

C Experimental Protocol

C.1 Data Preparation and Conditioning

We use RealEstate10K and DL3DV-10K for 15-second long-range generation and OmniWorld-Game for 5-second dynamic generation. DL3DV-10K uses its official evaluation set; RealEstate10K and OmniWorld-Game use scene-disjoint training and test splits. All methods receive the first image, the text prompt, and the target camera trajectory for the same clip.

RealEstate10K cameras are obtained from COLMAP reconstructions and aligned to the video frame indices. Their translation scale is aligned to a Depth Anything 3 (Lin et al., 2025) metric reference. DL3DV-10K uses VIPE camera estimates aligned to the extracted RGB frames. For OmniWorld-Game, we use the game subset from the xhigh tier of the SolarWM (Huang et al., 2026b) release, together with its camera annotations and captions. Spatial preprocessing applies the same crop and resize to RGB images and camera intrinsics. Camera trajectories are expressed relative to their first frame, and model-specific translation scaling is applied separately from the common evaluation trajectory.

C.2 Baseline Adaptation

The Wan baselines follow Appendix B. For GF, Gen3R, Voyager, Gen3C, and Uni3C, we use their released implementations and checkpoints with adapters for the shared image, prompt, and camera inputs. Each method retains its native generation procedure, including any depth estimation, geometry rendering, and history conditioning. Outputs are aligned to the common evaluation resolution of 832×480 and to 241 or 81 timestamps at 16 fps. Native frame counts and model resolutions can differ; the adapters transform camera intrinsics with the corresponding image geometry.

Long-range generation. GF retains its sparse-keyframe prediction and iterative interpolation procedure over the requested trajectory. Gen3R generates three 49-frame segments, conditioning each subsequent segment on the previous generated endpoint. Removing duplicate endpoints yields 145 native frames, which are aligned to the 241 evaluation timestamps by nearest-timestamp resampling. Voyager and Uni3C generate three 81-frame segments with one shared endpoint, yielding $81 + 80 + 80 = 241$ frames. Voyager propagates the previous generated RGB-D endpoint while retaining its initial world cache; its four internal prefix frames are removed from each segment's decoded output. Uni3C uses the previous generated endpoint as the next reference image and constructs its geometric conditions for the next segment. Gen3C uses its native long-video generation path over the requested camera sequence. Its adapter preserves the input field of view by resizing and padding to the native canvas, then removes the padding and resizes the output for evaluation. Segment continuation uses generated content rather than additional ground-truth RGB frames.

Dynamic generation. The OmniWorld-Game adapters cover the same 0–5-second interval. Voyager and Uni3C use one 81-frame segment, with Voyager retaining its internal prefix mechanism. GF retains its keyframe and interpolation procedure. Gen3R generates 49 native frames and resamples them to 81 output timestamps. Gen3C generates a 121-frame native window over the same interval and selects the nearest native timestamps for the 81-frame evaluation sequence. These conversions change the temporal sampling of the output, not the evaluated camera interval.

Camera-scale adaptation. Warping-based methods are sensitive to camera scale because their explicit scene initialization and geometric conditioning depend on depth estimated from the first frame. Camera translations must therefore be expressed at a scale consistent with this depth; a mismatch produces an inconsistent initial scene and incorrect warps, which can propagate errors into subsequent predictions. This dependence on depth–camera scale calibration is a practical limitation of these explicit geometric methods. To provide suitable geometric inputs for a fair comparison, we carefully align camera translations with the depth scale used by each baseline, using the dataset- and model-specific corrections described below. For Gen3C and Voyager, the DL3DV adapter applies the recorded VIPE-to-MoGe scale, approximately 0.370913. On RealEstate10K, their input translations use the per-scene COLMAP-to-DA3 scale followed by the fixed DA3-to-MoGe factor, approximately 0.632710. GF and Gen3R retain their internal camera normalization after the common input conversion. For Uni3C on RealEstate10K, camera translations are aligned to the DA3 depth scale using a per-scene scale factor.

For OmniWorld-Game, the supplied metric poses are scaled by approximately 0.227987 for Gen3C/Voyager and 0.597926 for Uni3C. These factors were estimated from training-side depth comparisons and fixed before evaluation. GF and Gen3R receive the metric poses without an additional dataset-specific multiplier. Scale factors affect translations only; rotations and intrinsics are unchanged. Common evaluation poses are stored separately and do not include model-specific scale factors.

C.3 Evaluation Metrics

Generation quality. We exclude the first conditioning frame from the target-frame quality measurements. FID uses 2048-dimensional Inception features from frames $4, 8, \ldots, T-1$, pooled across the complete evaluation set. FVD uses 400-dimensional I3D features from the target-frame video clips and is also computed at the complete-set level. CLIP uses ViT-B/32 image and text embeddings, sampling every fourth target frame and averaging $\max(100 \cos(\mathbf{e}_{\text{image}}, \mathbf{e}_{\text{text}}), 0)$.

Camera accuracy. Depth Anything 3 (Lin et al., 2025) estimates camera poses from the generated video. Predicted and reference trajectories are each transformed relative to their first camera, and their translations are independently normalized by the maximum distance from that camera. We average the rotation geodesic distance and normalized translation L2 distance over frames. Rotation distances are reported in degrees, as in Table 1; translation distances are dimensionless. This normalization evaluates relative trajectory agreement rather than absolute metric scale.

Cross-view consistency. LoMa-B (Nordström et al., 2026) extracts up to 2,048 keypoints per frame. We compare first-frame-to-target-frame matches in generated and reference videos at frames $4, 8, \ldots, T-1$. Reference matches are associated with generated matches by their source positions, using a 20-pixel threshold and consuming an assigned generated match only once. A hit additionally requires the target positions to differ by less than 20 pixels. M_{rate} is the number of hits divided by the number of reference matches; M_{dist} averages target-position distances over source-assigned matches, including those that miss the target threshold. Dataset-level scores aggregate the corresponding counts and distance sums, rather than averaging scene-level ratios.

Dynamic consistency and quality. We use VBench's custom-input evaluation for the five dimensions in Table 2: subject consistency, background consistency, motion smoothness, aesthetic quality, and imaging quality. Generated and ground-truth videos are evaluated separately with the same implementation, and each dimension is averaged over videos.

C.4 Inference Cost

The costs in Table 2 are measured on one NVIDIA A800 using three common OmniWorld-Game scenes, with methods run sequentially. The Wan variants generate 81 frames at 832×480 using 50 sampling steps, guidance scale 5, and seed 0. Each external method retains its native resolution, temporal sampling, precision, and offloading policy. The comparison therefore measures the cost of completing the common generation task under each method's evaluation protocol.

Runtime includes required preprocessing, generation, output postprocessing, and generated-video encoding, but excludes model loading and uninstrumented process overhead. GPU memory is monitored with NVML over the scene's process tree, with a target sampling interval of 50 ms and decimal GB units. Average memory is the time-weighted process-tree memory over the full scene-command duration, including loading and export; peak memory is the maximum sampled value in that run. The table reports the arithmetic mean of each quantity over the three scenes, so its peak-memory column is the mean of the per-scene peaks. For **GAD**, the trained DiT replaces the corresponding backbone DiT, and no teacher or aggregation MLP is loaded.

D Additional Quantitative Results

Tables 6 and 7 report the complete results corresponding to the compact comparisons in Tables 3 and 4, including rendering metrics and the Wan baseline (R1).

Loss assignment configurations. The Loss Assignment column in Table 6 specifies where the pairwise and trackwise auxiliary losses are applied during training. The flow-matching loss remains active in every configuration and at every noise timestep. Here, $t \in [0, 1000]$ is the model noise timestep, with larger values indicating higher noise, and Lk denotes student layer k using zero-based indexing. The eight configurations are:

- **R1 (Off @ all t):** The Wan baseline, with no auxiliary attention supervision at any noise timestep.
- **R2 (L5 @ all t):** Both auxiliary losses are applied at layer 5 throughout the noise range, testing a fixed shallow supervision layer.
- **R3 (L15 @ all t):** Both auxiliary losses are applied at layer 15 throughout the noise range, testing a fixed intermediate supervision layer.
- **R4 (L25 @ all t):** Both auxiliary losses are applied at layer 25 throughout the noise range, testing a fixed deep supervision layer.
- **R5 (L5 @ $t \geq 900$):** Auxiliary supervision is applied only at layer 5 for high-noise samples with $t \geq 900$; samples with $t < 900$ receive only the flow-matching loss.
- **R6 (L25 @ $t < 900$):** Auxiliary supervision is applied only at layer 25 for lower-noise samples with $t < 900$; samples with $t \geq 900$ receive only the flow-matching loss.
- **R7 (L25 / L5 @ 900):** Auxiliary supervision is routed to layer 25 for $t < 900$ and layer 5 for $t \geq 900$. This schedule assigns a deeper layer to lower noise and a shallower layer to higher noise.
- **R8 (L5 / L15 / L25 @ 850, 950):** Our schedule routes auxiliary supervision to layer 5 for $t < 850$, layer 15 for $850 \leq t < 950$, and layer 25 for $t \geq 950$, assigning progressively deeper layers as noise increases.

For every supervised sample, both auxiliary objectives use the same selected layer; the listed layers are not supervised simultaneously. At each routing boundary, equality belongs to the higher-noise interval. The auxiliary loss weights follow Appendix B.3.

The routed configuration improves all four geometric metrics over the tested fixed-layer and alternative schedules. These results support adapting supervision depth to the noise level. Across durations, **GAD** improves the geometric metrics at 5, 10, and 15 seconds, while rendering metrics show small trade-offs at the shorter durations. Each duration uses a separately trained Wan/**GAD** pair, so this comparison assesses applicability across training lengths.

Table 6: **Complete noise-timestep routing results on DL3DV-10K with Wan2.2 5B.** R1–R8 follow Table 4; R8 is **GAD**. Layer sequences run from low to high noise; values after @ are model-timestep boundaries, assigned to the higher-noise interval. For inequalities, supervision is off outside the stated interval. Best results are bold.

Exp. ID	Loss Assignment	FID↓	FVD↓	CLIP↑	R_{dist} ↓	T_{dist} ↓	M_{rate} ↑	M_{dist} ↓
R1 (Wan)	Off @ all t	38.60	322.1	33.85	9.573	0.108	0.257	48.90
R2	L5 @ all t	38.03	321.0	**34.24**	7.440	0.086	0.295	43.89
R3	L15 @ all t	38.18	319.1	34.12	7.346	0.090	0.286	42.71
R4	L25 @ all t	38.07	320.3	34.17	6.540	0.075	0.289	43.55
R5	L5 @ $t \geq 900$	38.64	323.9	34.09	8.932	0.098	0.262	46.36
R6	L25 @ $t < 900$	38.80	322.9	34.03	9.290	0.106	0.258	47.17
R7	L25 / L5 @ 900	38.48	320.8	34.12	8.755	0.104	0.268	45.77
R8 (ours)	**L5 / L15 / L25 @ 850, 950**	**37.93**	**318.0**	**34.24**	**4.890**	**0.065**	**0.335**	**38.85**

Table 7: **Complete generation-duration results on DL3DV-10K with Wan2.2 5B.** Separately trained Wan/**GAD** pairs show geometric gains at every duration, with small FID or FVD trade-offs at 5 and 10 seconds. Best results within each duration are bold.

Duration	FID↓	FVD↓	CLIP↑	R_{dist} ↓	T_{dist} ↓	M_{rate} ↑	M_{dist} ↓
5s	**26.85**	285.7	34.22	3.020	0.064	0.529	21.84
5s + **GAD**	27.63	**285.1**	**34.26**	**1.684**	**0.053**	**0.571**	**19.31**
10s	33.68	**287.4**	34.03	5.948	0.072	0.371	38.34
10s + **GAD**	**33.49**	288.4	**34.14**	**3.376**	**0.056**	**0.453**	**33.33**
15s	38.60	322.1	33.85	9.573	0.108	0.257	48.90
15s + **GAD**	**37.93**	**318.0**	**34.24**	**4.890**	**0.065**	**0.335**	**38.85**

E PCA AND ATTENTION ANALYSIS DETAILS

E.1 VIDEO FEATURES AND ATTENTION

Generation and feature capture. The analysis in Fig. 2a uses Wan2.2 5B on a RealEstate10K example, generating 81 frames at 832×480 with 50 sampling steps, guidance scale 5, scheduler shift 5, and seed 0. Features and attention are captured from the positive classifier-free-guidance branch during the same complete generation trajectory. The source query is fixed in the first latent frame at spatial token $(x, y) = (14, 8)$ on the 26×15 grid. The target is latent frame 10, displayed over frame 40 of the final generated video. This displayed frame is a visualization representative, distinct from the temporal representatives used to extract training labels.

PCA projection. For each inspected layer and denoising step, we collect the target frame's block-output features and project the spatial tokens onto three principal components. PCA is fitted independently for each layer–step pair, with a fixed random seed and deterministic component signs. Each projected channel is scaled using its first and 99th percentiles for RGB display. Because each projection is fitted independently, we compare the spatial organization of regions within each visualization rather than matching colors across cells. As discussed in Section 3.1, these maps reveal where recognizable scene structure emerges across noise levels and network depth: little coherent structure is visible in the noise-dominated region, scene regions begin to take shape in the structure-formation region, and the scene layout becomes more clearly defined in the structure-established region. The onset of recognizable structure occurs in deeper layers at higher noise and in shallower layers at lower noise. Viewed alongside the attention maps, the PCA features help distinguish the emergence of scene structure from the development of cross-frame associations, showing that attention can still favor competing regions even when the scene layout is recognizable. This joint analysis motivates placing correspondence supervision near the structure-formation stage and adapting the supervised layer to the noise timestep.

Attention readout. At the same layer and step, we extract queries and keys after normalization and rotary positional encoding. For the fixed source query, each head's dot-product scores are softmax-normalized over all video keys. We then retain the target-frame slice and average the resulting probabilities over heads. This readout differs from both the block-output features used for PCA and the learned aggregation MLP used during training. Target-frame responses are displayed after normalization by each map's maximum, with color mapping and interpolation used only for visualization. Thus, the color scale indicates relative spatial response within a map, not an absolute probability shared across layers or noise levels.

The denoising axis uses the actual model noise timesteps along the sampled inference trajectory. The noise-dependent shift in structure formation motivates our layer-routing schedule, whose generation performance is evaluated against fixed-layer and alternative schedules in Table 6.

E.2 Geometry Attention

The examples in Fig. 2b inspect VGGT-Ω responses to a fixed first-frame query across target frames. The static example uses the generated indoor video from the video-attention analysis, while the dynamic example uses the bike-packing sequence from TAPVid-DAVIS. These examples provide qualitative correspondence visualizations. The teacher receives the complete clean input sequence: 81 frames for the static example and 69 frames for the dynamic example.

For training-label extraction, we use the fixed global-attention layer 13, head 10 (zero-based indices), as described in Appendix B.1. Source queries are mapped to the 26×15 grid, and target patch responses are resampled and normalized within each target frame. The static query lies on the stair railing; the dynamic query lies on the person's hat. Target frames are selected across the sequence for visualization.

Heatmaps are normalized within each target frame and displayed with contrast adjustment to highlight the spatial concentration of attention responses. The examples illustrate that selected heads can associate observations of static surfaces and moving object parts.

F Limitations and Future Work

GAD improves scene and dynamic consistency across two video backbones by transferring geometric correspondences during training, while preserving the original inference procedure. Its current design suggests three directions for further improvement.

More reliable correspondence targets. Our results demonstrate that attention distillation with geometric correspondence targets improves video consistency. The current implementation extracts these targets from VGGT-Ω attention, making their quality dependent on the teacher's matching ability and spatial resolution. This raises a natural question: can more reliable correspondences further improve the effectiveness of attention distillation? Stronger correspondence estimators or complementary sources of matching evidence could provide more accurate and complete supervision, particularly for ambiguous appearance, occlusion, and moving objects. The correspondence-based formulation of **GAD** provides a basis for exploring such alternatives to the current geometry teacher.

Adaptive supervision placement. Our noise-timestep routing improves geometric metrics over fixed-layer supervision, demonstrating the value of adapting supervision depth to the denoising state. The current layer assignment depends only on the noise timestep and remains fixed throughout training. As the video model learns, the layers that benefit most from correspondence supervision may also change. Conditioning routing jointly on noise timestep and training progress could better match supervision to the model's evolving representations, extending the current mechanism while preserving the original inference procedure.

Extension to autoregressive video generation. **GAD** improves static and dynamic consistency in Wan's bidirectional video models. Autoregressive video generation offers a natural next setting, where preserving scene identity over long rollouts and maintaining coherent object motion remain important challenges. The principle of guiding attention with correspondences also applies to associations between newly generated content and preceding frames. Adapting the distillation objectives to the available temporal context under causal attention could help autoregressive models

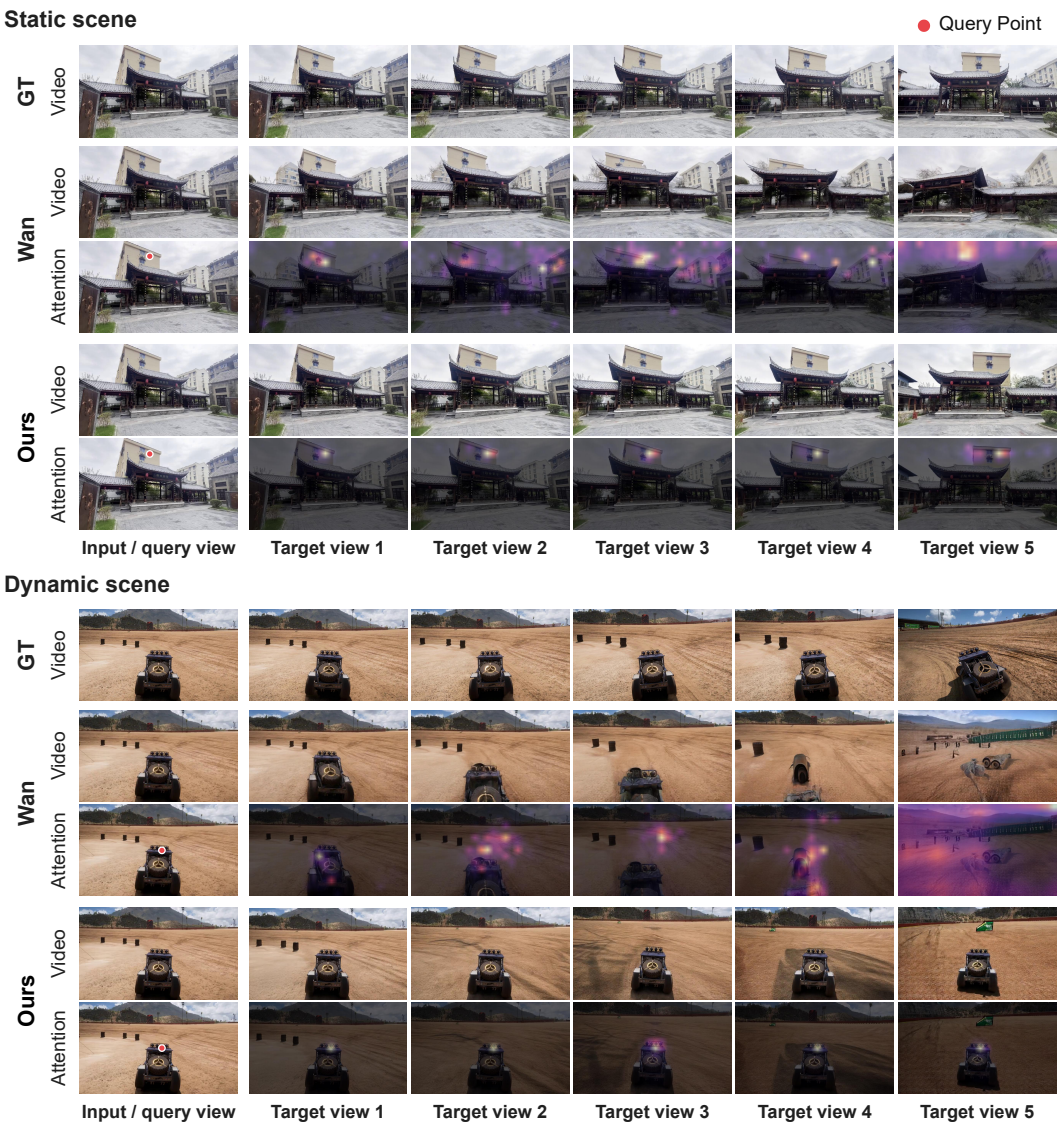


Figure 5: **Video attention and generation results with geometry attention distillation.** Compared with the Wan baseline, **GAD** produces more localized attention responses around corresponding regions and better preserves scene and object structure in static (top) and dynamic (bottom) examples. Red dots mark source queries in the input frame; heatmaps show their attention responses in subsequent target views. GT denotes reference frames.

retain scene structure and moving-object identity as generation proceeds. This offers a promising route to extending the benefits demonstrated here to autoregressive generation.

G More Qualitative Results and Videos

Figs. 5 and 6 provide additional visualizations of video attention with **GAD** and VGGT-Ω correspondence reliability, respectively. Figs. 7 and 8 provide additional qualitative comparisons for long-range static scenes and dynamic scenes, respectively. All examples use first-frame, text, and camera conditioning. Video comparisons are included in a separately submitted supplementary archive.

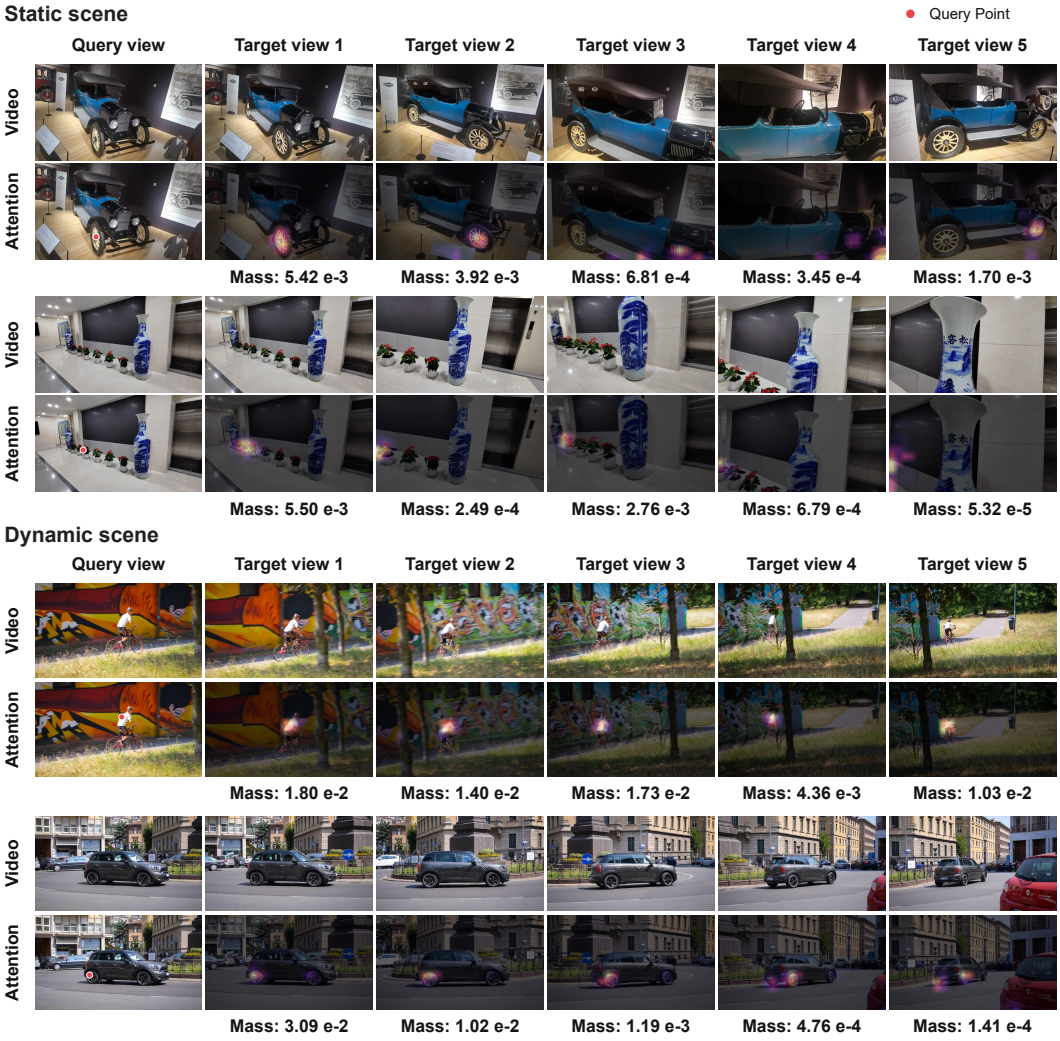


Figure 6: **VGGT-Ω attention and correspondence reliability.** Static (top) and dynamic (bottom) examples show that ambiguous or incorrect responses can have low attention mass despite visible peaks, supporting its use as a reliability cue. Red dots mark source queries; heatmaps show target-view responses, with the total attention mass assigned to each target frame reported below.

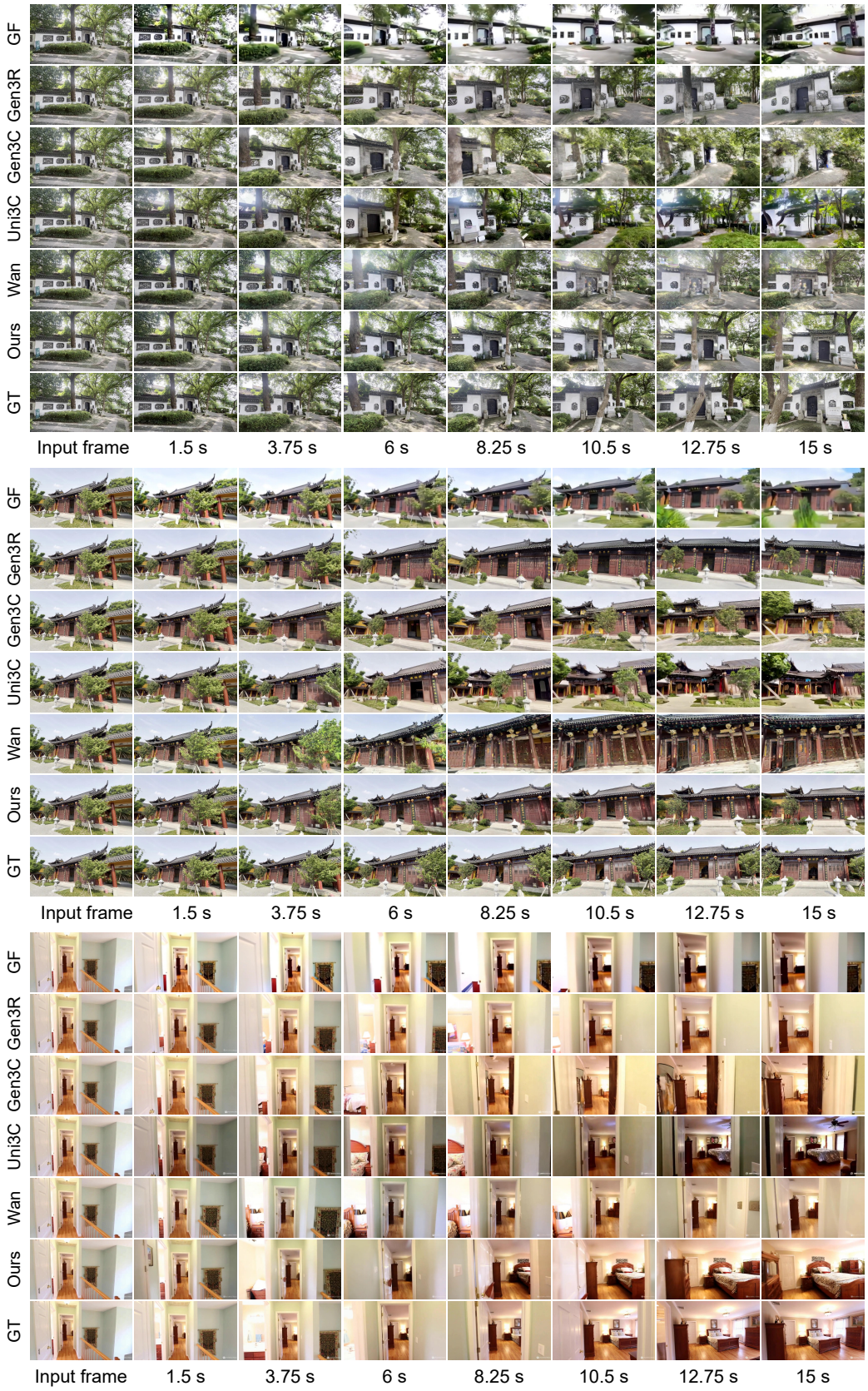


Figure 7: **More 15-second camera-controlled generation results in static scenes.**



Figure 8: **More 5-second camera-controlled generation results in dynamic scenes.**