WARP-AS-HISTORY: GENERALIZABLE CAMERA-CONTROLLED VIDEO GENERATION FROM ONE TRAINING VIDEO

Anonymous authors
Paper under double-blind review

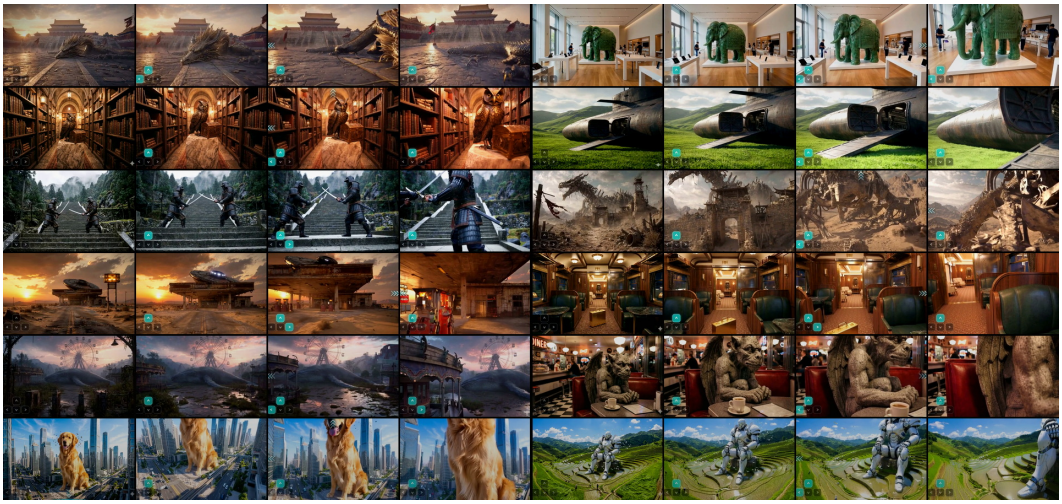

Figure 1: *Warp-as-History* generalizes to unseen scenes and unseen trajectories after finetuning on one video and one camera trajectory.

ABSTRACT

Camera-controlled video generation has made substantial progress, enabling generated videos to follow prescribed viewpoint trajectories. However, existing methods usually learn camera-specific conditioning through camera encoders, control branches, or attention and positional-encoding modifications, which often require post-training on large-scale camera-annotated videos. Training-free alternatives avoid such post-training, but often shift the cost to test-time optimization or extra denoising-time guidance. We propose *Warp-as-History*, a simple interface that turns camera-induced warps into camera-warped pseudo-history with target-frame positional alignment and visible-token selection. Given a target camera trajectory, we construct camera-warped pseudo-history from past observations and feed it through the model's visual-history pathway. Crucially, we align its positional encoding with the target frames being denoised and remove warped-history tokens without valid source observations. Without any training, architectural modification, or test-time optimization, this modification reveals a non-trivial zero-shot capability of a frozen video generation model to follow camera trajectories. Moreover, lightweight offline LoRA finetuning on only one camera-annotated video further improves this capability and generalizes to unseen videos, improving camera adherence, visual quality, and motion dynamics without test-time optimization or target-video adaptation. With systems-level optimizations, the method supports real-time interactive inference on a single H200 GPU. Extensive experiments on diverse datasets confirm the effectiveness of our method.

1 INTRODUCTION

Camera motion is a primary control signal for interactive video generation. It determines not only the viewpoint, but also which regions become visible, how objects move relative to the observer, and

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



Figure 2: From zero-shot history conditioning to one-training-video finetuning. Given the first image and a predefined camera trajectory, the four rows show ground truth, the camera-induced warp, zero-shot Warp-as-History, and one-training-video finetuning. The frozen model already turns the warp into visible camera-follow behavior through the pretrained history interface; one-training-video finetuning further stabilizes this behavior using a single separate training video, without test-time fitting to the shown video.

whether a generated scene can be explored beyond its initial frame. This makes camera control in dynamic video more demanding than static novel-view synthesis: the model must enforce a prescribed camera trajectory while preserving appearance, disoccluding new content, and allowing foreground objects to move independently of the camera.

Recent progress in camera-controlled and interactive video generation has largely been driven by dedicated camera-control mechanisms. Training-based methods inject camera information through camera encoders, control branches, attention or positional-encoding modifications, or related architectural changes, and typically require post-training on camera-annotated videos (He et al., 2024; Li et al., 2025; Zhang et al., 2025; Ren et al., 2025; Yu et al., 2024; Huang et al., 2025a). Training-free methods avoid such post-training, but often enforce the desired trajectory at inference time through test-time optimization, denoising-time guidance, warp-and-repaint procedures, or other sampling-time constraints (Hou and Chen, 2024; You et al., 2024; Liu et al., 2024; Zhou et al., 2025; Song et al., 2025a). At the same time, recent video generation models already exhibit surprisingly rich camera-motion behavior, suggesting that camera control may be latent in video generation models. The challenge is therefore to expose and reliably steer this capability with minimal additional machinery, ideally without collecting large-scale camera-annotated videos, adding camera-specific modules, or imposing extra inference-time objectives.

We approach this question from the perspective of history-conditioned video generation. Many video generation models already condition on visual history to continue a scene from previously observed frames. This history pathway is usually treated as temporal context, but it is also a learned interface for interpreting appearance continuity, motion evidence, and incomplete observations. We ask whether camera-induced geometric evidence can be presented through this existing interface. Specifically, can warped observations induced by a target camera trajectory be used as history evidence, rather than as a dedicated adapter, camera-aware attention or positional encoding, or inference-time guidance objective?

Our answer is yes, when the geometric cue is expressed as history-conditioned evidence. We construct target-frame-aligned, visibility-aware warped observations: source-visible regions provide history evidence, while newly revealed regions are left to the pretrained generator for completion. Warping itself is not new; it appears in prior camera-control, view-synthesis, guidance, and repainting methods. Our distinction is where the warp enters generation: through the visual-history pathway, rather than as a sampling-time constraint or repainting signal.

Given the first frame and a pre-defined camera trajectory, Figure 2 presents examples of: ground truth observation, the camera-induced warp, zero-shot Warp-as-History, and one-training-video finetuning. The warp captures the prescribed camera-induced motion but remains an imperfect geometric cue.

When encoded through the pretrained history-conditioning pathway, this imperfect cue already elicits zero-shot camera-following capability from the frozen model, even in scenes with substantial foreground motion. Although this zero-shot effect is not robust enough to serve as a final method on its own, it reveals a useful latent capability: pretrained video generators can interpret camera-induced geometric evidence when provided as history-conditioned visual evidence.

This observation motivates *Warp-as-History*, a low-resource camera-control framework rather than a new camera-conditioned video generation model. It keeps the control signal visual and geometry-aware, injecting it through the model's native history-conditioning pathway rather than converting it into a hard rendering target or an inference-time guidance objective. We further enhance this capability with lightweight offline LoRA finetuning on only one single camera-annotated video, improving camera adherence, visual quality, and motion dynamics without test-time optimization or target-video adaptation. Section 3 gives the exact construction.

The same view also clarifies the role of finetuning. If zero-shot Warp-as-History already produces measurable camera-follow behavior, lightweight finetuning can be studied as behavior stabilization: it adjusts when the model follows visible warp evidence, when it ignores unreliable warp regions, and when it relies on its generative prior for dynamics and disocclusion. We use one-training-video finetuning as a diagnostic: when a single separate training video improves camera adherence on unrelated test videos, it supports the view that the proposed history interface exposes behavior partially supported by pretraining. In our experiments, one-training-video finetuning makes the zero-shot behavior visibly clearer, as shown in Figure 2. This update is trained offline on a video separate from the test videos; it is not test-time fitting, per-video optimization, or adaptation to the test instance.

Our contributions are:

- We show that pretrained history-conditioned video models contain a weak camera-follow prior, and introduce Warp-as-History to expose it: target camera trajectories are converted into camera-warped pseudo-history with temporal alignment and visibility-aware evidence selection, allowing the frozen model to produce measurable zero-shot camera-following behavior through its native history pathway.
- We demonstrate one-training-video activation: offline LoRA finetuning on a single separate video stabilizes the exposed behavior and generalizes to unseen videos, supporting the view that finetuning amplifies an existing prior rather than learning camera control from scratch.
- Experiments on WorldScore, RE10K, and DAVIS show that Warp-as-History, after finetuning on only one separate video, is competitive with recent state-of-the-art camera-control baselines trained on orders of magnitude more data, with comparable camera adherence and strong visual-quality and consistency metrics.

2 Related Work

Camera-controlled video generation. Camera-controlled video generation has largely followed two routes. Camera-matrix conditioning methods such as CameraCtrl (He et al., 2024), PRoPE (Li et al., 2025), and UCPE (Zhang et al., 2025) inject camera parameters through control branches, camera-aware attention, or positional encodings. Warp- and geometry-conditioned methods such as Gen3C (Ren et al., 2025), ViewCrafter (Yu et al., 2024), and Voyager (Huang et al., 2025a) instead provide target-view evidence through warps, geometry representations, or rendered views. These methods provide strong trajectory control, but often rely on camera-aware modules, geometric representations, or large-scale camera-related training data. Our goal is different: we ask whether an existing history-conditioned video generation model can read camera motion through its native video-history interface.

Training-free camera control. Training-free methods avoid camera-specific post-training and are therefore an important comparison class. Examples include Training-free Camera Control (Hou and Chen, 2024), NVS-Solver (You et al., 2024), video-diffusion-prior novel-view extrapolation (Liu et al., 2024), Latent-Reframe (Zhou et al., 2025), and WorldForge (Song et al., 2025a). Many such methods still pay for control at inference time through test-time optimization, denoising guidance, latent repainting, recursive rollout, or related sampling-time procedures. Warp-as-History instead

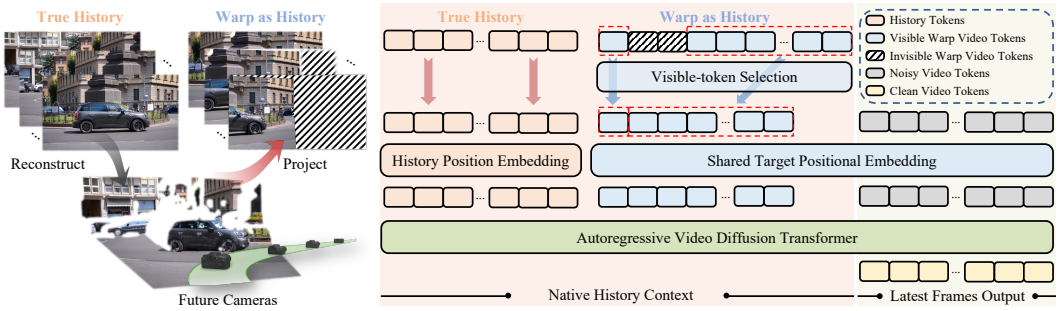

Figure 3: Conditioning a video diffusion model on camera motion. Warp-as-History packs camera-warped pseudo-history into the native history stream, aligns it to target-frame positions, and applies visible-token selection.

constructs camera-induced history once and then follows the native sampler, without per-sample optimization or extra denoising-time guidance.

History-conditioned video generation. History-conditioned video generation uses previous frames as visual context for predicting future frames. Recent methods (Song et al., 2025b; Huang et al., 2025b; Yu et al., 2025; Wu et al., 2025) explore how visual history and retrieved context can improve generation, rollout behavior, and scene consistency. Helios (Yuan et al., 2026) is a recent state-of-the-art history-conditioned backbone with a native history interface. We build on this interface but change its role: history is no longer only temporal context, but an aligned camera-control signal.

3 METHOD

3.1 OVERVIEW: ONE-TRAINING-VIDEO WARP-AS-HISTORY

We first describe the pretrained interface that our method will reuse. Write a video as $X = (x_1, \ldots, x_T)$ and let p be the text prompt. Let $p_\theta(\cdot \mid \cdot)$ denote the conditional sampling distribution induced by the pretrained history-conditioned video generation model and its sampling procedure. For a chunk starting at time t, $X_{<t}$ denotes the available past frames and $X_{t:t+K}$ denotes the future chunk generated by the backbone. The model consumes history through its native construction operator $\mathcal{H}$, which selects, encodes, and temporally packs past visual evidence into a history condition H_t. In history-conditioned video generation, this history may be processed by a transform η_t that corrupts, masks, or drops parts of the past. With this notation, the model predicts future chunks from visual history:

$$\begin{aligned} \bar{X}_{<t} &= \eta_t(X_{<t}), \\ H_t &= \mathcal{H}(\bar{X}_{<t}), \\ X_{t:t+K} &\sim p_\theta(\cdot \mid H_t, p). \end{aligned} \tag{1}$$

This notation highlights the interface we reuse: the model receives processed visual history through H_t and samples the next chunk conditioned on that history and the text prompt.

Warp-as-History is the conditioning interface used by our one-training-video method. It converts a target camera trajectory into camera-warped pseudo-history and feeds it through the native history pathway, with target-frame positional alignment and visible-token selection. Applied directly to the frozen model, the same interface produces the zero-shot behavior discussed in the introduction; we use this behavior as diagnostic evidence that pretrained history-conditioned models can read camera-induced visual evidence from history. The final model uses offline LoRA finetuning on one separate camera-annotated video to stabilize this behavior and improve quality, foreground dynamics, and disocclusion. The resulting weights are shared across test videos; no test-time fitting or per-video optimization is used. Figure 3 illustrates how Warp-as-History conditions the video diffusion model on camera motion.

3.2 WARP-AS-HISTORY CONDITIONING

We first define the conditioning interface that turns camera geometry into visual history evidence. Geometric warps and rendered target views are already common camera-control signals; our use of a warp is not the novelty. The design question is how a history-conditioned video generation model should receive such a signal without a new control branch, a learned camera encoder, or a sampling-time optimization loop. We answer by converting the warp into the same kind of visual evidence the pretrained model already consumes as history, then aligning and filtering that evidence, as summarized in Figure 3.

Camera-warped pseudo-history. Let $C = (c_1, \ldots, c_T)$ be the target camera trajectory for the generated window. A camera-induced warp video W_C renders the available observation under the target camera trajectory, producing an image-space camera-motion cue. We first reconstruct the scene with an off-the-shelf reconstruction model (Wang et al., 2025), then project the reconstruction to each target camera in C to obtain a 2D warp video. Using it as a hard render target would encourage copying warp errors, while learning a new warp-conditioning branch would require extra camera-specific training. We therefore route the warp through the native history interface, corresponding to the warp construction and history-packing path in Figure 3. Let $\tilde{H}_t^C$ denote the camera-warped pseudo-history condition:

$$\tilde{H}_t^C = \mathcal{S}_{M_C}(\mathcal{H}(W_C)). \tag{2}$$

Here M_C is the warp validity mask and $\mathcal{S}_{M_C}$ is visible-token selection applied after native history construction; $\mathcal{H}$ itself does not take a mask input. The construction $\mathcal{H}$ is the same native history construction used by the backbone: the warped frames are patchified, encoded, and packed as ordinary visual history. This condition differs from ordinary history only in how these history tokens are temporally positioned and in which tokens are retained as valid evidence. With ordinary history placement, the warp is presented as past visual context, so the frozen model can apply its pretrained history-to-future continuation behavior to the camera-induced motion. On the frozen model, this also serves as a diagnostic interface: if the base model can continue camera-induced visual motion from history, this condition should produce measurable camera-follow behavior before finetuning. Section 4.3 tests this zero-shot behavior directly.

Target-frame positional alignment. The pretrained continuation behavior separates past history from the current noisy chunk through both the history patchification path and temporal rotary positional embedding (RoPE) positions (Su et al., 2024). If warp tokens keep ordinary history positions, they remain valid history evidence for a motion trace to continue, but the j-th warp frame is still interpreted as past context rather than evidence for the j-th frame being denoised. We therefore keep the warp in the history patchification path, but give each warp latent the same temporal position as the corresponding current noisy latent by assigning it the RoPE index of the target latent at the same frame order, as shown by the shared target positional embedding in Figure 3. Because these tokens are still inserted as history evidence, this remapping does not replace or overwrite the noisy target tokens. Empirically, it is critical: normal denoising remains stable, and Figure 5 shows that the zero-shot output immediately starts to follow the warp after target-frame alignment. The same effect also makes unreliable or invisible warp regions easier to copy, motivating the visible-token selection described next.

Visible-token selection. Camera motion creates newly visible areas that a first-frame warp cannot observe, and imperfect geometry can produce holes, stretched textures, or unreliable regions. Rather than adding a separate conditioning input for the invisible mask, we make invalid evidence resemble the incomplete histories seen during history-conditioned pretraining. Dropping invisible warp tokens from the DiT history stream, shown as visible-token selection in Figure 3, leaves disocclusions to the pretrained completion behavior while still using reliable warped regions as camera-motion evidence. In practice, the warp validity mask is mapped to the latent-token grid, and tokens with insufficient valid support are removed from the history stream. Figure 5 shows the resulting zero-shot jump: after visible-token selection, the frozen model follows the target camera while completing regions that were invisible in the warp. Section 4.3 ablates this chain, and Figure 2 shows the same behavior in the main qualitative example. The behavior is still imperfect: the model may over-copy warped dynamic objects and produce unnatural boundaries near visibility changes, motivating the one-training-video finetuning used by the final model.

This pseudo-history can coexist with ordinary history:

$$\hat{X}_{t:t+K} \sim p_\theta(\cdot \mid H_t, \tilde{H}_t^C, p). \tag{3}$$

Both H_t and $\tilde{H}_t^C$ are inserted through the model's native history stream; no new camera branch or sampling-time guidance loss is introduced. In the first-window setting or ablations without clean history, H_t can be empty. We use this expression only to state the conditioning interface: camera control is represented by populating the existing history pathway with camera-warped pseudo-history. The same conditioning form is used by the frozen-model diagnostic and by the one-video finetuned model.

3.3 One-training-video LoRA finetuning

The final Warp-as-History model keeps the conditioning interface above and finetunes the backbone with a lightweight LoRA update on one separate camera-annotated video. The frozen-model diagnostic reveals camera-follow behavior, but does not solve all aspects of dynamic video generation. In practice, the frozen model can still over-trust the camera-warped pseudo-history: dynamic foreground objects can be copied too rigidly from the warp, and visibility boundaries can remain unnatural. We therefore use lightweight LoRA (Hu et al., 2022) finetuning as the adaptation step.

The goal of finetuning is not to learn a new camera-control branch. Instead, it adjusts how the pretrained history reader balances two sources of evidence: the visible warp tokens, which provide camera-induced motion cues, and the model's generative prior, which is needed for independent dynamics and disocclusion completion. The training loss is the same video-generation objective used by the backbone; only the low-rank update is optimized. The role of the update is to mitigate zero-shot artifacts and reduce the remaining distribution shift from natural histories H_t to camera-warped pseudo-history $\tilde{H}_t^C$.

One-training-video finetuning is treated as a diagnostic for low-resource finetuning. If a single held-out training video improves camera adherence across unrelated test videos, it suggests that the history-conditioning interface is exposing behavior already partially supported by the pretrained model. Which single videos are effective for this finetuning is an empirical question, not part of the method definition; Section 4.4 studies it explicitly. We treat additional training videos as a sensitivity check rather than a main method claim; Section 4.4 and Appendix C report the current multi-video setting.

3.4 Implementation details

All experiments in this paper are built on Helios (Yuan et al., 2026), a real-time long-video generation model with native history conditioning. Unless otherwise stated, the zero-shot experiments use the distilled Helios checkpoint. The main recipe keeps the adaptation localized: aligned warp history and LoRA are inserted only in the first, lowest-resolution Helios stage, while later stages use the native refinement path. The training loss is unchanged from the backbone, and inference uses the standard sampler without test-time optimization or extra denoising-time guidance. In our runs, one-training-video LoRA finetuning uses 1000 iterations and takes about one hour on a single A800 GPU, already producing useful camera-control behavior when mounted on the distilled inference model. Once the warp video is available, inserting Warp-as-History adds less than one second of overhead for generating a 33-frame chunk, since it only packs the camera-warped history condition and does not introduce an optimization loop or extra denoising-time guidance. Appendix C provides the checkpoint, LoRA, and packing details used in the experiments.

4 Analysis and Experiments

The experiments are grouped by the claim they test. First, we compare against prior camera-control systems on the public benchmarks used for static and dynamic video evaluation. Second, we ask whether the frozen model can be induced to follow camera motion at all, and which interface choices make this behavior appear. Third, we analyze how the choice and amount of small finetuning data affect activation of the camera-follow prior.

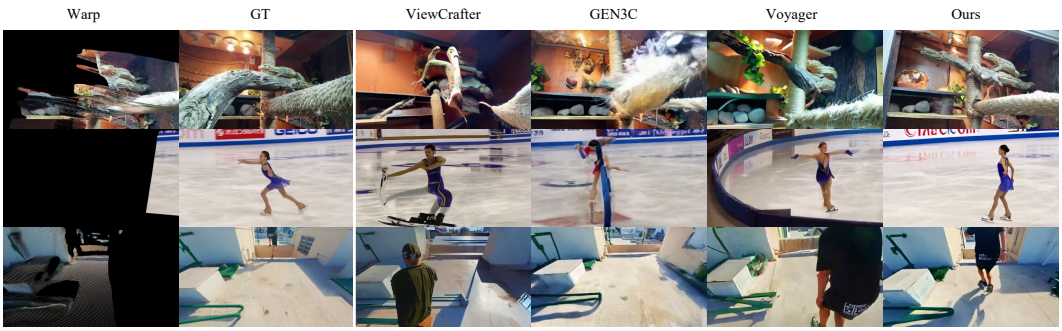

Figure 4: Qualitative comparison with external camera-control methods on in-the-wild videos. Columns show the camera-induced warp, ground truth, ViewCrafter, Gen3C, Voyager, and ours under the same target camera setting.

4.1 Evaluation Datasets

We evaluate on WorldScore (Duan et al., 2025), RealEstate10K (RE10K) (Zhou et al., 2018), and DAVIS (Perazzi et al., 2016). WorldScore provides a static world-generation benchmark, RE10K provides real static scenes with camera motion, and DAVIS provides dynamic videos with foreground motion. Unless otherwise specified, *Ours (one-shot)* denotes a single LoRA finetuning run on the DAVIS *car-roundabout* video, evaluated without per-test-video adaptation. The main text reports compact metrics for camera adherence, quality, consistency, and dynamics, including DOVER (Wu et al., 2023) and VBench (Huang et al., 2024) axes; full metric columns and evaluation details are provided in Appendix C.

4.2 Comparison on Diverse Benchmarks

The external comparisons use our one-shot setting unless otherwise stated, while full report cards are deferred to the appendix.

WorldScore comparison. Table 1 reports representative WorldScore (Duan et al., 2025) results together with our Helios evaluations. Helios-Distilled and Ours rows report native 33-frame full-static WorldScore results.

Table 1: WorldScore results. Helios-Distilled and Ours rows report native 33-frame full-static WorldScore evaluation. Best values are bolded and second-best values are underlined.

Method	Avg.	Camera Control	Object Control	Content Align.	3D Cons.	Photo. Cons.	Style Cons.	Subjective Quality
CogVideoX-I2V (Yang et al., 2024)	62.15	38.27	40.07	36.73	86.21	88.12	83.22	<u>62.44</u>
Voyager (Huang et al., 2025a)	<u>77.62</u>	**85.95**	<u>66.92</u>	**68.92**	81.56	85.99	84.89	**71.09**
FantasyWorld-1.0 (Dai et al., 2025)	**80.45**	<u>81.45</u>	**87.90**	<u>66.94</u>	84.62	**94.07**	86.69	61.46
Helios-Distilled (text-only)	62.42	26.42	42.66	37.75	**92.54**	<u>93.93</u>	<u>90.41</u>	53.21
Ours (zero-shot)	63.26	61.32	33.07	39.92	87.27	88.18	85.67	47.37
Ours (one-shot)	65.64	62.00	32.82	38.60	<u>89.36</u>	90.43	**91.46**	54.83

On WorldScore, Warp-as-History substantially increases camera controllability over the text-only Helios-Distilled baseline. Camera Control rises from 26.42 to 61.32 in the zero-shot setting and 62.00 after one-shot finetuning, corresponding to relative gains of 132.1% and 134.7%, respectively. One-shot finetuning mainly improves visual quality over the zero-shot interface: Subjective Quality increases from 47.37 to 54.83, a 15.7% relative gain, while the average score also improves from 63.26 to 65.64.

Long-video comparison with HyWorldPlay. Table 5 and Figure 6 in Appendix A compare with HyWorldPlay (Sun et al., 2025) on 30-second WorldScore-sampled trajectories. Ours is slightly better on Flicker and Motion Smoothness, while HyWorldPlay is stronger on VBench Overall, Imaging Quality, Dynamic Degree, and scene consistency.

RE10K and DAVIS comparisons. Table 2 reports camera-following metrics on RE10K and DAVIS, and Table 3 reports visual quality, temporal consistency, and dynamics. Within each dataset, all methods use the same evaluation protocol; the exact subset construction is deferred to Appendix C. The external baselines are Gen3C (Ren et al., 2025), Voyager (Huang et al., 2025a), and ViewCrafter (Yu et al., 2024).

Table 2: Camera-following evaluation on DAVIS and RE10K.

Dataset	Method	Training scale	PSNR ↑	SSIM ↑	LPIPS ↓	Vis. LPIPS ↓	R-Err ↓	T-Err ↓
DAVIS	Gen3C	∼90K videos	**16.29**	**0.5267**	**0.3539**	**0.1930**	**2.24**	**0.0663**
	Voyager	∼78K videos → ∼100K clips	14.75	0.3983	0.4431	0.2558	3.05	0.0706
	ViewCrafter	∼85K videos → ∼632K clips	14.72	0.4133	0.3925	0.2308	3.85	0.1031
	Ours (one-shot)	1 video → 4 clips	15.21	0.3976	0.3794	0.2236	2.97	0.0942
RE10K	Gen3C	∼90K videos	**20.10**	**0.7775**	**0.1523**	**0.0828**	**0.62**	**0.0158**
	Voyager	∼78K videos → ∼100K clips	19.03	0.6914	0.2304	0.1268	0.86	0.0322
	ViewCrafter	∼85K videos → ∼632K clips	15.86	0.6765	0.2636	0.2015	0.83	0.0237
	Ours (one-shot)	1 video → 4 clips	17.15	0.6214	0.2343	0.1426	1.28	0.0454

Table 3: Visual quality, temporal consistency, and dynamics on DAVIS and RE10K.

Dataset	Method	FID ↓	FVD ↓	DOVER ↑	Flicker ↑	Motion ↑	Subject ↑	Backgr. ↑	Dynamic ↑	Imaging ↑
DAVIS	Gen3C	72.71	64.98	0.326	0.946	0.977	0.926	0.936	0.701	58.06
	Voyager	90.68	80.28	0.371	**0.948**	0.976	0.925	0.930	0.727	58.61
	ViewCrafter	76.54	65.76	**0.507**	0.933	0.961	0.885	0.903	**0.818**	**67.33**
	Ours (one-shot)	**68.18**	**57.95**	0.448	0.941	**0.977**	**0.941**	**0.940**	0.714	62.20
RE10K	Gen3C	**16.55**	**12.66**	0.298	0.951	0.989	0.917	0.949	**1.000**	58.60
	Voyager	30.65	22.49	0.382	**0.963**	**0.992**	0.955	0.947	0.960	56.80
	ViewCrafter	30.67	23.44	0.436	0.947	0.983	0.881	0.935	**1.000**	63.44
	Ours (one-shot)	22.11	17.92	**0.442**	0.955	0.991	**0.956**	**0.958**	0.980	**65.97**

RE10K is a domain-stress test for our low-resource setting: our one-shot update is finetuned on DAVIS rather than on RE10K, whereas the external baselines use large-scale training data from the same real-estate domain. Even under this mismatch, Ours remains in a comparable camera-following range and obtains stronger visual quality, including the best DOVER, Subject Consistency, Background Consistency, and Imaging scores. On DAVIS, the visual-quality advantage is more pronounced: Ours has the best FID/FVD, Subject Consistency, and Background Consistency, while maintaining camera-following accuracy comparable to the external baselines.

Qualitative comparison. Figure 4 compares our one-training-video setting against the three external baselines on seven in-the-wild videos. Together with the DAVIS quantitative results, these examples show that Warp-as-History preserves scene content and foreground motion more cleanly than prior warp-based baselines, which often expose warp artifacts, blur, or distorted objects.

4.3 Ablating Warp-as-History

Interface ablation. Table 4 compares NoAlign, NoVisDrop, and the full interface in both zero-shot and one-shot regimes. Figure 5 visualizes the frozen-model chain: native warp history, target-frame positional alignment, and visible-token selection.

Table 4: Interface ablations on DAVIS and RE10K. VisLPIPS is visible-region LPIPS; Dyn. and Img. are VBench Dynamic Degree and Imaging Quality.

		DAVIS						RE10K					
Regime	Setting	R-Err ↓	T-Err ↓	VisLPIPS ↓	DOVER ↑	Dyn. ↑	Img. ↑	R-Err ↓	T-Err ↓	VisLPIPS ↓	DOVER ↑	Dyn. ↑	Img. ↑
Zero-shot	NoAlign	7.33	0.1343	0.343	**0.384**	0.740	**58.7**	4.26	0.1109	0.469	0.390	0.980	59.0
	NoVisDrop	4.37	**0.1151**	0.275	0.341	**0.779**	56.1	**2.08**	0.0589	0.249	0.391	**1.000**	62.4
	Full	**3.41**	0.1336	**0.274**	0.364	**0.779**	57.3	2.20	**0.0579**	**0.236**	**0.399**	**1.000**	**62.5**
One-shot	ChFusion	9.10	0.1902	0.289	**0.478**	0.831	**65.4**	5.38	0.2178	0.257	0.422	0.790	66.7
	SeqConcat	5.86	0.1339	0.318	0.346	0.662	56.6	3.20	0.1451	0.331	0.241	0.710	41.3
	NoAlign	7.06	0.1128	0.312	0.471	**0.896**	63.1	4.26	0.0806	0.392	0.445	**1.000**	66.3
	NoVisDrop	**2.94**	0.1111	0.231	0.452	0.753	62.5	1.44	0.0548	0.157	**0.449**	0.990	**67.5**
	Full	2.97	**0.0942**	**0.224**	0.448	0.714	62.2	**1.28**	**0.0454**	**0.143**	0.442	0.980	66.0



Figure 5: Zero-shot interface ablation with the frozen model.

4.4 Small-Data Sensitivity

We treat one-shot source selection and few-shot scaling as diagnostics rather than extra method components. The fixed *car-roundabout* source was chosen before the retrospective sweep; Appendix D shows that it is high-performing but not the best source overall. Appendix C further shows that the clearest gain comes from zero-shot invocation to one-training-video finetuning, while adding more videos gives smaller, non-monotonic changes. Thus the main evidence is not test-set-tuned source selection or many-video scaling, but that one separate video already activates the camera-follow prior across unseen videos.

5 Conclusion

We presented Warp-as-History, a one-training-video camera-control method that routes camera-warped pseudo-history through the native history pathway with target-frame positional alignment and visible-token selection. The same interface reveals zero-shot camera-following behavior in a frozen model, and a lightweight LoRA update on one separate video stabilizes it without test-time optimization or per-video fitting. Across WorldScore, RE10K, and DAVIS, the method is competitive with larger camera-control systems while preserving the pretrained backbone's generative behavior.

AI use statement

In preparing this manuscript, the authors used generative AI tools to assist with organizing and editing manuscript text, converting the manuscript to the official ICLR LaTeX format, and checking consistency among reported results, tables, and source files. All numerical results were supplied by the authors' existing experimental materials; generative AI was not used to produce experimental data or bibliography entries. The authors reviewed and verified all AI-assisted text, tables, claims, and LaTeX changes, and take responsibility for the final content of this work.

References

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

Haoyi Duan, Hong-Xing Yu, Sirui Chen, Li Fei-Fei, and Jiajun Wu. WorldScore: A unified evaluation benchmark for world generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 27713–27724, 2025.

Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. CameraCtrl: Enabling camera control for text-to-video generation. *arXiv preprint arXiv:2404.02101*, 2024.

Chen Hou and Zhibo Chen. Training-free camera control for video generation. *arXiv preprint arXiv:2406.10126*, 2024.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3, 2022.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21807–21818, 2024.

Ruilong Li, Brent Yi, Junchen Liu, Hang Gao, Yi Ma, and Angjoo Kanazawa. Cameras as relative positional encoding. *arXiv preprint arXiv:2507.10496*, 2025.

Kunhao Liu, Ling Shao, and Shijian Lu. Novel view extrapolation with video diffusion priors. *arXiv preprint arXiv:2411.14208*, 2024.

Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 724–732, 2016.

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

Chenxi Song, Yanming Yang, Tong Zhao, Ruibo Li, and Chi Zhang. WorldForge: Unlocking emergent 3d/4d generation in video diffusion model via training-free guidance. *arXiv preprint arXiv:2509.15130*, 2025a.

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

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Wenqiang Sun, Haiyu Zhang, Haoyuan Wang, Junta Wu, Zehan Wang, Zhenwei Wang, Yunhong Wang, Jun Zhang, Tengfei Wang, and Chunchao Guo. Worldplay: Towards long-term geometric consistency for real-time interactive world modeling. *arXiv preprint arXiv:2512.14614*, 2025.

Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. π^3: Permutation-equivariant visual geometry learning. *arXiv preprint arXiv:2507.13347*, 2025.

Haoning Wu, Erli Zhang, Liang Liao, Chaofeng Chen, Jingwen Hou, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. Exploring video quality assessment on user generated contents from aesthetic and technical perspectives. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20144–20154, 2023.

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

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. *arXiv preprint arXiv:2408.06072*, 2024.

Meng You, Zhiyu Zhu, Hui Liu, and Junhui Hou. NVS-Solver: Video diffusion model as zero-shot novel view synthesizer. *arXiv preprint arXiv:2405.15364*, 2024.

Jiwen Yu, Jianhong Bai, Yiran Qin, Quande Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Xihui Liu. Context as memory: Scene-consistent interactive long video generation with memory retrieval. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, pages 1–11, 2025.

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*, 2024.

Shenghai Yuan, Yuanyang Yin, Zongjian Li, Xinwei Huang, Xiao Yang, and Li Yuan. Helios: Real real-time long video generation model. *arXiv preprint arXiv:2603.04379*, 2026.

Cheng Zhang, Boying Li, Meng Wei, Yan-Pei Cao, Camilo Cruz Gambardella, Dinh Phung, and Jianfei Cai. Unified camera positional encoding for controlled video generation. *arXiv preprint arXiv:2512.07237*, 2025.

Min Zhao, Hongzhou Zhu, Bokai Yan, Zihan Zhou, Yimin Chen, Wenqiang Sun, Kaiwen Zheng, Guande He, Xiao Yang, Chongxuan Li, et al. minWM: A full-stack open-source framework for real-time interactive video world models. *arXiv preprint arXiv:2605.30263*, 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.

Zhenghong Zhou, Jie An, and Jiebo Luo. Latent-Reframe: Enabling camera control for video diffusion models without training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12779–12789, 2025.

A HyWorldPlay Evaluation Details

For Table 5, we randomly sample 50 images from WorldScore. For each image, we sample three random camera directions, generate 30-second videos, and evaluate the results with VBench. This protocol is separate from the native 33-frame WorldScore evaluation in Table 1.

Table 5: VBench comparison with HyWorldPlay on 30-second WorldScore-sampled trajectories.

Method	Flicker ↑	Motion ↑	Subject ↑	Background ↑	Dynamic ↑	Imaging ↑	Overall ↑
Ours	**0.984301**	**0.994419**	0.850141	0.908713	0.340000	54.5424	0.770500
HyWorldPlay (Sun et al., 2025)	0.979760	0.990693	**0.862521**	**0.918601**	**0.380000**	**62.8726**	**0.793383**

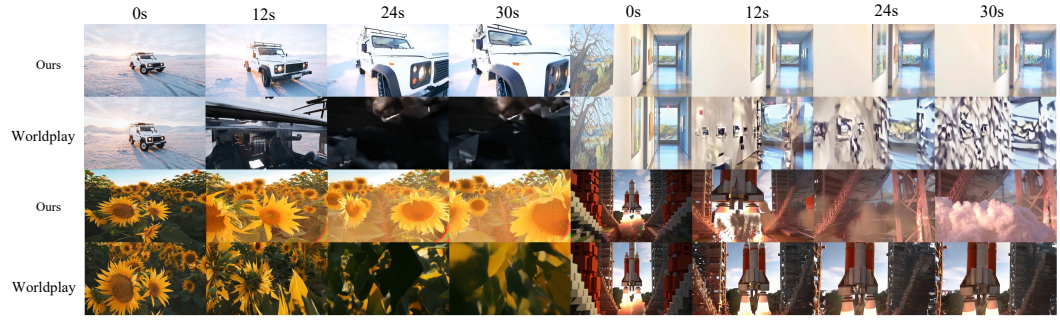

Figure 6: Qualitative comparison with HyWorldPlay on 30-second trajectories sampled from WorldScore images. Frames are shown at 0, 12, 24, and 30 seconds.

B Additional Qualitative Comparisons

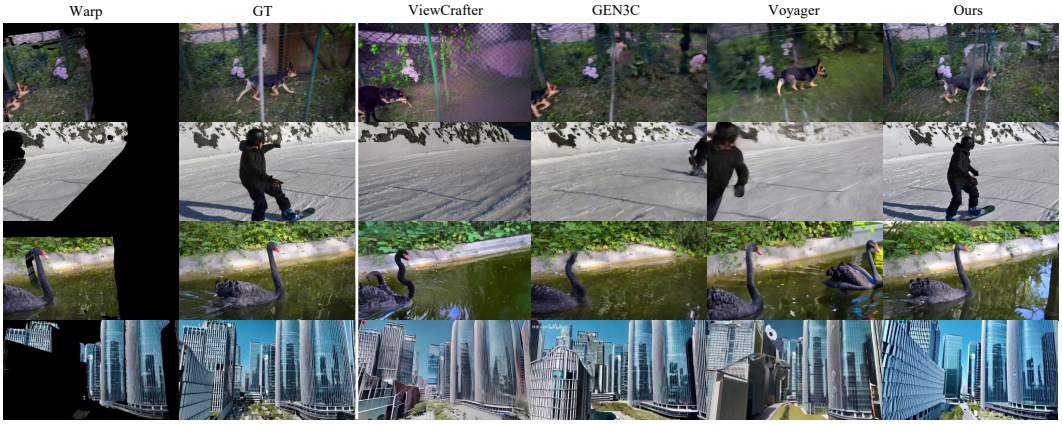

Figure 7: Additional qualitative comparison with external camera-control methods on in-the-wild videos. Columns follow the same layout as Figure 4.

C Interface Ablation Settings and Full Tables

The main text reports compact interface-ablation results in Table 4 and external-baseline results in Tables 2 and 3. Here we provide the exact evaluation settings, full interface-ablation tables, and auxiliary external-baseline metrics omitted from the compact main tables.

Training and sequence selection. Training videos are sampled from sequences disjoint from all evaluation videos. When evaluation is reported on a subset for compute reasons, the subset is randomly selected once before evaluation and then fixed for all compared rows in that table. Thus every method in a table is evaluated on the same videos and target camera trajectories.

Implementation details. For LoRA finetuning, we train LoRA parameters on the Helios-Mid checkpoint and mount the resulting update on the distilled checkpoint at inference time. The distilled inference model uses six denoising steps grouped as $2 + 2 + 2$ across three resolution stages, with the first step of each later stage upsampling noise tokens from the previous stage. Aligned warp history and LoRA are applied only in the first stage; later higher-resolution stages are left unchanged. Within the first stage, LoRA adapters are mounted on the self-attention query, key, value, and output projections with rank 32 and scaling factor $\alpha = 32$. For camera-warped pseudo-history, we do not apply the temporal compression used by ordinary history; the warped video is kept at the original VAE latent resolution before being patchified and inserted into the history stream. The current main

recipe uses a generic camera-control trigger, trains for 1000 iterations on one source video, and uses no test-time optimization.

Shared row definitions. For the interface-ablation tables, all rows use Helios-Distilled as the base inference model. The text-only baseline receives no camera-control condition at inference time, but camera metrics are still computed against the same target camera reference used by the camera-conditioned rows. `Full` denotes warp-as-history with RoPE alignment and visible-token dropping; `NoAlign` removes RoPE alignment; `NoVisDrop` keeps invisible warp tokens; `SeqConcat` and `ChFusion` are non-history conditioning baselines. `ChFusion` is a Gen3C-like channel-fusion baseline: the warped target-view latents are concatenated with the noisy target latents along the channel dimension and fused before denoising, rather than being routed through the history stream. `SeqConcat` is a sequence-concatenation baseline: warp tokens are appended to the denoising-token sequence as ordinary condition tokens, following the same sequence-level path as the noisy target tokens instead of the pretrained history pathway. Zero-shot rows use no LoRA, while one-shot rows use the same stage0-only distilled inference protocol with the corresponding one-shot LoRA update.

WorldScore setting. The WorldScore (Duan et al., 2025) report uses the `static_cc_dev32` subset under native 33-frame evaluation. It contains 32 deterministic samples selected from metadata before evaluation: one sample for each combination of 2 visual styles, 2 scene types, and 8 single-camera motions. All listed rows are evaluated on the same 32 samples.

DAVIS setting. The DAVIS (Perazzi et al., 2016) report uses a 77-video, common-33-frame first-chunk protocol. It contains 77 DAVIS videos with at least 33 common frames; videos without enough frames are excluded. For each video, the evaluated window starts at frame 0 and uses the first 33-frame chunk. Camera-control rows use Pi3X-estimated pseudo-ground-truth camera trajectories from the original DAVIS videos as the target camera condition; if a model requires intrinsics, it uses the paired Pi3X intrinsics for the same 33 frames. Evaluation follows the same DAVIS camera-control protocol for all compared rows.

RE10K setting. The RE10K (Zhou et al., 2018) report uses a fixed 100-sequence subset of the test split for the DAVIS-aligned ablation. It follows the same row definitions as the DAVIS table and reports the available camera, DOVER, and VBench axes. Metrics unavailable in the source report are omitted rather than filled with proxy values.

DAVIS external baseline setting. The external-baseline report compares Ours (one-shot) against Gen3C (Ren et al., 2025), Voyager (Huang et al., 2025a) with the corrected crop, and ViewCrafter (Yu et al., 2024) on DAVIS. The Ours row uses Warp-as-History after one-training-video finetuning, while the external rows use the matched baseline outputs. All rows use the common 33-frame DAVIS evaluation and the same camera, DOVER, and VBench report card.

RE10K external baseline setting. The RE10K external-baseline report compares Ours (one-shot) against Gen3C (Ren et al., 2025), Voyager (Huang et al., 2025a), and ViewCrafter (Yu et al., 2024) on the same 99 RE10K sequences. One sequence is excluded because the corresponding ViewCrafter output is unavailable. Camera metrics use 33 frames with Pi3X frame stride 4, and all rows share the same camera, DOVER, and VBench report card.

Additional-data sensitivity. The main paper centers on one-training-video finetuning. We include the multi-video runs only as a sensitivity check, because increasing the training set in the current range, up to 12 videos, does not produce a clear monotonic improvement over the one-training-video setting. Table 6 reports DAVIS+RE10K mean metrics, while Tables 7 and 8 provide the full per-dataset report cards.

Table 6: Few-shot sensitivity (DAVIS+RE10K mean). The 0 row is zero-shot without LoRA.

Tuning videos	PSNR ↑	LPIPS ↓	R-Err ↓	T-Err ↓	DOVER ↑	Img. ↑
0	13.38	0.4178	2.81	0.0958	0.381	59.93
1	**16.02**	0.3136	2.25	0.0766	0.447	64.47
3	15.89	**0.3115**	1.84	0.0779	0.461	65.09
5	15.04	0.3564	1.90	0.0851	0.471	65.71
7	15.63	0.3163	**1.72**	**0.0751**	0.475	65.76
10	15.49	0.3242	1.99	0.0772	0.473	65.30
12	15.08	0.3483	2.12	0.0807	**0.478**	**65.92**

Table 7: DAVIS few-shot sensitivity full metrics. Rows differ in the number of source videos used for LoRA finetuning under the same small-data recipe.

Videos	PSNR ↑	SSIM ↑	LPIPS ↓	VisLPIPS ↓	R-Err ↓	T-Err ↓	FID ↓	FVD ↓	DOVER ↑	Tech. ↑	Aesth. ↑	Flicker ↑	Motion ↑	Subject ↑	Backgr. ↑	Dyn. ↑	Img. ↑
1	15.14	0.3949	0.3820	0.2252	3.06	0.1076	67.37	58.52	0.453	0.0760	0.9509	0.9397	0.9765	0.9403	0.9395	0.714	62.26
3	15.02	0.3921	0.3737	0.2219	2.54	0.1051	64.00	49.48	0.472	0.0797	0.9617	0.9360	0.9754	0.9279	0.9340	0.831	61.93
5	14.59	0.3780	0.3911	0.2340	2.18	0.1013	64.43	48.70	0.472	0.0799	0.9620	0.9332	0.9734	0.9198	0.9303	0.883	61.93
7	14.66	0.3802	0.3815	0.2262	2.27	0.1055	63.15	49.32	0.478	0.0813	0.9630	0.9354	0.9759	0.9228	0.9310	0.818	61.76
10	14.55	0.3763	0.3907	0.2367	2.53	0.1052	66.07	51.02	0.479	0.0802	0.9635	0.9320	0.9728	0.9162	0.9313	0.883	62.09
12	14.41	0.3723	0.3991	0.2417	2.68	0.1083	65.13	54.68	0.481	0.0809	0.9634	0.9320	0.9740	0.9102	0.9291	0.961	61.94

Table 8: RE10K few-shot sensitivity full metrics. The same LoRA updates as Table 7 are evaluated on RE10K without per-test-video adaptation.

Videos	PSNR ↑	SSIM ↑	LPIPS ↓	VisLPIPS ↓	R-Err ↓	T-Err ↓	FID ↓	FVD ↓	DOVER ↑	Tech. ↑	Aesth. ↑	Flicker ↑	Motion ↑	Subject ↑	Backgr. ↑	Dyn. ↑	Img. ↑
1	16.91	0.6108	0.2453	0.1514	1.45	0.0456	21.96	17.88	0.442	0.0696	0.9639	0.9540	0.9910	0.9571	0.9587	0.980	66.68
3	16.77	0.6098	0.2492	0.1558	1.15	0.0506	34.54	30.37	0.451	0.0714	0.9663	0.9510	0.9900	0.9441	0.9491	0.990	68.24
5	15.50	0.5395	0.3216	0.2068	1.61	0.0688	33.21	30.32	0.471	0.0738	0.9701	0.9474	0.9889	0.9409	0.9485	0.990	69.49
7	16.61	0.5972	0.2511	0.1553	1.17	0.0448	30.40	29.64	0.471	0.0741	0.9703	0.9506	0.9903	0.9484	0.9509	0.990	69.76
10	16.43	0.6004	0.2577	0.1641	1.45	0.0493	24.81	22.43	0.466	0.0730	0.9700	0.9498	0.9904	0.9477	0.9530	1.000	68.51
12	15.75	0.5705	0.2976	0.1952	1.56	0.0531	43.98	38.81	0.474	0.0729	0.9731	0.9472	0.9894	0.9363	0.9460	0.970	69.89

D Supplementary One-Shot Source Diagnostics

Table 9 reports a compact subset of the one-shot source sweep. The full sweep records PSNR, SSIM, LPIPS, visible-region LPIPS, R-Err, T-Err, FID/FVD, DOVER, and VBench axes for each source clip. Tables 10 and 11 report the main reconstruction, camera-following, distribution, and quality axes for the full one-shot source sweep. For each source clip, we also profile the mean/max invisible ratio, mean/max camera rotation, translation-direction angle, foreground centroid motion and area, and source/warp visual quality. The useful criterion is balanced, stable camera-induced parallax rather than raw motion magnitude or the largest invisible region. Effective sources provide clear camera-induced parallax, moderate and stable disocclusion, limited foreground self-motion, and clean source/warp visual quality. The omitted source rows follow the same pattern as the compact table: weak camera-signal clips such as *breakdance* and unstable foreground/warp cases underperform even though the LoRA recipe is unchanged. For example, *breakdance* and *drift-chicane* have very low invisible ratios but almost no camera rotation, so they provide a weak camera-control signal; *motocross-bumps* and *bmx-bumps* contain larger motion and invisible regions, but their unstable motion and disocclusion are not consistently beneficial.

Table 9: One-shot source sensitivity. The LoRA recipe is fixed and only the source video changes. Mean rank aggregates the full DAVIS/RE10K source sweep, with lower values better.

Source	Source diagnostics			Downstream results		
	Inv. mean	Rot. mean	Fg motion	DAVIS LPIPS ↓	RE10K LPIPS ↓	Mean rank ↓
train	0.105	2.82°	0.0046	**0.372**	**0.228**	**2.13**
horsejump-high	0.258	12.14°	0.0098	<u>0.375</u>	0.236	<u>3.06</u>
car-roundabout	0.322	11.15°	0.0024	0.377	<u>0.235</u>	3.94
drift-chicane	0.014	0.03°	0.0103	0.407	0.307	8.38
classic-car	0.252	29.18°	0.0072	0.409	0.316	8.75
surf	0.228	1.45°	0.0174	0.452	0.479	10.88

Table 10: DAVIS one-shot source sweep. Rows differ only in the source video used for one-shot finetuning. VisLPIPS is visible-region LPIPS; Dyn. and Img. are VBench Dynamic Degree and Imaging Quality.

Source	PSNR ↑	SSIM ↑	LPIPS ↓	VisLPIPS ↓	R-Err ↓	T-Err ↓	FID ↓	FVD ↓	DOVER ↑	Dyn. ↑	Img. ↑
bmx-bumps	14.40	0.3668	0.4003	0.2378	2.89	0.0976	67.97	53.48	0.431	0.857	60.26
breakdance	14.62	0.3839	0.4145	0.2501	3.54	0.1110	71.26	57.76	0.433	0.909	59.90
car-roundabout	15.21	0.3980	0.3772	0.2224	3.03	0.1033	68.69	56.02	0.453	0.727	62.41
classic-car	14.64	0.3707	0.4085	0.2481	2.76	0.0905	73.64	64.07	0.415	0.779	59.86
dance-twirl	15.11	0.3970	0.3786	0.2264	2.87	0.1042	64.66	49.25	0.456	0.831	60.65
drift-chicane	15.04	0.3942	0.4069	0.2461	3.51	0.1038	68.13	56.87	0.427	0.740	60.16
horsejump-high	15.38	0.4073	0.3746	0.2225	2.76	0.0933	65.42	53.18	0.440	0.766	60.95
kite-surf	15.17	0.3959	0.3992	0.2389	3.06	0.1032	70.00	60.24	0.418	0.740	60.00
motocross-bumps	15.10	0.3936	0.3884	0.2311	2.60	0.1054	67.37	60.31	0.443	0.740	60.60
parkour	15.04	0.3870	0.3864	0.2333	2.91	0.1045	64.98	55.89	0.446	0.766	62.33
surf	14.08	0.3516	0.4515	0.2842	3.97	0.0998	71.38	61.29	0.413	0.805	59.28
train	15.31	0.4034	0.3717	0.2205	2.72	0.0968	68.59	55.81	0.447	0.753	61.39

Table 11: RE10K one-shot source sweep. Rows use the same source-video LoRA updates as Table 10, evaluated on RE10K without per-test-video adaptation.

Source	PSNR ↑	SSIM ↑	LPIPS ↓	VisLPIPS ↓	R-Err ↓	T-Err ↓	FID ↓	FVD ↓	DOVER ↑	Dyn. ↑	Img. ↑
bmx-bumps	15.81	0.5658	0.2872	0.1828	1.43	0.0541	22.55	20.94	0.451	1.000	67.92
breakdance	15.78	0.5752	0.3024	0.1992	1.72	0.0610	50.22	46.21	0.407	1.000	63.93
car-roundabout	17.12	0.6187	0.2353	0.1426	1.32	0.0456	21.60	18.16	0.442	0.980	66.37
classic-car	15.48	0.5621	0.3162	0.2088	1.38	0.0613	23.09	19.71	0.433	1.000	65.73
dance-twirl	16.78	0.6074	0.2514	0.1582	1.31	0.0478	36.59	30.23	0.431	0.990	66.23
drift-chicane	15.86	0.5755	0.3069	0.2015	1.90	0.0554	24.47	21.29	0.417	0.990	63.62
horsejump-high	17.13	0.6274	0.2356	0.1466	1.29	0.0440	28.03	25.72	0.418	0.980	64.65
kite-surf	16.23	0.5883	0.2848	0.1827	1.65	0.0535	22.30	18.52	0.423	0.990	64.31
motocross-bumps	17.04	0.6220	0.2425	0.1485	1.28	0.0510	29.80	24.02	0.425	0.990	64.04
parkour	16.33	0.5858	0.2760	0.1790	1.56	0.0505	21.23	18.97	0.434	0.990	65.94
surf	13.31	0.4960	0.4788	0.3340	4.26	0.0693	23.50	23.34	0.416	1.000	61.51
train	17.22	0.6271	0.2282	0.1396	1.27	0.0481	21.21	17.45	0.442	0.980	66.49

Table 12: Additional DAVIS interface-ablation metrics from the same source as Table 4.

Regime	Setting	PSNR ↑	SSIM ↑	LPIPS ↓	Vis. LPIPS ↓	R-Err ↓	T-Err ↓	FID ↓	FVD ↓	DOVER ↑	Flicker ↑	Motion ↑	Subject ↑	Backgr. ↑	Dynamic ↑	Imaging ↑
Text-only	Base	14.38	0.3627	0.4199	0.2626	7.96	0.1968	71.55	80.87	0.463	0.989	0.994	0.979	0.971	0.091	65.21
Zero-shot	NoAlign	12.03	0.2752	0.5439	0.3430	7.33	0.1343	94.86	83.98	0.384	0.948	0.973	0.912	0.929	0.740	58.75
	NoVisDrop	12.31	0.3162	0.5013	0.2749	4.37	0.1151	89.69	72.28	0.341	0.957	0.978	0.901	0.920	0.779	56.05
	Full	13.28	0.3441	0.4629	0.2735	3.41	0.1336	83.37	72.16	0.364	0.941	0.969	0.923	0.931	0.779	57.33
One-shot	ChFusion	13.59	0.3150	0.4600	0.2891	9.10	0.1902	72.54	67.15	0.478	0.935	0.972	0.942	0.938	0.831	65.44
	SeqConcat	12.42	0.3177	0.5019	0.3178	5.86	0.1339	100.52	90.31	0.346	0.942	0.969	0.919	0.937	0.662	56.59
	NoAlign	13.47	0.3190	0.4826	0.3115	7.06	0.1128	69.62	64.32	0.471	0.936	0.975	0.927	0.932	0.896	63.08
	NoVisDrop	15.02	0.3924	0.3863	0.2312	2.94	0.1111	68.58	56.87	0.452	0.940	0.976	0.937	0.935	0.753	62.54
	Full	15.21	0.3976	0.3794	0.2236	2.97	0.0942	68.18	57.95	0.448	0.941	0.977	0.941	0.940	0.714	62.20

Table 13: Additional RE10K interface-ablation metrics from the same source as Table 4.

Regime	Setting	PSNR ↑	SSIM ↑	LPIPS ↓	Vis. LPIPS ↓	R-Err ↓	T-Err ↓	FID ↓	FVD ↓	DOVER ↑	Flicker ↑	Motion ↑	Subject ↑	Backgr. ↑	Dynamic ↑	Imaging ↑
Zero-shot	NoAlign	9.46	0.3258	0.6834	0.4691	4.26	0.1109	85.95	68.27	0.390	0.964	0.991	0.933	0.952	0.980	58.96
	NoVisDrop	13.09	0.5048	0.3906	0.2493	2.08	0.0589	36.81	29.79	0.391	0.960	0.988	0.953	0.955	1.000	62.40
	Full	13.47	0.5220	0.3726	0.2358	2.20	0.0579	35.70	28.30	0.399	0.957	0.987	0.954	0.953	1.000	62.53
One-shot	ChFusion	14.52	0.5247	0.3822	0.2566	5.38	0.2178	29.50	30.78	0.422	0.961	0.987	0.968	0.961	0.790	66.71
	SeqConcat	12.24	0.5497	0.5289	0.3307	3.20	0.1451	94.80	82.01	0.241	0.973	0.991	0.911	0.944	0.710	41.25
	NoAlign	12.37	0.4365	0.5585	0.3924	4.26	0.0806	33.89	35.58	0.445	0.950	0.989	0.936	0.948	1.000	66.28
	NoVisDrop	16.78	0.6082	0.2523	0.1573	1.44	0.0548	21.61	18.33	0.449	0.953	0.990	0.956	0.957	0.990	67.50
	Full	17.15	0.6214	0.2343	0.1426	1.28	0.0454	22.11	17.92	0.442	0.955	0.991	0.956	0.958	0.980	65.97

Table 14: Additional external-baseline metrics omitted from Tables 2 and 3.

Dataset	Method	Visible %	DOVER tech. ↑	DOVER aesth. ↑
DAVIS	Ours	66.0	0.0757	0.9479
	Gen3C	66.0	0.0513	0.8954
	Voyager	66.0	0.0649	0.9159
	ViewCrafter	66.0	0.0799	0.9571
RE10K	Ours	67.0	0.0697	0.9638
	Gen3C	67.0	0.0409	0.9282
	Voyager	67.0	0.0626	0.9446
	ViewCrafter	67.0	0.0607	0.9658

E Runtime Analysis

Warp-as-History adds camera-warped pseudo-history to the native history stream, so it inevitably increases the number of tokens processed by the transformer. We measure runtime for generating one 33-frame chunk on a single NVIDIA A800 GPU. These measurements profile the standard, higher-quality inference recipe rather than the optimized real-time configuration described in Appendix G. Table 15 profiles this cost relative to the original image-to-video sampler under two visibility regimes. The visible-token percentage denotes the fraction of warped-history tokens retained after visibility filtering; invisible-token dropping removes tokens without valid source observations before they enter the transformer.

The main overhead comes from the transformer/sampling stage rather than from geometry or warp preparation. With 86% visible tokens, transformer/sampling increases by 7.59s and accounts for almost all of the 7.81s end-to-end increase. When only 47% of the warped tokens are visible, the transformer/sampling overhead drops to 3.38s, and the end-to-end overhead drops to 4.62s. By contrast, camera rendering, warp VAE encoding, and warp/mask preparation together contribute only about 1–2 seconds. This confirms that Pi3X-based geometry estimation and warp preparation are not the bottleneck in this setting; the bottleneck is the longer transformer sequence induced by using warps as history. The same trend also supports invisible-token dropping: discarding unobserved warp tokens not only avoids conditioning on invalid evidence, but also reduces the sequence length that dominates runtime.

Table 15: Runtime profile for generating one 33-frame chunk with Warp-as-History under different visible-token ratios on a single NVIDIA A800 GPU. Times are reported in seconds; `original` is the native image-to-video sampler and `ours` adds camera-warped pseudo-history.

Setting	Part	Original	Ours	Increase
Warp with 86% visible tokens	Camera render	0.00s	1.02s	+1.02s
Warp with 86% visible tokens	Warp VAE encode	0.00s	0.47s	+0.47s
Warp with 86% visible tokens	Transformer / sampling	14.34s	21.94s	+7.59s
Warp with 86% visible tokens	End-to-end	15.83s	23.63s	+7.81s
Warp with 47% visible tokens	Camera render, assumed	0.00s	1.02s	+1.02s
Warp with 47% visible tokens	Warp VAE encode	0.00s	0.47s	+0.47s
Warp with 47% visible tokens	Warp/mask prepare	0.00s	0.19s	+0.19s
Warp with 47% visible tokens	Transformer / sampling	14.37s	17.75s	+3.38s
Warp with 47% visible tokens	End-to-end	15.78s	20.40s	+4.62s

F Additional Cross-Backbone Results

F.1 Transfer to a second autoregressive backbone

To test whether Warp-as-History is specific to Helios, we additionally implement the same interface in the minWM framework (Zhao et al., 2026), using its HunyuanVideo-1.5 teacher-forcing autoregressive text-and-image-to-video checkpoint whose clean-history/noisy-target interface permits independent temporal-position assignment. Table 16 reports a matched comparison on this second backbone. Warp-as-History transfers without architectural changes: the zero-shot variant substantially improves all three reported measures over both the text-only checkpoint and its camera-trained Action2V/PRoPE counterpart, and one-video LoRA further improves camera following and reconstruction quality. This result supports cross-backbone transfer within the intended model class; it does not claim compatibility with arbitrary video generators.

Table 16: Cross-backbone evaluation using the HunyuanVideo-1.5 checkpoint provided by minWM. All rows use the same autoregressive backbone family and evaluation protocol.

Method	Camera data	R-Err ↓	T-Err ↓	PSNR ↑
HunyuanVideo-1.5 text-only	0	3.426°	0.0930	16.07
Action2V/PRoPE	$\sim$4.6K videos	3.136°	0.0457	16.04
Ours (zero-shot)	0	0.903°	0.0401	18.08
Ours (one-shot)	1 video	$\mathbf{0.750^\circ}$	**0.0325**	**18.42**

The quality diagnostics in Table 17 further indicate that the camera-control gain is not produced by a broad perceptual-quality collapse. One-shot Warp-as-History remains close to the text-only and Action2V/PRoPE checkpoints on VBench Aesthetic Quality and Imaging Quality, while increasing Dynamic Degree from 0.6250 for Action2V/PRoPE to 0.7917. These aggregate metrics do not replace visual inspection, and we interpret the result as a quality–control–data–latency trade-off rather than uniform superiority on every quality axis.

Table 17: Perceptual-quality diagnostics on the HunyuanVideo-1.5 comparison. A dash denotes an unavailable value.

Method	Aesthetic ↑	Imaging ↑	Dynamic ↑
HunyuanVideo-1.5 text-only	**0.5161**	**62.34**	–
Action2V/PRoPE	0.5133	62.29	0.6250
Ours (one-shot)	0.5141	62.22	**0.7917**

G Optimized Real-Time Inference

Beyond the standard recipe profiled in Appendix E, we implement an optimized real-time configuration for interactive camera control on a single H200 GPU. The system initializes the prompt, first frame, and model state once, then keeps the autoregressive state resident across user inputs and consumes one new camera-pose or pre-rendered-warp chunk per interaction. This avoids rebuilding the generation state between camera commands while retaining the same incremental camera-control interface.

The optimized configuration uses a matching efficient LoRA trained with `patch_mid` warp-history downsampling, one denoising step at each of the three pyramid stages, no first-chunk amplification, a fast target-fill camera renderer, a 130,000-pixel geometry-preprocessing limit, and two mesh samples per axis. It also preloads the pipeline, uses optimized attention kernels when available, disables progress-display overhead, and employs a lightweight video decoder for interactive output. Together, these systems optimizations enable real-time interactive inference while preserving the method's stateful camera-control interface. The standard preset remains available when output quality is preferred over latency; consequently, the A800 timings in Table 15 should not be read as the latency of this optimized deployment.

Table 18 reports peak GPU memory measured on a clean single-GPU setup with Helios-Distilled in BF16 at 384×640 resolution and 33 frames, without CPU or group offloading. Pre-rendered-warp inference has essentially the same memory footprint as the original Helios image-to-video pipeline. Online camera-pose inference uses more memory because the reconstruction and camera-warp components remain resident alongside Helios.

Table 18: Peak GPU memory for the evaluated configurations. All measurements use a clean single-GPU process; the training row reports one Helios-Mid LoRA step.

Configuration	Peak VRAM ↓
Original Helios I2V	46.1 GB
Ours, pre-rendered warp input	46.1 GB
Ours, online camera-pose input	53.6 GB
Helios-Mid LoRA training (one step)	48.7 GB

H Limitations

Warp-as-History depends on the quality and cost of the warp construction step. In our implementation, the warp is produced online by reconstructing the observed scene with an external reconstruction model and projecting the reconstruction to the target future cameras. This avoids training a camera-specific control branch, but it adds preprocessing cost and inherits the reconstruction model's failure modes, including errors in geometry, visibility, and disoccluded regions. The history interface itself is also not free: even though we insert warped history only in the first Helios stage, where the spatial resolution is low and inference uses only two denoising steps, the additional history tokens still increase runtime relative to the native image-to-video sampler. Finally, the method is an invocation interface rather than a new video generator. Its generalization is therefore bounded by the pretrained backbone's existing ability to interpret visual history, preserve dynamics, and complete unobserved content; when the base model lacks these capabilities, lightweight LoRA can stabilize the behavior but cannot fully remove the limitation.

I Broader Impacts

This work studies camera control for pretrained video generation models. It may be useful for creative editing, virtual cinematography, simulation, and controllable scene visualization. Like other video generation and editing methods, it could also be misused to create misleading videos or to modify private or sensitive footage without consent. Any release or deployment should therefore follow the safety policies of the underlying video model, clearly label generated or edited content, and respect

dataset licenses and subject consent. This paper does not introduce a new dataset or a deployed user-facing system.