GRAFT: GEOMETRY-FIRST VIDEO RETAKING VIA APPEARANCE-FIRST TRAINING

Anonymous authors
Paper under double-blind review

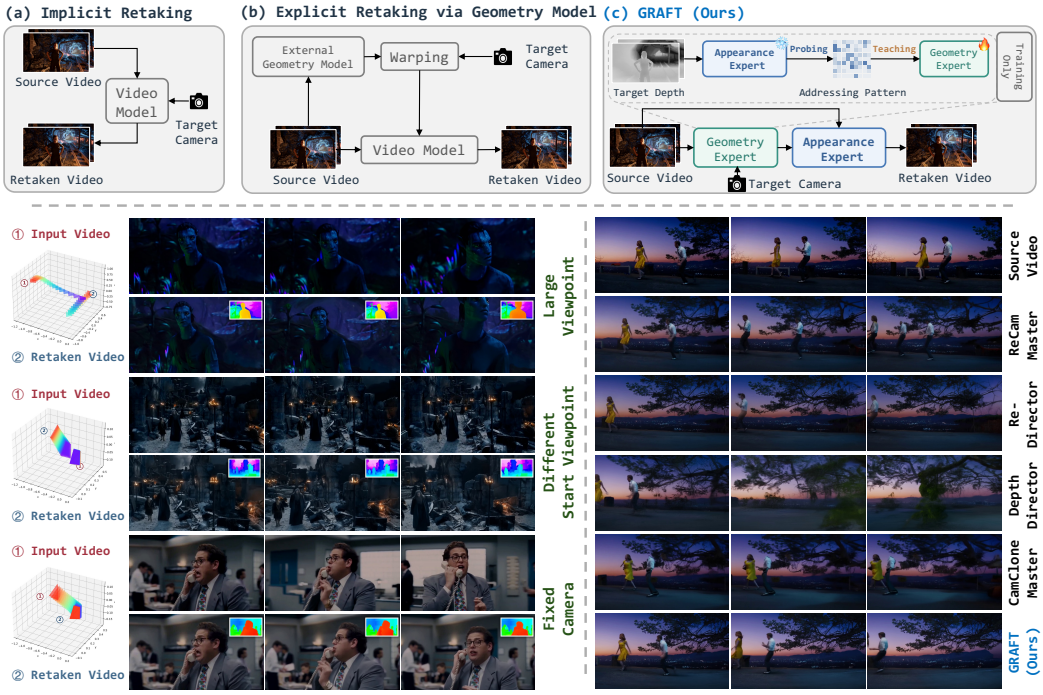


Figure 1: **Overview and qualitative results of GRAFT.** Addressing Distillation is used only during training; at inference, the geometry expert provides the target-view depth required by the appearance expert, without ground-truth target depth. We also show qualitative comparisons. The upper-right insets in the retaken frames of the left-hand examples show the predicted depth.

ABSTRACT

Video retaking aims to reproduce an observed dynamic scene along a new camera trajectory while preserving its content and motion. A central challenge is locating relevant source evidence across frames as viewpoint changes, motion, and occlusion alter visibility. We introduce **GRAFT**, a video retaking framework built around **Addressing Distillation**, which transfers source-addressing behavior from appearance synthesis to target-view geometry prediction. Our key insight is that appearance synthesis conditioned on known target geometry provides a task for learning where to retrieve source evidence. Through **Appearance-First Training**, we first train an appearance expert to synthesize target RGB video from source observations and ground-truth target depth. We then probe the frozen expert at the pure-noise endpoint under the same depth conditioning, extracting its target-to-source attention distributions as soft addressing patterns. These distributions supervise a geometry expert's source addressing alongside its depth generation objective, without constructing explicit geometric correspondence labels. At inference, **Geometry-First Retaking** reverses the order: the geometry expert predicts target-view metric depth, which guides the same appearance expert in synthesizing the retaken video. Evaluations on dynamic-scene and human-centric video retaking demonstrate improved camera trajectory following, geometric consistency, and subject motion preservation.

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.

1 INTRODUCTION

Video retaking Bai et al. (2025); Yu et al. (2025); Chen et al. (2026); Bai et al. (2026); Hu et al. (2025); Luo et al. (2025); Park et al. (2026) aims to reproduce an observed dynamic scene along a new camera trajectory while preserving its content and motion. Visual realism alone does not ensure that the generated video follows the requested camera trajectory or preserves the original motion. Faithful retaking therefore requires locating relevant evidence in the source video. As viewpoint, motion, and occlusion change what is visible, this evidence may be distributed across different source regions and frames. We refer to this process as *target-to-source addressing*.

Existing retaking methods differ in how they incorporate geometry into generation. Camera-conditioned methods such as ReCamMaster Bai et al. (2025) and ReDirector Park et al. (2026) directly synthesize target RGB videos from source observations and camera conditions, leaving target-view geometry implicit. Geometry-guided methods instead provide explicit spatial cues: TrajectoryCrafter Yu et al. (2025) warps source observations into the target view, while DepthDirector Chen et al. (2026) uses warped depth as guidance. These cues place observed content in the target view, but their coverage and accuracy depend on visibility and reconstruction quality. Beyond such input-level guidance, recent work uses geometric relationships to supervise attention or organize source retrieval, as in CAMEO Kwon et al. (2026), MVTrack4Gen Lee et al. (2026), PE-Field 4D Bai et al. (2026), and World in World Song et al. (2026). This motivates us to explore a complementary source of addressing supervision: target-to-source attention learned through appearance synthesis, without constructing explicit geometric correspondence labels.

We focus on target-view depth prediction as an intermediate step that establishes scene structure for subsequent RGB synthesis. As subjects move and become occluded, the source evidence relevant to preserving the original scene dynamics can be distributed across different frames. Depth generation supervision specifies *what to predict*, but does not explicitly specify *where to look*. Our analysis shows that a predictor trained with depth generation supervision alone can rely excessively on the initial source frame and fail to preserve subsequent subject motion. We therefore supervise source addressing alongside depth generation to guide how the geometry predictor uses the source video.

Our key insight is that appearance synthesis under known target geometry provides a task for learning this addressing behavior. Ground-truth (GT) target depth specifies the target layout, while the source video provides the appearance and scene content to be preserved. Training an appearance expert to synthesize the target video teaches it to use source evidence for the given target layout. Its target-to-source attention distributions provide soft addressing patterns over source regions and frames. The geometry expert must infer the target structure that is given to the appearance expert. We introduce **Addressing Distillation** to transfer addressing learned under this known structure to the task of predicting it, without constructing explicit geometric correspondence labels.

We implement this idea in **GRAFT** through **Appearance-First Training**. We train the appearance expert with GT target depth, then freeze it as the teacher for geometry training. To extract addressing supervision, we probe the teacher at the pure-noise endpoint with depth conditioning. This removes residual target RGB content from the noisy input, so the teacher's addressing distributions are obtained from source observations and the known target layout rather than partially revealed target appearance. We evaluate the geometry expert with identical noise and match its target-to-source addressing distributions to the teacher's alongside the depth generation objective. The geometry expert learns to predict target depth and to allocate attention across source evidence.

At inference, GRAFT follows **Geometry-First Retaking**. The geometry expert predicts target-view metric depth from the source video, camera conditions, and an initial depth reference. This prediction conditions the same appearance expert to synthesize the retaken RGB video, without GT target depth or addressing distillation. The appearance expert therefore serves as both the training-time teacher and the final RGB generator: it teaches source addressing under known target geometry, while the geometry expert learns to provide the geometry needed for retaking.

Our contributions are threefold. **(i)** We introduce **Addressing Distillation**, which probes an appearance teacher conditioned on GT target depth and transfers its soft target-to-source addressing distributions to a geometry predictor, without constructing explicit geometric correspondence labels. **(ii)** We develop **GRAFT**, a video retaking framework that combines **Appearance-First Training** with **Geometry-First Retaking**: the appearance expert provides addressing supervision during training, and the geometry expert supplies predicted depth to the same appearance expert during inference.

(iii) Experiments on dynamic-scene and human-centric video retaking demonstrate improvements in camera trajectory following, geometric consistency, and subject motion preservation.

2 Related Work

Camera-Controlled Video Retaking. Video retaking aims to change the camera trajectory of an observed video while preserving its content and dynamics. ReCamMaster Bai et al. (2025) conditions RGB generation on source observations and camera motion, while ReDirector Park et al. (2026) introduces rotary camera encoding. CameraAnything Li et al. (2026b) extends camera control to intrinsic and extrinsic parameters through Plücker ray conditioning and resolution-aware positional encoding. Geometry-guided methods such as GEN3C Ren et al. (2025) and TrajectoryCrafter Yu et al. (2025) condition generation on scene renderings, while DepthDirector Chen et al. (2026) uses warped depth. Reshoot-Anything Paliwal et al. (2026) constructs pseudo multi-view training triplets from monocular videos and uses point-cloud-rendered anchors at inference. SierpinskiCam Wizadwongsa et al. (2026) augments geometric guidance with Sierpinski texture cues in newly revealed regions and uses negative RoPE indices for source-video conditioning. MoCam Liu et al. (2026) switches from rendered geometric guidance to source appearance within RGB denoising. GRAFT instead predicts target-view metric depth before RGB synthesis and trains the geometry predictor through addressing distillation from the appearance expert.

Correspondence Learning and Source Retrieval. Correspondence learning provides supervision for improving consistency in generation. Track4Gen Jeong et al. (2025) jointly trains point tracking and video generation to strengthen temporal correspondence features. CAMEO Kwon et al. (2026) directly supervises multi-view attention with geometric correspondences, while MVTrack4Gen Lee et al. (2026) uses multi-view tracks to supervise both point prediction and attention across views and time. Complementary approaches use geometry to organize source retrieval: PE-Field 4D Bai et al. (2026) equips source tokens with projected target-view positional encodings, and World in World Song et al. (2026) uses point identities and geometry to route attention during inference. These approaches derive correspondence supervision or retrieval guidance from tracks and geometric relationships. AD instead learns addressing supervision from an appearance synthesis task, without constructing physical point matches for the supervised target tokens. Although MoAI Kwak et al. (2026) targets image novel-view synthesis, its use of cross-modal attention instillation provides a relevant point of comparison and warrants detailed discussion.The motivation differs: MoAI shares attention to align jointly generated RGB images and pointmaps, whereas AD teaches a geometry predictor where to retrieve source evidence. We construct this supervision through an appearance task conditioned on GT target depth, transferring its learned addressing to a geometry expert that must infer that depth. The predicted depth then guides RGB synthesis, establishing a sequential dependency between the outputs.

3 Method

3.1 Task Definition and Overview

Given a source video X and camera conditions $\mathcal{C}$ containing the source and target trajectories, video retaking aims to synthesize a target video Y that preserves the original scene and motion under the requested viewpoint. We denote the target-view metric depth sequence by D. The geometry expert additionally receives a target-view first-frame depth reference a, obtained through external geometry estimation Kong et al. (2026); Wang et al. (2026b). This reference provides a geometric cue for depth prediction; its construction and conditioning mechanism are described in Appendix B.2.

GRAFT centers on **Addressing Distillation (AD)**, which transfers source-addressing distributions from an appearance expert with parameters ϕ to a geometry expert with parameters θ. Under **Appearance-First Training**, we first train the appearance expert to synthesize target RGB video conditioned on source observations and GT target depth. We then freeze it and train the geometry expert with addressing distillation alongside depth generation supervision. At inference, **Geometry-First Retaking** predicts target-view depth with the geometry expert and supplies it to the same appearance expert for RGB synthesis.

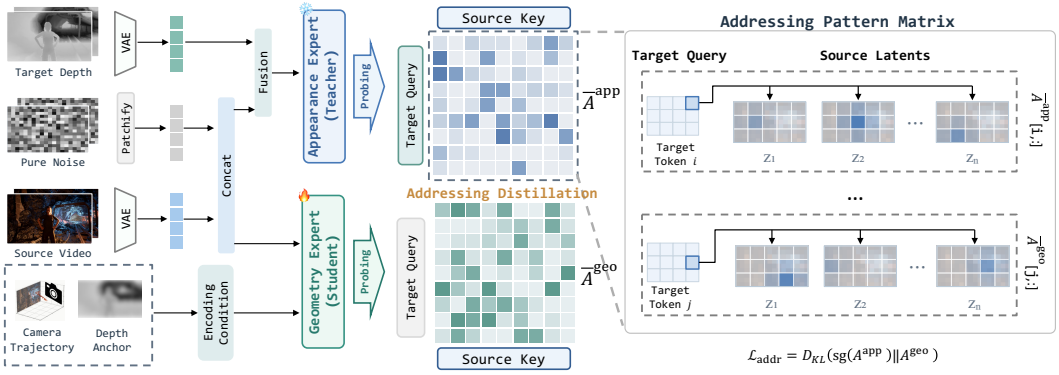


Figure 2: We probe both experts at the pure-noise endpoint using the same target noise. The frozen appearance teacher receives GT target depth paired with its RGB training target. The rightmost panel illustrates the teacher's addressing matrix, which is distilled into the geometry student.

Both experts share the frozen weights of a pretrained Wan2.2 video diffusion backbone Wan et al. (2025), based on a diffusion transformer Peebles & Xie (2023), and use separate low-rank adaptation (LoRA) adapters Hu et al. (2022). Source and target RGB videos are encoded with the backbone's video variational autoencoder (VAE) Kingma & Welling (2014); Wan et al. (2025). Across the two experts, the source token grids match in spatiotemporal layout, as do the target token grids, allowing their target-to-source addressing distributions to be aligned.

We represent target geometry as metric depth to retain a physical scale consistent with camera translation. To encode depth with the same VAE, we adopt the color mapping Ψ from Vision Banana Gabeur et al. (2026), which maps metric distances to RGB values and supports decoding back to depth. Letting $\mathcal{E}$ denote the VAE encoder, the depth latent is $Z^D = \mathcal{E}(\Psi(D))$. Ground-truth target depths are rendered directly in Unreal Engine Epic Games (2026), avoiding noise introduced by external geometry estimation in the training targets. Details are provided in Appendix B.4.

3.2 Addressing Distillation

Addressing Distillation (AD) supervises a geometry expert using target-to-source addressing distributions extracted from a frozen appearance expert conditioned on GT target depth (Figure 2). We extract these distributions from the attention heads of selected DiT blocks, denoted by the set $\mathcal{B}$, and average them across heads within each block before distillation.

Target-to-Source Addressing. Consider one attention head in a selected block; we omit block and head indices for clarity. Let N_{tgt} and N_{src} denote the numbers of target and source tokens. Let $q_i, k_j \in \mathbb{R}^d$ be the position-encoded query of target token i and key of source token j, where d is the head dimension. We define the addressing matrix $A \in \mathbb{R}^{N_{\text{tgt}} \times N_{\text{src}}}$ by normalizing query–key scores over source tokens:

$$A_{ij} = \frac{\exp(q_i^\top k_j / \sqrt{d})}{\sum_{j'=1}^{N_{\text{src}}} \exp(q_i^\top k_{j'} / \sqrt{d})}. \tag{1}$$

Each row $A_{i,:} \in \mathbb{R}^{N_{\text{src}}}$ satisfies $\sum_{j=1}^{N_{\text{src}}} A_{ij} = 1$ and describes the relative allocation of attention among source tokens for target query i. This distribution spans source regions and frames without requiring a unique point match. It is defined directly on the spatiotemporally compressed token grid produced by the video VAE Wan et al. (2025). Addressing supervision therefore requires no explicit mapping of frame-level correspondences onto latent tokens. For a block with H heads, we average the individually normalized distributions as $\bar{A} = \frac{1}{H} \sum_{h=1}^{H} A^{(h)}$, where $A^{(h)}$ denotes the distribution from head h. The two experts have matching source and target token layouts, allowing their head-averaged addressing matrices to be compared directly.

Probing and Distilling Addressing. We first train the appearance expert to synthesize target RGB video from source observations and GT target depth, then freeze it as the teacher. To obtain addressing supervision, we probe both experts at the pure-noise endpoint using the same noise realization $\epsilon \sim \mathcal{N}(0, I)$, with the shape of the target latent. The teacher receives source video X and GT target depth D, while the geometry student receives X, camera conditions $\mathcal{C}$, and the target-view

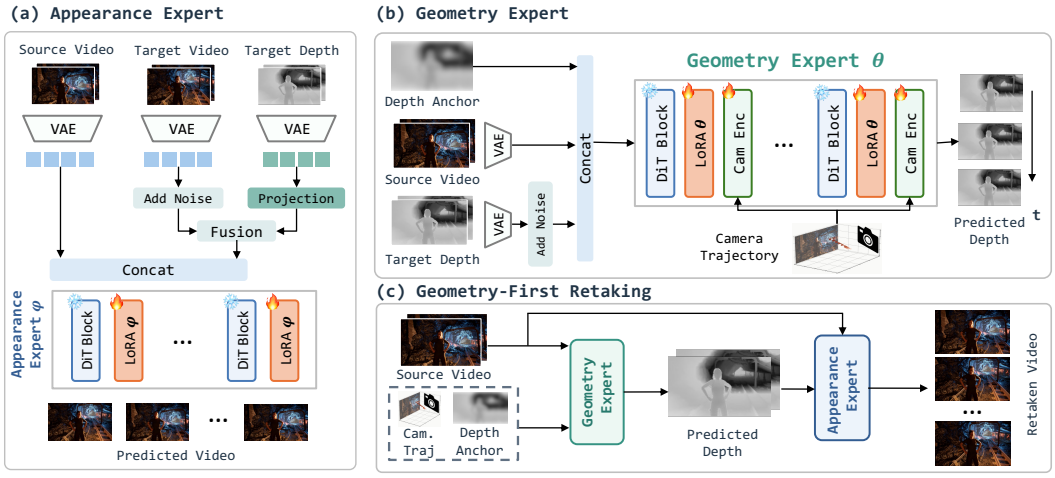


Figure 3: **Training and inference in GRAFT.** Appearance-First Training first trains an appearance expert with GT target depth, then freezes it to supervise geometry learning through Addressing Distillation. Geometry-First Retaking predicts target depth and supplies it to the same appearance expert for RGB synthesis. GT target depth and addressing distillation are used only during training.

first-frame depth reference a. Let $\mathrm{Addr}^{(b)}$ denote the operation that extracts the head-averaged addressing matrix at block $b \in \mathcal{B}$. For teacher parameters ϕ and student parameters θ, we obtain:

$$\bar{A}^{\mathrm{app},b} = \mathrm{Addr}_{\phi}^{(b)}(X, D, \epsilon), \qquad \bar{A}^{\mathrm{geo},b} = \mathrm{Addr}_{\theta}^{(b)}(X, \mathcal{C}, a, \epsilon). \tag{2}$$

At this endpoint, the teacher retains the target-depth condition while its noisy target input contains no residual target RGB content. Its addressing distributions therefore provide supervision under known target geometry for the student learning to predict that geometry.

We distill these distributions using row-wise Kullback–Leibler divergence, denoted by D_{KL}, over source tokens. With sg denoting stop-gradient, the addressing loss is:

$$\mathcal{L}_{\mathrm{addr}} = \frac{1}{|\mathcal{B}|} \sum_{b \in \mathcal{B}} \frac{1}{N_{\mathrm{tgt}}} \sum_{i=1}^{N_{\mathrm{tgt}}} D_{\mathrm{KL}}\left(\mathrm{sg}\left(\bar{A}_{i,:}^{\mathrm{app},b} \right) \,\|\, \bar{A}_{i,:}^{\mathrm{geo},b} \right). \tag{3}$$

The loss matches the student's source-token distribution to the teacher's for each target query, then averages over target queries and selected blocks. Combined with depth generation supervision, AD transfers addressing learned under known target geometry to geometry prediction without constructing explicit geometric correspondence labels.

3.3 Appearance-First Training

As illustrated in Figure 3, we train the appearance expert first, then freeze it to supervise the geometry expert through Addressing Distillation.

Training the appearance expert. Given source video X and GT target depth D, the appearance expert learns to synthesize the paired target RGB video Y. We project the clean depth latent $Z^D = \mathcal{E}(\Psi(D))$ to the backbone's hidden dimension and add it to the corresponding target RGB token features. Noisy target RGB tokens and clean source video tokens are concatenated as $[\mathrm{tgt}, \mathrm{src}]$. The appearance expert is trained with the standard flow matching objective Lipman et al. (2023).

Training the geometry expert. The geometry expert predicts Z^D from X, camera conditions $\mathcal{C}$, and the target-view first-frame depth reference a. We encode $\mathcal{C}$ using Unified Camera Positional Encoding (UCPE) Zhang et al. (2026) and use a as a geometric cue. Clean source video tokens and noisy target depth tokens are concatenated as $[\mathrm{src}, \mathrm{tgt}]$. In both experts, rotary positional embeddings (RoPE) Su et al. (2024) are applied independently to the source and target subsequences, with indices starting from zero within each. AD extracts target-query/source-key distributions according to token roles, aligning their rows and columns despite the different concatenation orders.

With the appearance expert frozen, the geometry training objective is:

$$\mathcal{L}_{\text{geo-train}} = \mathcal{L}_{\mathrm{geo}} + \beta \mathcal{L}_{\mathrm{addr}}, \tag{4}$$

Table 1: Quantitative comparison on **SpatialVID**, evaluating camera following, geometric consistency, and source fidelity. FVD is reported with I3D and VideoMAE features. Best and second-best results are shown in **bold** and underlined, respectively.

Method	Venue	Camera Control		Geometry		Visual Fidelity			
		Rot.↓	Trans.↓	Dyn-MEt3R↑	MEt3R↓	CLIP-V↑	FID↓	FVD_{I3D}↓	FVD_{VMAE}↓
ReCamMaster	ICCV'25	3.1781	0.0316	0.8193	0.2312	0.9148	30.76	567.37	878.27
TrajCrafter	ICCV'25	1.0248	0.0208	0.8483	0.2062	0.9207	31.96	517.68	826.56
CamCloneMaster	SIGGA'25	3.6689	0.0384	0.8335	0.2749	0.8895	37.62	566.29	1034.21
EX-4D	arXiv	1.5287	0.0191	0.8472	0.1944	0.8924	31.68	513.93	878.39
ReDirector	CVPR'26	3.2554	0.0195	0.8426	0.1943	0.9235	21.27	535.97	894.31
DepthDirector	ECCV'26	1.3734	0.0225	0.8374	0.1852	0.9232	21.08	490.66	844.57
Ours	–	**0.7627**	**0.0107**	**0.8777**	**0.1745**	**0.9357**	**18.03**	**439.42**	**809.63**

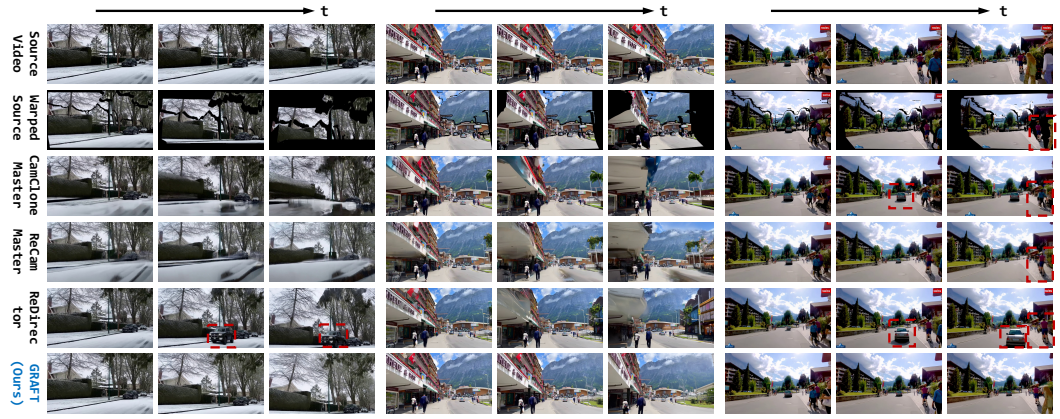


Figure 4: **Qualitative comparisons on SpatialVID.** Red boxes highlight regions with structural or content-placement errors. Warped source frames provide a visual reference for the requested camera motion and approximate scene layout; they are not inputs to GRAFT.

where $\mathcal{L}_{\text{geo}}$ is the flow matching loss for target depth generation, $\mathcal{L}_{\text{addr}}$ is defined in Equation 3, and β weights the addressing loss. We evaluate $\mathcal{L}_{\text{geo}}$ at noise levels sampled from the training schedule. For $\mathcal{L}_{\text{addr}}$, we evaluate both experts at the pure-noise endpoint with the same target noise, as described in Section 3.2.

3.4 Geometry-First Retaking at Inference

At inference, the geometry expert first predicts target-view metric depth $\widehat{D}$. The appearance expert then synthesizes the retaken RGB video $\widehat{Y}$ conditioned on this prediction and source video X:

$$\widehat{D} \sim p_\theta(D \mid X, \mathcal{C}, a), \qquad \widehat{Y} \sim p_\phi(Y \mid X, \widehat{D}), \tag{5}$$

where p_θ and p_ϕ denote the conditional generative distributions of the geometry and appearance experts, respectively. The generated depth latent is decoded through the VAE decoder and the depth color decoding procedure to obtain $\widehat{D}$. This depth supplies the appearance expert's depth condition, replacing GT target depth at inference. The two experts run sequentially, with no inference-time addressing distillation. This use of an explicit intermediate representation relates to Latent Forcing Baade et al. (2026), Semantics Lead the Way Pan et al. (2026), and Representation Forcing Wang et al. (2026c); in GRAFT, that representation is the target-view metric depth sequence.

4 Experiments

4.1 Experimental Setup

Evaluation benchmarks. We construct 400 source–target camera trajectory combinations from SpatialVID Wang et al. (2026a) to evaluate scene geometric consistency and camera following,

Table 2: Quantitative comparison on **HumanVid**, evaluating camera following, subject motion, identity preservation, and source fidelity.

Method	Camera		Motion			Identity			Visual Fidelity		
	Rot.↓	Trans.↓	Traj.↑	Flow↑	Amp.↑	FaceID↑	F-Temp↑	F-IFS↑	CLIP-V↑	FID↓	FVD↓
ReCamMaster	5.015	0.0542	0.237	0.253	0.614	0.7247	0.9280	0.9298	0.8968	44.03	410.28
TrajCrafter	2.537	0.0698	0.308	0.282	0.615	0.5473	0.9120	0.9129	0.8961	48.54	398.13
CamCloneMaster	6.850	0.0719	0.210	0.169	0.485	0.5131	0.7962	0.8177	0.7773	97.70	644.08
EX-4D	2.124	0.0613	0.294	0.225	0.582	0.4507	0.9170	0.9347	0.8482	50.23	411.25
ReDirector	5.123	0.0538	0.237	0.251	0.599	0.7208	0.9153	0.9096	0.9002	38.03	442.18
DepthDirector	8.112	0.0770	0.277	0.258	0.615	0.7509	0.8989	0.9078	0.8795	54.47	450.69
GRAFT (Ours)	**1.709**	**0.0400**	**0.379**	**0.319**	**0.676**	**0.7550**	**0.9437**	**0.9476**	**0.9346**	**26.93**	**346.15**



Figure 5: **Qualitative comparisons on HumanVid.** Frames progress from left to right within each example. Dashed boxes in the left example mark the expected subject locations; the middle example highlights the unfolding of the blanket.

and 300 combinations from HumanVid Wang et al. (2024a) to evaluate human identity and motion preservation. Both evaluations use monocular source videos without paired GT RGB or depth for the requested target trajectories. At inference, the first-frame depth reference a is constructed through geometry estimation using MoGe-3 Kong et al. (2026) and Pi3X Wang et al. (2026b), followed by warping; the geometry expert predicts the target depth sequence without GT target depth. During training, a is constructed from Unreal Engine-rendered depth. Reference construction details are provided in the appendix. We further evaluate GRAFT on DAVIS (Appendix A.3) and test its extension to differing source and target initial viewpoints on iPhonePTZ Yin et al. (2026) (Appendix A.4).

Evaluation metrics. We evaluate 5 aspects: **(i)** *Camera following and geometric consistency*. We estimate camera trajectories using ViPE Huang et al. (2025) and report rotation and translation errors (Rot. and Trans.) relative to the requested trajectories. MEt3R Asim et al. (2025) measures cross-view geometric consistency between corresponding source and generated frames, while Dyn-MEt3R Park et al. (2025) evaluates background geometric consistency across generated frames. **(ii)** *Source fidelity*. CLIP-V Radford et al. (2021) measures semantic consistency with the source video. Using source videos as the reference, FID Heusel et al. (2017) and FVD Unterthiner et al. (2018) measure image-level and video-level feature-distribution distances, respectively. **(iii)** *Subject motion preservation*. After camera-motion compensation, CoTracker-based Karaev et al. (2025) Trajectory Fidelity (Traj.), RAFT-based Teed & Deng (2020) Flow Fidelity (Flow), and Motion Amplitude Fidelity (Amp.) assess preservation of subject trajectories, residual motion fields, and motion magnitude, respectively. **(iv)** *Facial identity preservation*. Using ArcFace embeddings Deng et al. (2019), FaceID measures identity similarity to the source reference, while Face IFS (F-IFS) and Face Temp (F-Temp) measure identity consistency between adjacent generated frames. **(v)** *Visual quality and temporal consistency*. We report VBench Huang et al. (2024) subject consistency, background consistency, motion smoothness, and imaging quality. Detailed metric definitions and evaluation procedures are in the Appendix B.1.

Table 3: **Ablation study on SpatialVID and HumanVid.** GFR: Geometry-First Retaking; AD: Addressing Distillation; Warp: warped-depth completion followed by RGB synthesis. Best and second-best distinct values are **bold** and underlined. Avg. rank averages ranks over eight metrics at the reported precision, using average ranks for ties.

ID	Variant	SpatialVID				HumanVid				Avg. rank↓
		Rot.↓	Trans.↓	Dyn-MEt3R↑	MEt3R↓	Rot.↓	Trans.↓	Traj.↑	Amp.↑	
I. GFR and AD										
(a)	RGB	1.919	0.018	0.845	0.197	3.774	0.054	0.348	0.647	5.313
(b)	GFR	1.883	0.014	0.875	0.200	2.503	0.049	0.366	0.663	3.813
(c)	RGB + AD	1.174	0.012	**0.878**	0.176	2.116	0.049	**0.379**	**0.678**	2.000
II. External geometry injection										
(d)	Warp	1.954	0.018	0.863	0.188	3.355	0.052	0.342	0.645	5.188
(e)	Warp + AD	1.138	0.012	0.877	0.183	2.745	0.053	0.344	0.668	3.438
Full model: shared reference for I and II										
(f)	**GRAFT**	**0.763**	**0.011**	**0.878**	**0.175**	**1.709**	**0.040**	**0.379**	0.676	**1.250**

Compared methods. We compare GRAFT with six video retaking baselines: ReCamMaster Bai et al. (2025), ReDirector Park et al. (2026), TrajectoryCrafter (TrajCrafter) Yu et al. (2025), DepthDirector Chen et al. (2026), CamCloneMaster Luo et al. (2025), and EX-4D Hu et al. (2025). These methods cover direct camera-conditioned RGB generation, implicit end-to-end approaches, and approaches that incorporate explicit geometric guidance.

4.2 Comparison with Existing Methods

Scene geometry on SpatialVID. Table 1 evaluates scene geometric consistency, camera following, and source fidelity. GRAFT achieves the best MEt3R and Dyn-MEt3R scores, indicating stronger source–target and temporal geometric consistency. It also obtains the lowest rotation and translation errors, together with the highest CLIP-V and lowest FID and FVD scores. These results show improved geometric consistency and camera following while retaining source content and appearance. Figure 4 provides qualitative comparisons, with red boxes highlighting local errors in scene structure and content placement. GRAFT better preserves the scene layout across the illustrated frames.

Human identity and motion on HumanVid. Table 2 evaluates the preservation of human subjects under camera changes. GRAFT achieves the highest trajectory, flow, and motion amplitude fidelity scores, together with the best scores across the reported identity metrics. It also yields the lowest camera errors and improves CLIP-V, FID, and FVD. Figure 5 illustrates the corresponding differences in subject placement and action preservation. In the left example, the marked regions indicate the expected subject locations under the requested viewpoint. In the middle example, GRAFT preserves the unfolding of the blanket, whereas the compared methods failed.

General video quality. GRAFT also performs strongly on general video quality, achieving the best scores on most VBench metrics across both benchmarks. Appendix A.1 reports the full results.

4.3 Ablation Studies

We evaluate six variants in two groups using the same training data, pretrained backbone, and evaluation protocols, without GT target depth at evaluation. Geometry-First Retaking (GFR) predicts target depth before RGB synthesis, and Addressing Distillation (AD) supervises source addressing using the appearance teacher. Camera-conditioned variants build on UCPE Zhang et al. (2026), providing a strong starting point for evaluating our contributions. All depth-based variants use the same appearance expert. Table 3 reports the results.

I. Contributions of GFR and AD. This group compares (a) direct RGB generation, (b) GFR without AD, (c) RGB generation with AD, and (f) GRAFT, which combines GFR and AD. All four use explicit camera conditions; (b) and (f) additionally receive the first-frame depth reference a. Comparing (a) with (b), GFR improves camera following and motion fidelity, although MEt3R slightly worsens. AD improves all reported metrics in both prediction settings: (a) versus (c), and (b) versus



Figure 6: **Qualitative ablations.** RGB, GFR, Warp, and RGB+AD correspond to variants (a), (b), (d), and (c) in Table 3. Time progresses from left to right. White boxes mark approximate vehicle locations in warped references; red and blue boxes highlight discrepancies in generated results. Warped frames provide references for camera motion and layout.

(f). GRAFT achieves the best average rank and the lowest camera errors and MEt3R. RGB + AD ties GRAFT on Dyn-MEt3R and Trajectory Fidelity at the reported precision and retains a small advantage in Amplitude Fidelity.

II. Comparison with External Geometry Injection. This group tests warped-depth injection as an alternative to direct camera-conditioned depth prediction. Variants (d) and (e) replace explicit camera conditions with warped target-view depth and learn to complete it before RGB synthesis, without and with AD, respectively. We compare them with (b) and (f), which use camera conditions and the first-frame reference a. Direct prediction performs better on most metrics without AD and on all reported metrics with AD. Within the completion pipeline, AD improves geometric consistency and motion fidelity, but slightly increases camera errors on HumanVid. These results favor direct camera-conditioned depth prediction with a first-frame reference in our setting.

Figure 6 illustrates differences in vehicle motion and pedestrian placement. GRAFT better preserves the motion and target-view layout, while the ablated variants exhibit misplaced or missing subjects.

4.4 User Study

We compare GRAFT with three baselines in a user study. Participants rate Visual Quality (Vis. Qual.), Camera-Motion Plausibility (Cam. Plaus.), Subject Motion Preservation (Mot. Pres.), and Scene Structural Integrity (Struct. Int.) on a 1–5 scale. GRAFT achieves the highest mean rating on all four criteria, outperforming the strongest baseline for each by 0.59–0.93 points (Table 4). Appendix B.6 provides the details.

Table 4: User ratings (↑). Gray subscripts report standard deviations.

Method	Vis. Qual.	Cam. Plaus.	Mot. Pres.	Struct. Int.
ReCamMaster	$3.21_{\pm 1.12}$	$3.41_{\pm 1.11}$	$3.30_{\pm 1.14}$	$3.20_{\pm 1.17}$
ReDirector	$3.30_{\pm 1.17}$	$3.30_{\pm 1.04}$	$3.18_{\pm 1.01}$	$3.20_{\pm 1.15}$
DepthDirector	$2.64_{\pm 1.38}$	$2.98_{\pm 1.31}$	$2.86_{\pm 1.21}$	$2.48_{\pm 1.14}$
GRAFT	$\mathbf{4.23}_{\pm 0.74}$	$\mathbf{4.00}_{\pm 0.71}$	$\mathbf{4.02}_{\pm 0.82}$	$\mathbf{4.11}_{\pm 0.80}$

5 Conclusion

We presented **GRAFT**, a video retaking framework that improves source evidence retrieval through **Addressing Distillation**. By first training an appearance expert with ground-truth target depth, we obtain soft target-to-source addressing patterns that supervise a geometry expert without requiring explicit correspondence labels. At inference, the predicted target-view depth from the geometry expert guides the same appearance expert to synthesize the video. Experiments on dynamic-scene and human-centric settings show that GRAFT improves camera trajectory following, geometric consistency, and subject motion preservation. We hope this work provides a useful perspective on transferring learned source-addressing behavior across geometry and appearance generation tasks.

REFERENCES

Mohammad Asim, Christopher Wewer, Thomas Wimmer, Bernt Schiele, and Jan Eric Lenssen. Met3r: Measuring multi-view consistency in generated images. In *IEEE/CVF Computer Vision and Pattern Recognition (CVPR)*, 2025.

Alan Baade, Eric Ryan Chan, Kyle Sargent, Changan Chen, Justin Johnson, Ehsan Adeli, and Fei-Fei Li. Latent forcing: Reordering the diffusion trajectory for pixel-space image generation. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, 2026.

Sherwin Bahmani, Ivan Skorokhodov, Guocheng Qian, Aliaksandr Siarohin, Willi Menapace, Andrea Tagliasacchi, David B Lindell, and Sergey Tulyakov. Ac3d: Analyzing and improving 3d camera control in video diffusion transformers. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22875–22889. IEEE, 2025.

Jianhong Bai, Menghan Xia, Xiao Fu, Xintao Wang, Lianrui Mu, Jinwen Cao, Zuozhu Liu, Haoji Hu, Xiang Bai, Pengfei Wan, and Di Zhang. Recammaster: Camera-controlled generative rendering from a single video. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14834–14844, 2025. doi: 10.1109/ICCV51701.2025.01376.

Yunpeng Bai, Haoxiang Li, and Qixing Huang. Pe-field 4d: Video generation models as canvas, 2026. URL https://arxiv.org/abs/2607.15667.

Jonathan T. Barron. A power transform, 2025. URL https://arxiv.org/abs/2502.10647.

Dong-Yu Chen, Yixin Guo, Shuojin Yang, Taijiang Mu, and Shi-Min Hu. Beyond inpainting: Unleash 3d understanding for precise camera-controlled video generation. In *European Conference on Computer Vision (ECCV)*, 2026.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

Epic Games. Unreal engine. https://www.unrealengine.com/, 2026.

Valentin Gabeur, Shangbang Long, Songyou Peng, Paul Voigtlaender, Shuyang Sun, Yanan Bao, Karen Truong, Zhicheng Wang, Wenlei Zhou, Jonathan T. Barron, Kyle Genova, Nithish Kannen, Sherry Ben, Yandong Li, Mandy Guo, Suhas Yogin, Yiming Gu, Huizhong Chen, Oliver Wang, Saining Xie, Howard Zhou, Kaiming He, Thomas Funkhouser, Jean-Baptiste Alayrac, and Radu Soricut. Image generators are generalist vision learners, 2026. URL https://arxiv.org/abs/2604.20329.

Jia Guo, Jiankang Deng, Alexandros Lattas, and Stefanos Zafeiriou. Sample and computation redistribution for efficient face detection. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=RhB1AdoFfGE.

Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. Cameractrl: Enabling camera control for video diffusion models. In *International Conference on Learning Representations*, volume 2025, pp. 100433–100464, 2025.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

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

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

Tao Hu, Haoyang Peng, Xiao Liu, and Yuewen Ma. Ex-4d: Extreme viewpoint 4d video synthesis via depth watertight mesh, 2025. URL https://arxiv.org/abs/2506.05554.

Jiahui Huang, Qunjie Zhou, Hesam Rabeti, Aleksandr Korovko, Huan Ling, Xuanchi Ren, Tianchang Shen, Jun Gao, Dmitry Slepichev, Chen-Hsuan Lin, et al. Vipe: Video pose engine for 3d geometric perception. *arXiv preprint arXiv:2508.10934*, 2025.

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.

Hyeonho Jeong, {Chun Hao P.} Huang, Jong {Chul Ye}, {Niloy J.} Mitra, and Duygu Ceylan. Track4gen: Teaching video diffusion models to track points improves video generation. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 7276–7287, 2025. ISSN 1063-6919. doi: 10.1109/CVPR52734.2025.00682. Publisher Copyright: © 2025 IEEE.; 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025 ; Conference date: 11-06-2025 Through 15-06-2025.

Nikita Karaev, Yuri Makarov, Jianyuan Wang, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Cotracker3: Simpler and better point tracking by pseudo-labeling real videos. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1–10. IEEE, 2025.

Diederik P. Kingma and Max Welling. Auto-encoding variational Bayes. In *International Conference on Learning Representations*, 2014.

Lingyu Kong, Ruicheng Li, Ruicheng Wang, Sicheng Xu, Chengtang Yao, Jianfeng Xiang, and Jiaolong Yang. Moge-3: Fine-detail monocular geometry estimation with self-guided sparse volumetric refinement, 2026. URL https://arxiv.org/abs/2607.17967.

Min-Seop Kwak, Junho Kim, Sangdoo Yun, Dongyoon Han, Taekyung Kim, Seungryong Kim, and Jin-Hwa Kim. Aligned novel view image and geometry synthesis via cross-modal attention instillation. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=vjvwYexMQn.

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

JoungBin Lee, Jaewoo Jung, Jongmin Lee, Tongmin Kim, Hyunsung Kim, Takuya Narihira, Kazumi Fukuda, Jahyeok Koo, Jisang Han, Yuki Mitsufuji, and Seungryong Kim. Mvtrack4gen: Multi-view point tracking as geometric supervision for 4d video generation, 2026. URL https://arxiv.org/abs/2606.26087.

Ruilong Li, Brent Yi, Junchen Liu, Hang Gao, Yi Ma, and Angjoo Kanazawa. Cameras as relative positional encoding. *Advances in Neural Information Processing Systems*, 38:15984–16009, 2026a.

Yixuan Li, Yanhong Zeng, Ka Leong Cheng, Jiayi Zhu, Hanlin Wang, Wen Wang, Yihao Meng, Hao Ouyang, Qiuyu Wang, Yue Yu, et al. Cameraanything: Refilming videos with arbitrary camera control. In *European Conference on Computer Vision (ECCV)*, 2026b.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

Haofeng Liu, Yang Zhou, Ziheng Wang, Zhengbo Xu, Zhan Peng, Jie Ma, Jun Liang, Shengfeng He, and Jing Li. Mocam: Unified novel view synthesis via structured denoising dynamics. In *European Conference on Computer Vision (ECCV)*, 2026.

Xingchao Liu, Chengyue Gong, and qiang liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=XVjTT1nw5z.

Yawen Luo, Xiaoyu Shi, Jianhong Bai, Menghan Xia, Tianfan Xue, Xintao Wang, Pengfei Wan, Di Zhang, and Kun Gai. Camclonemaster: Enabling reference-based camera control for video generation. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, pp. 1–10, 2025.

Christopher Mei and Patrick Rives. Single view point omnidirectional camera calibration from planar grids. In *Proceedings 2007 IEEE International Conference on Robotics and Automation*, pp. 3945–3950. IEEE, 2007.

Takeru Miyato, Bernhard Jaeger, Max Welling, and Andreas Geiger. Gta: A geometry-aware attention mechanism for multi-view transformers. In *International Conference on Learning Representations*, volume 2024, pp. 8172–8208, 2024.

Avinash Paliwal, Adithya Iyer, Shivin Yadav, Muhammad Ali Afridi, and Midhun Harikumar. Reshoot-anything: A self-supervised model for in-the-wild video reshooting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2026.

Yue-Ming Pan, Ruoyu Feng, Qi Dai, Yuqing Wang, Wenfeng Lin, Mingyu Guo, Chong Luo, and Nanning Zheng. Semantics lead the way: Harmonizing semantic and texture modeling with asynchronous latent diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Byeongjun Park, Hyojun Go, Hyelin Nam, Byung-Hoon Kim, Hyungjin Chung, and Changick Kim. Steerx: Creating any camera-free 3d and 4d scenes with geometric steering. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 27326–27337. IEEE, 2025.

Byeongjun Park, Byung-Hoon Kim, Hyungjin Chung, and Jong Chul Ye. Redirector: Creating any-length video retakes with rotary camera encoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

William Peebles and Saining Xie. Scalable Diffusion Models with Transformers . In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, Los Alamitos, CA, USA, October 2023. IEEE Computer Society. doi: 10.1109/ICCV51070.2023.00387. URL https://doi.ieeecomputersociety.org/10.1109/ICCV51070.2023.00387.

Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation, 2018. URL https://arxiv.org/abs/1704.00675.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

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*, 2025.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 10674–10685. ieee, 2022.

Vincent Sitzmann, Semon Rezchikov, Bill Freeman, Josh Tenenbaum, and Fredo Durand. Light field networks: Neural scene representations with single-evaluation rendering. *Advances in neural information processing systems*, 34:19313–19325, 2021.

Chenxi Song, Yanming Yang, and Chi Zhang. World in world: Explore the world with world models, 2026. URL https://arxiv.org/abs/2609.11548.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021a. URL https://openreview.net/forum?id=St1giarCHLP.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021b. URL https://openreview.net/forum?id=PxTIG12RRHS.

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

Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *European conference on computer vision*, pp. 402–419. Springer, 2020.

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

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

Jiahao Wang, Yufeng Yuan, Rujie Zheng, Youtian Lin, Jian Gao, Lin-Zhuo Chen, Yajie Bao, Chang Zeng, Yanxi Zhou, Xiao-Xiao Long, et al. Spatialvid: A large-scale video dataset with spatial annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 42592–42603, 2026a.

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. In *The Fourteenth International Conference on Learning Representations*, 2026b. URL https://openreview.net/forum?id=DTQIjngDta.

Yuqing Wang, Zhijie Lin, Ceyuan Yang, Yang Zhao, Fei Xiao, Hao He, Qi Zhao, Zihan Ding, Fuyun Wang, Shuai Wang, Youliang Zhang, Haoqi Fan, and Xihui Liu. Representation forcing for bottleneck-free unified multimodal models, 2026c. URL https://arxiv.org/abs/2605.31604.

Zhenzhi Wang, Yixuan Li, Yanhong Zeng, Youqing Fang, Yuwei Guo, Wenran Liu, Jing Tan, Kai Chen, Tianfan Xue, Bo Dai, et al. Humanvid: Demystifying training data for camera-controllable human image animation. *Advances in Neural Information Processing Systems*, 37:20111–20131, 2024a.

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

Suttisak Wizadwongsa, Hyelin Nam, Supasorn Suwajanakorn, and Jeong Joon Park. Sierpinskicam: Camera-controlled video retaking with sierpinski triangle pattern cues, 2026. URL https://arxiv.org/abs/2606.17310.

Kejia Yin, Zhihao Shi, Weilin Wan, Yuhongze Zhou, Yuanhao Yu, Xinxin Zuo, Qiang Sun, and Juwei Lu. Camdirector: Towards long-term coherent video trajectory editing. *arXiv preprint arXiv:2603.02256*, 2026.

Mark Yu, Wenbo Hu, Jinbo Xing, and Shan Ying. Trajectorycrafter: Redirecting camera trajectory for monocular videos via diffusion models. *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 100–111, 2025. URL https://api.semanticscholar.org/CorpusID:276885325.

Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=Sks9_ajex.

Cheng Zhang, Boying Li, Meng Wei, Yan-Pei Cao, Camilo Gambardella, Dinh Phung, and Jianfei Cai. Unified camera positional encoding for controlled video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 38027–38037, 2026.

Supplementary Material

A Additional Results and Analysis

A.1 General Video Quality Evaluation

Table 5 reports VBench evaluation on the same 400 SpatialVID and 300 HumanVid source–trajectory combinations used in the main experiments. All methods are evaluated on the same samples, with no generation failures or excluded samples. We use the official VBench implementation to evaluate subject consistency, background consistency, imaging quality, and motion smoothness within the generated videos; these metrics do not compare the outputs with the source videos. GRAFT achieves the best results on most metrics across both datasets, demonstrating strong visual quality and temporal consistency.

A.2 Additional Qualitative Results

Removing source camera motion. Figure 7 demonstrates video stabilization through retaking. GRAFT removes the original camera motion while preserving the subjects' actions and scene appearance. Each example compares source and retaken frames at matching time points, illustrating the separation of camera motion from the observed scene dynamics.

A.3 Additional Evaluation on DAVIS

We evaluate GRAFT and six baselines on 300 source–target camera trajectory combinations constructed from DAVIS Pont-Tuset et al. (2018). All methods use the same source videos and target camera trajectories, following the evaluation protocol in Section 4.1. As shown in Table 6, GRAFT achieves the best results on eight of the nine metrics, further demonstrating its effectiveness in camera following, geometric consistency, and source fidelity.

A.4 Retaking with Different Initial Viewpoints

First-frame reference. To accommodate a target camera whose initial position differs from that of the source camera, we first estimate depth from the first source RGB frame. Using the camera intrinsics and relative pose, we warp this depth into the first target view and inject the resulting depth reference a and its validity mask into the geometry expert through additive feature conditioning. The geometry expert then predicts the target depth sequence, which guides RGB synthesis. Appendix B.2 provides the conditioning details.

Temporal reversal augmentation. Following the temporal reversal strategy of ReDirector Park et al. (2026), we reverse the temporal order of paired source and target sequences together with their associated depth and camera sequences. A pair that shares the same initial viewpoint can have different final viewpoints because its two camera trajectories diverge. After reversal, these final viewpoints become the initial viewpoints of the new pair. This produces training examples with different source and target starting viewpoints while preserving temporal correspondence between the paired observations. For forward-order training pairs, we use the GT target first-frame depth directly; for reversed pairs, we construct the reference by reprojecting the first source depth into the first target view.

Qualitative extension. Figure 8 demonstrates retaking with an initial target-camera position different from the source camera's initial position. We construct the first-frame depth reference by reprojection, with implementation details deferred to Appendix B.2. The source and generated frames are shown at matching time points. These results illustrate GRAFT's applicability beyond a shared initial camera position while retaining the source scene content and subject actions.

Evaluation on iPhonePTZ. We evaluate our method on the iPhonePTZ Yin et al. (2026) benchmark, each with different initial camera positions and synchronized ground-truth target RGB video. All seven metrics in Table 7 use the ground-truth target videos as the reference. GRAFT achieves

Table 5: VBench evaluation on **SpatialVID** and **HumanVid**. We report subject consistency (Sub.), background consistency (BG.), imaging quality (Imaging), and motion smoothness (Smooth). Higher is better for all metrics. Best results are in **bold** and second-best distinct results are underlined; ties at the reported precision receive the same marking.

Method	SpatialVID				HumanVid			
	Sub.↑	BG.↑	Imaging↑	Smooth↑	Sub.↑	BG.↑	Imaging↑	Smooth↑
DepthDirector	0.9438	0.9598	0.6642	**0.9934**	0.8661	0.9208	0.6663	0.9876
ReCamMaster	0.9348	0.9247	0.6277	0.9931	0.8981	0.9091	0.6485	**0.9915**
TrajectoryCrafter	0.9453	0.9584	0.5748	0.9820	0.9083	0.9349	0.5990	0.9762
CamCloneMaster	0.9185	0.9139	0.5870	0.9929	0.8098	0.8717	0.5589	0.9894
EX-4D	0.9392	0.9502	0.5030	0.9825	0.8977	0.9308	0.5514	0.9764
ReDirector	0.9496	0.9573	0.6699	0.9911	0.9010	0.9287	0.6768	0.9908
GRAFT (Ours)	**0.9582**	**0.9609**	**0.6918**	0.9896	**0.9344**	**0.9432**	**0.7159**	**0.9915**

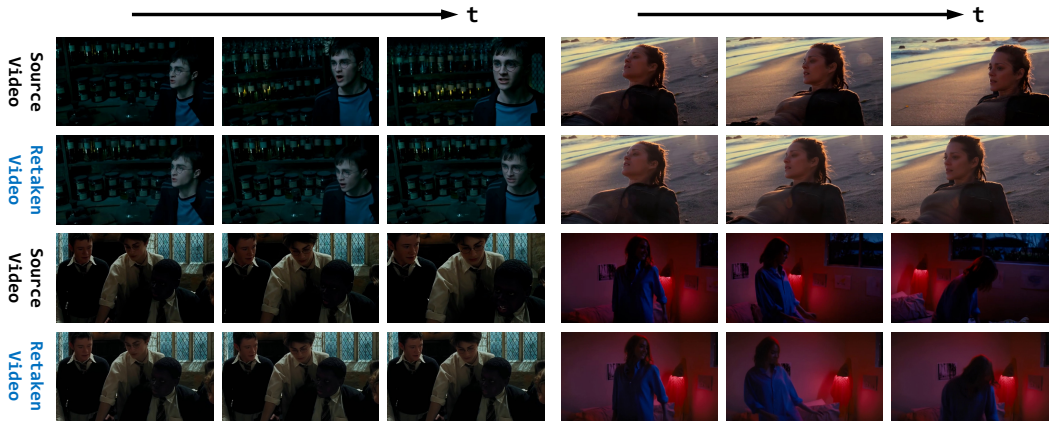


Figure 7: **Video stabilization through retaking.** GRAFT removes the original camera motion while preserving subject actions and scene appearance. For each example, the upper row shows the source video and the lower row shows the stabilized output. Columns show matching time points, progressing from left to right.

the best results on six of the seven metrics, demonstrating effective retaking with different initial viewpoints. Figure 9 presents qualitative comparisons at matching time points.

A.5 Source-Addressing Visualization

Figure 10 examines the relationship between source-addressing patterns and motion preservation. We compare GRAFT with GFR, the geometry-first variant without AD, under otherwise matched training and inference settings. In this example, GFR places excessive addressing weight on the first source frame, where the vehicle remains at its initial position. This emphasis on initial-frame evidence accompanies the vehicle's persistence at that position in the generated sequence. With AD, GRAFT reduces the initial-frame bias and preserves the vehicle's subsequent movement. The comparison links the illustrated motion-preservation failure to how the geometry expert allocates attention across the source video.

For this visualization, we probe the geometry expert at the pure-noise endpoint and average addressing patterns over blocks 15–30 in its latter part. The target queries comprise all spatial tokens in the 15th latent temporal slice, corresponding to target video frame 56. The temporal curve reports mean addressing weights, averaging over heads, the selected target queries, blocks, and source spatial positions at each source latent time. The vertical dashed line marks the target time in this plot. This visualization uses a broader block range than the three blocks selected for the training-time AD loss.

Table 6: Quantitative comparison on 300 source–target camera trajectory combinations from **DAVIS**. Rot. and Trans. denote camera rotation and translation errors; CLIP-V measures source semantic similarity; Dyn-MEt3R measures temporal background geometric consistency. Sub., BG., Smooth, and Imaging denote VBench subject consistency, background consistency, motion smoothness, and imaging quality; Traj. denotes Trajectory Fidelity. Best and second-best results are shown in **bold** and underlined, respectively.

Method	Rot.↓	Trans.↓	CLIP-V↑	Sub.↑	BG.↑	Smooth↑	Imaging↑	Dyn-MEt3R↑	Traj.↑
ReCamMaster	2.93673	0.07684	0.87224	0.88815	0.90050	0.99086	0.55525	0.79106	0.38799
TrajectoryCrafter	6.21844	0.08925	0.86253	0.87559	0.91946	0.97225	0.52252	0.74327	0.39137
CamCloneMaster	6.18720	0.07539	0.81416	0.85530	0.88585	**0.99147**	0.52140	0.74977	0.32270
EX-4D	6.03796	0.08432	0.83457	0.85708	0.90964	0.97315	0.44884	0.73200	0.37717
ReDirector	2.29791	0.06172	0.86938	0.89240	0.91860	0.98989	0.59692	0.79102	0.42761
DepthDirector	21.11072	0.14307	0.68335	0.66887	0.80861	0.96842	0.50465	0.60348	0.33293
GRAFT (Ours)	**1.83044**	**0.05558**	**0.88415**	**0.91130**	**0.92491**	0.98933	**0.62868**	**0.81161**	**0.42899**

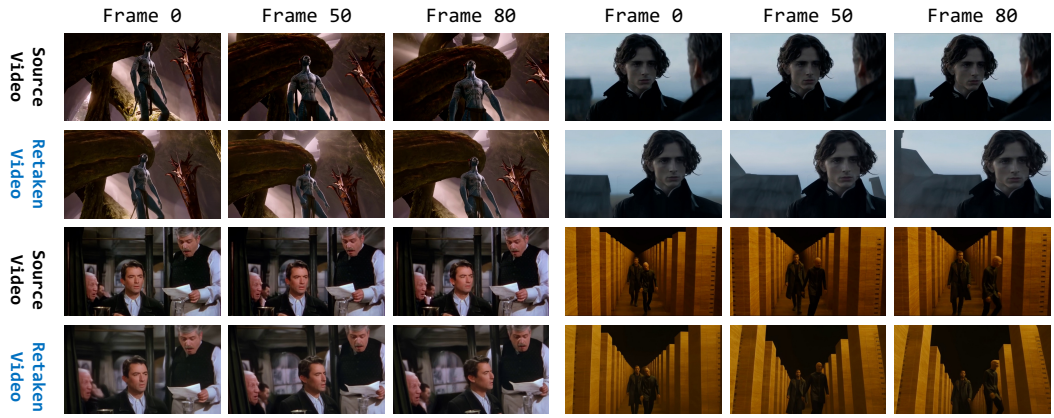


Figure 8: **Extension to different initial camera positions.** GRAFT retakes source videos with a target camera whose initial position differs from that of the source camera. Each example shows source frames in the upper row and generated frames at the same time points in the lower row. Columns correspond to frames 0, 50, and 80. These examples illustrate the applicability of GRAFT beyond a shared initial camera position.

A.6 Training Dynamics

Figure 11 shows the smoothed training losses over 20,000 appearance-training steps and 30,000 geometry-training steps. The appearance generation loss decreases and gradually levels off. During geometry training, both the depth generation loss and the addressing loss decrease, indicating improved depth prediction and closer agreement with the teacher's source-addressing distributions. The addressing curve reports the raw loss before multiplication by $\beta = 0.05$.

The depth generation loss is computed at every geometry-training step, while AD is activated once every five optimizer steps. Training runtime is discussed in Appendix B.7.

B Implementation and Evaluation Details

B.1 Implementation Details

Backbone and tokenization. We use the official `Wan-AI/Wan2.2-TI2V-5B` checkpoint Wan et al. (2025), loaded in BF16, with the Wan2.2 VAE and UMT5-XXL text encoder. Each video contains 81 frames at a resolution of 384×704 (height $\times$ width). The VAE produces a latent of shape $48 \times 21 \times 24 \times 44$ (channels $\times$ time $\times$ height $\times$ width). With a DiT patch size of $1 \times 2 \times 2$, each

Table 7: Quantitative comparison on 40 **iPhonePTZ** source–target video pairs with different initial camera positions. All metrics use the synchronized ground-truth target RGB videos as the reference. Arrows indicate the preferred direction for each metric. Best and second-best results are shown in **bold** and underlined, respectively.

Method	PSNR↑	SSIM↑	L1↓	LPIPS↓	CLIP↑	FID↓	FVD_{I3D}↓
ReCamMaster	11.175	0.299	0.219	0.714	0.829	111.612	727.283
CamCloneMaster	10.942	0.280	0.225	0.706	0.812	119.949	828.017
ReDirector	11.790	0.320	0.204	0.671	0.820	**111.293**	722.862
DepthDirector	11.847	0.307	0.208	0.693	0.742	170.771	938.045
GRAFT (Ours)	**13.383**	**0.356**	**0.168**	**0.562**	**0.838**	112.857	**605.454**

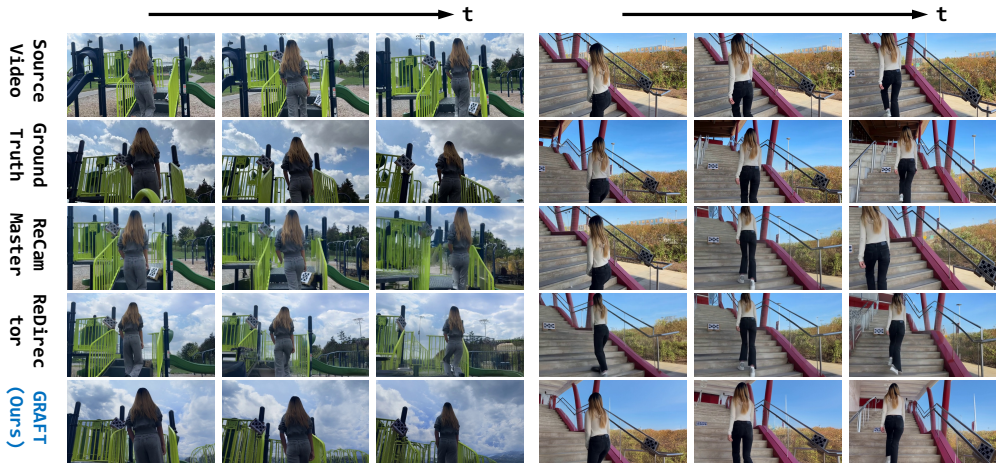


Figure 9: **Qualitative comparison on iPhonePTZ with different initial viewpoints.** From top to bottom: source video, ground-truth target video, ReCamMaster, ReDirector, and GRAFT. Each group shows one example at matching time points across rows, with time progressing from left to right. The source and target cameras begin at different positions; the ground-truth target row provides the reference for appearance, subject configuration, and viewpoint.

video stream contains $21 \times 12 \times 22 = 5{,}544$ tokens. The concatenated target and source streams therefore contain 11,088 tokens.

Geometry-stage optimization. The geometry student is trained for 30,000 steps and adapted using rank-32 LoRA Hu et al. (2022) on the attention projections `q`, `k`, `v`, and `o`, and the feed-forward layers `ffn.0` and `ffn.2`. Its trainable components are depth LoRA (600 parameter tensors) and target-only UCPE (240 parameter tensors), totaling 222,293,760 trainable parameters. The Wan backbone and the entire RGB teacher are frozen, and Shared RoPE has no trainable parameters. The reference validity mask is encoded with the video VAE and projected for additive conditioning. Table 8 summarizes the confirmed configuration.

Appearance teacher. Geometry training uses the frozen appearance checkpoint saved at step 20,000. Its conditioning includes the source RGB sequence, target depth, a white mask, and the first source RGB frame; it receives no camera or UCPE conditioning.

Training data. The Unreal Engine training data released by DepthDirector Chen et al. (2026) contain 1,217 scenes and 8,449 RGB videos captured along different camera trajectories. We use 50,336 directed source–target training pairs. Paired observations depict synchronized scene dynamics from different viewpoints, with rendered metric depth and calibrated camera parameters supporting target depth prediction and RGB synthesis.

Inference. The geometry expert first generates target-view depth from the source video, target camera conditions, and initial depth reference. The predicted depth then conditions the appearance

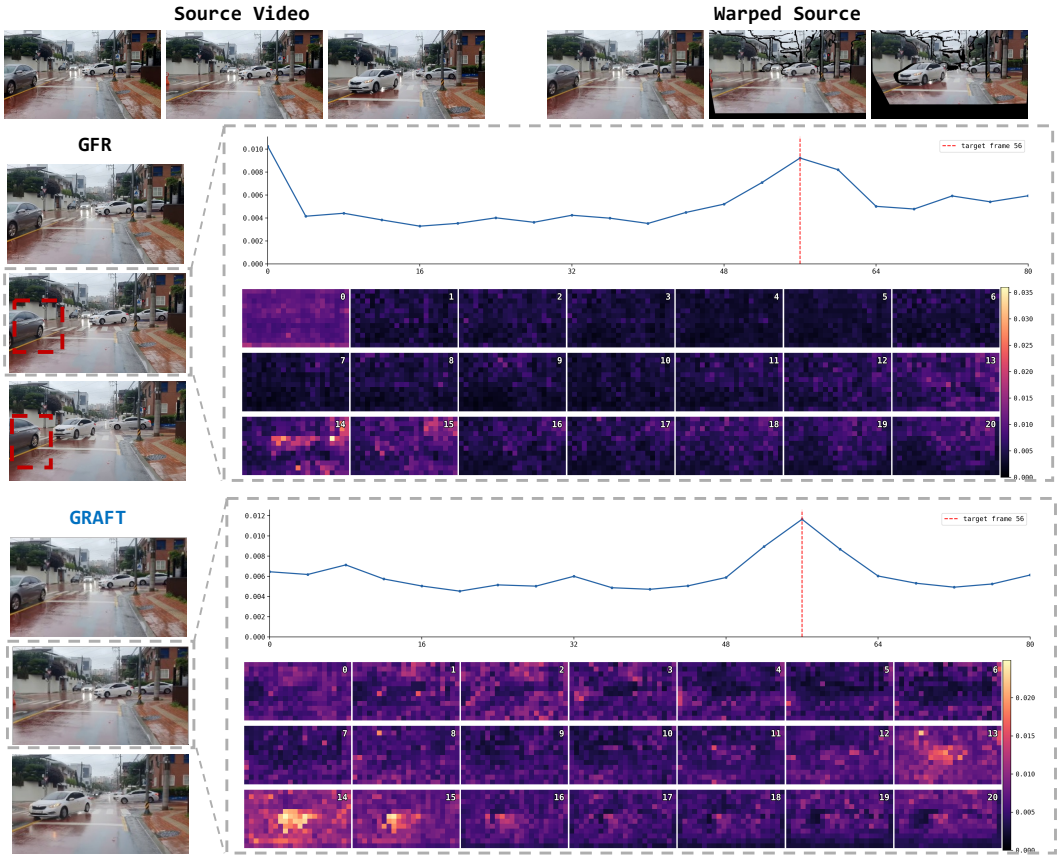


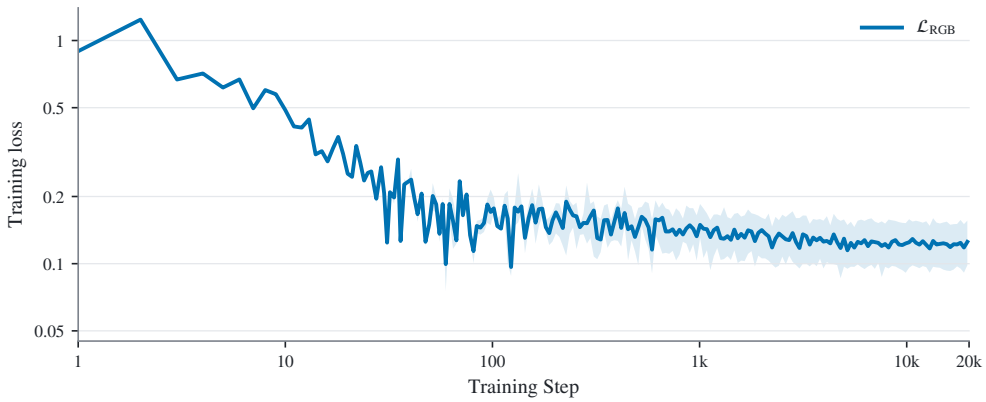
Figure 10: **Source-addressing patterns and motion preservation.** The top rows show source and warped source frames. The upper and lower comparison panels show GFR (without AD) and GRAFT, respectively, with generated frames on the left, temporal addressing curves above, and spatial addressing heatmaps across source latent times below. GFR assigns excessive weight to the first source frame, accompanying the vehicle's persistence at its initial position, highlighted by the red dashed boxes.

expert to synthesize the retaken RGB video. Both experts use 30 denoising steps. They run sequentially, and AD is not used at inference.

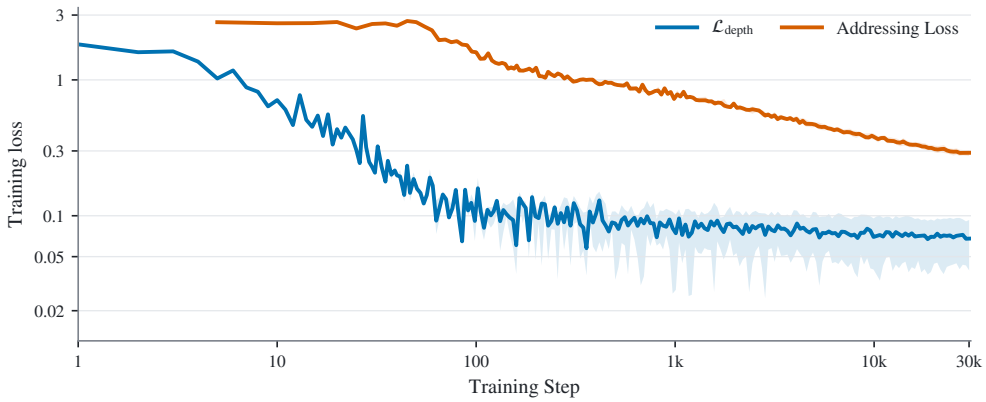
B.2 Conditioning and First-Frame Reference Construction

Appearance conditioning. The appearance expert receives the source RGB sequence, target depth, a white mask, and the first source RGB frame, without camera/UCPE conditioning. The clean target depth is color-encoded and passed through the same video VAE used for RGB, yielding $Z^D = \mathcal{E}(\Psi(D))$. A projection maps the depth latent to the backbone hidden dimension, and the resulting features are added to the corresponding target RGB token features. The appearance expert concatenates noisy target RGB tokens and clean source RGB tokens in the order $[\mathrm{tgt}, \mathrm{src}]$.

Geometry conditioning. The geometry expert receives the source RGB video, camera conditions $\mathcal{C}$ through target-only UCPE Zhang et al. (2026), and the target-view first depth reference a together with its validity mask. The depth-reference features are added only to the first target latent temporal slice. The binary validity mask is 1 at valid projected locations and 0 at projection holes. It is encoded with the video VAE, projected to the hidden dimension, and also added only to the first target latent temporal slice, distinguishing observed reference regions from missing ones. Its source RGB tokens and noisy target depth tokens are concatenated in the order $[\mathrm{src}, \mathrm{tgt}]$. In both experts, RoPE is applied independently to the source and target subsequences, with indices starting from



(a) Appearance expert training



(b) Geometry expert training

Figure 11: **Training dynamics of GRAFT.** Smoothed losses for (a) appearance training over 20,000 steps and (b) geometry training over 30,000 steps. Panel (b) shows the depth generation loss and the raw addressing loss before weighting by $\beta = 0.05$. AD is activated once every five optimizer steps. Both axes use logarithmic scales.

zero within each subsequence. AD extracts target queries and source keys according to their roles, so that the different concatenation orders do not change the alignment of the distilled distributions.

Reference construction. We combine MoGe-3 Kong et al. (2026) and Pi3X Wang et al. (2026b) to construct the first-frame depth reference. MoGe-3 provides fine geometric detail and sharp depth boundaries, while Pi3X estimates the overall scene scale and camera parameters. We align the MoGe-3 depth to the Pi3X depth by fitting a scale and an offset in depth space. Let D_{M} and D_{P} denote the corresponding MoGe-3 and Pi3X depth maps. The least-squares fit is

$$(s^\star, b^\star) = \arg\min_{s,b} \sum_{\mathbf{u}\in\Omega} \left(sD_{\text{M}}(\mathbf{u}) + b - D_{\text{P}}(\mathbf{u})\right)^2, \qquad D_{\text{aligned}} = s^\star D_{\text{M}} + b^\star, \tag{6}$$

where Ω denotes the pixel set used for alignment. We then reproject D_{aligned} into the first target view using the camera intrinsics and relative pose. The resulting target-view depth and its validity mask are injected into the geometry expert through additive feature conditioning. During training, forward-order samples use the GT target first-frame depth directly, while temporally reversed samples use source-depth reprojection to construct the target-view reference. The reprojection equations are given in Appendix C.2, and temporal reversal augmentation for differing initial viewpoints is described in Appendix A.4.

Table 8: **Confirmed geometry-stage training configuration.** The global batch size is the product of eight GPUs and one sample per GPU, with gradient accumulation set to one. AD is activated once every five optimizer steps.

Setting	Value
Backbone	`Wan-AI/Wan2.2-TI2V-5B`
Model loading precision	BF16
Video frames and resolution	81; 384×704
LoRA rank	32
LoRA target modules	`q, k, v, o, ffn.0, ffn.2`
Learning rate	10^{-4}
Weight decay	0.01
GPU count / samples per GPU	8 / 1
Gradient accumulation	1
Global batch size	8
Student trainable parameters	222,293,760
AD blocks (zero-based)	$\{17, 23, 27\}$
AD interval / active-step weight	5 optimizer steps / 0.05

B.3 Addressing Distillation Details

Matched probing inputs. Both experts are probed with the same target-shaped Gaussian noise realization at noise level $\sigma = 1$. This is the pure-noise endpoint ($t = 0$ under the noise-to-data convention in Appendix C.1). The teacher retains its source-video, target-depth, white-mask, and first-source-frame conditions. The student receives the source RGB sequence, the target-view first-frame depth reference and validity mask, and target-only UCPE conditions. Thus, the noisy target streams contain no residual clean target latent, while each expert retains its own conditioning inputs. The teacher is frozen throughout geometry training.

Source-only attention recomputation. AD uses the zero-indexed DiT blocks

$$\mathcal{B} = \{17, 23, 27\}, \tag{7}$$

corresponding to the 18th, 24th, and 28th blocks. For each block, we extract the target queries and source keys after the RMSNorm/projection/Shared-RoPE processing. Their shapes are

$$Q_{\mathrm{tgt}}, K_{\mathrm{src}} \in \mathbb{R}^{B_{\mathrm{mb}} \times 24 \times 5544 \times 128}, \tag{8}$$

where B_{mb} denotes the local minibatch size. For each example m, block b, and head h, we recompute

$$A^{(b,h)}_{m,i,j} = \frac{\exp\Big(\langle Q^{(b,h)}_{m,i}, K^{(b,h)}_{m,j}\rangle / \sqrt{128}\Big)}{\sum_{j'=1}^{5544} \exp\Big(\langle Q^{(b,h)}_{m,i}, K^{(b,h)}_{m,j'}\rangle / \sqrt{128}\Big)}. \tag{9}$$

The softmax is taken over all 5,544 source tokens, so source frames and spatial locations compete within the same normalization. There is no separate per-frame normalization. This distribution is recomputed from the post-RoPE query and key tensors; it is not obtained by cropping a source region from the full self-attention weight matrix.

Head averaging and KL reduction. We first average the 24 independently normalized head distributions:

$$\bar{A}^{(b)}_{m,i,j} = \frac{1}{24} \sum_{h=1}^{24} A^{(b,h)}_{m,i,j}. \tag{10}$$

Let $P^{(b)} = \mathrm{sg}(\bar{A}^{(b)}_{\mathrm{teacher}})$ and $S^{(b)} = \bar{A}^{(b)}_{\mathrm{student}}$. The loss for block b is

$$L_b = \frac{1}{B_{\mathrm{mb}} N_{\mathrm{tgt}}} \sum_{m=1}^{B_{\mathrm{mb}}} \sum_{i=1}^{N_{\mathrm{tgt}}} \sum_{j=1}^{N_{\mathrm{src}}} P^{(b)}_{m,i,j} \log \frac{P^{(b)}_{m,i,j}}{S^{(b)}_{m,i,j}}, \qquad N_{\mathrm{tgt}} = N_{\mathrm{src}} = 5544. \tag{11}$$

We then average the three block losses with equal weights:

$$\mathcal{L}_{\mathrm{addr}} = \frac{L_{17} + L_{23} + L_{27}}{3}. \tag{12}$$

Thus, KL is evaluated after head averaging, rather than evaluated separately per head and then averaged. Its direction is teacher-to-student, with a sum over source tokens and a mean over examples, target queries, and selected blocks. This matches the reduction in Equation 3.

Update schedule. Let n be the geometry-training optimizer-step index. The implemented objective is

$$\mathcal{L}^{(n)}_{\text{geo-train}} = \begin{cases} \mathcal{L}^{(n)}_{\text{geo}} + 0.05\mathcal{L}^{(n)}_{\text{addr}}, & n \bmod 5 = 0, \\ \mathcal{L}^{(n)}_{\text{geo}}, & \text{otherwise.} \end{cases} \tag{13}$$

The depth flow-matching objective is evaluated at every optimizer step. AD is evaluated once every five optimizer steps, with weight 0.05 applied directly on those steps and no factor-of-five compensation.

B.4 Depth Encoding

To represent metric depth with Wan's RGB video VAE Wan et al. (2025), we adopt the depth-to-color mapping from Vision Banana Gabeur et al. (2026). This mapping converts nonnegative metric depth into three-channel values through a bounded power transform followed by interpolation along an RGB-cube path. It allows RGB and depth to use the same encoder and latent token layout without introducing a separate depth autoencoder.

Depth normalization. For depth $d \geq 0$ measured in meters, we use the rescaled power transform Barron (2025)

$$f(d, \lambda, c) = 1 - \left(1 - \frac{d}{\lambda c}\right)^{\lambda+1}, \qquad \lambda < -1, \quad c > 0, \tag{14}$$

which maps $[0, \infty)$ monotonically onto $[0, 1)$. Following Vision Banana, we set $\lambda = -3$ and $c = 10/3$, giving

$$f(d) = 1 - (1 + d/10)^{-2}, \qquad f^{-1}(s) = 10\left((1 - s)^{-1/2} - 1\right). \tag{15}$$

The transform allocates more of the normalized range to nearby geometry and compresses larger distances. For example, $f(5) = 0.56$, $f(10) = 0.75$, and $f(40) = 0.96$. The same mapping is applied across videos without per-video depth normalization.

RGB encoding. The normalized depth is mapped onto a seven-segment path through the RGB cube with ordered vertices

$$\begin{aligned} &\mathbf{c}_0 = (0,0,0), \quad \mathbf{c}_1 = (1,0,0), \quad \mathbf{c}_2 = (1,1,0), \quad \mathbf{c}_3 = (0,1,0), \\ &\mathbf{c}_4 = (0,1,1), \quad \mathbf{c}_5 = (0,0,1), \quad \mathbf{c}_6 = (1,0,1), \quad \mathbf{c}_7 = (1,1,1). \end{aligned} \tag{16}$$

Let $u = 7f(d)$, $k = \lfloor u \rfloor$, and $\tau = u - k$. The encoded color is

$$\Psi(d) = (1 - \tau)\mathbf{c}_k + \tau\mathbf{c}_{k+1}. \tag{17}$$

Each segment varies one color channel. Finite depths map uniquely onto this path, with white approached as $d \to \infty$. The mapping is invertible on its image, rather than on the entire RGB cube.

Recovering metric depth. A VAE-decoded color $\mathbf{r} \in \mathbb{R}^3$ may lie off the encoding path. We project it onto the nearest segment:

$$(k^\star, \tau^\star) \in \underset{\substack{k \in \{0,\ldots,6\} \\ \tau \in [0,1]}}{\arg\min} \left\| \mathbf{r} - ((1 - \tau)\mathbf{c}_k + \tau\mathbf{c}_{k+1}) \right\|_2^2. \tag{18}$$

The recovered normalized value is $\hat{s} = (k^\star + \tau^\star)/7$, yielding

$$\widehat{d} = f^{-1}(\hat{s}) = 10\left((1 - \hat{s})^{-1/2} - 1\right), \qquad \hat{s} < 1. \tag{19}$$

The endpoint $\hat{s} = 1$ corresponds to infinite depth. We use Ψ^{-1} to denote this projection-based decoding operation. Projection assigns an off-path color a depth value, but does not guarantee recovery of the original depth after reconstruction errors.

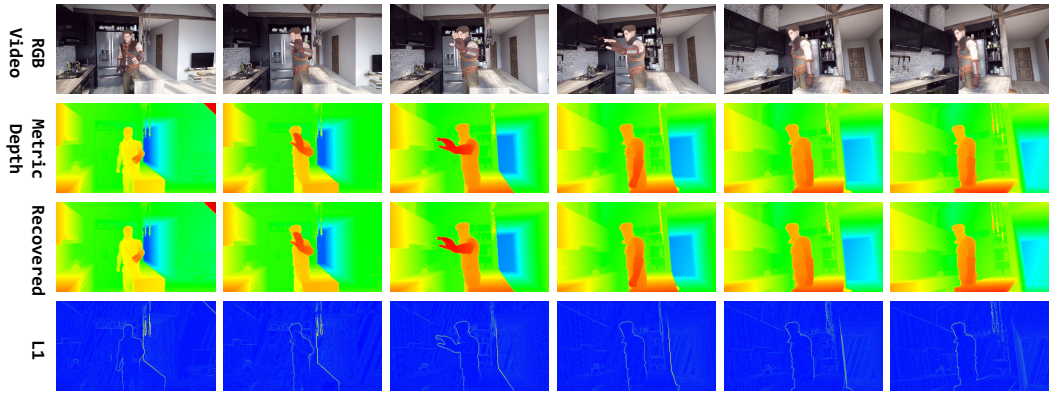


Figure 12: **Depth encoding and VAE round-trip.** From top to bottom: corresponding RGB frames, original metric depth D, recovered depth $\widetilde{D}$ after depth-to-color encoding, VAE encoding/decoding, and color-to-depth decoding, and per-pixel absolute error $|\widetilde{D} - D|$ in meters. Columns show successive frames. This reconstruction uses only the depth representation and video VAE, without geometry-expert prediction.

VAE round-trip. Let $\mathcal{E}$ and $\mathcal{D}$ denote the video VAE encoder and decoder. Applying Ψ pointwise to a depth sequence D, the encoded latent and reconstructed metric depth are

$$Z^D = \mathcal{E}(\Psi(D)), \qquad \widetilde{D} = \Psi^{-1}\big(\mathcal{D}(Z^D)\big)\,. \tag{20}$$

Although the color mapping is analytically invertible on its path, the VAE introduces reconstruction error. Figure 12 evaluates this representation through a depth-to-color encoding, VAE encoding/decoding, and color-to-depth decoding round-trip, without running the geometry expert. It compares the original and recovered metric depth across successive frames and reports per-pixel absolute reconstruction errors in meters.

B.5 Identity and Motion Metrics

Face detection and representation. We use InsightFace's `buffalo_l` models: the SCRFD detector Guo et al. (2022) (`det_10g.onnx`) and ArcFace ResNet-50 Deng et al. (2019) (`w600k_r50.onnx`). Detected faces are aligned using five landmarks and resized to 112×112. ArcFace embeddings are L2-normalized, so their dot product is cosine similarity. The detector threshold is 0.5. Source faces used to construct an identity prototype additionally require confidence ≥ 0.6 and a bounding-box short side ≥ 24 pixels, with at least three valid source detections per video.

Face-ID. Face-ID measures identity preservation relative to a multi-frame source identity prototype. Let e_j^s be the embeddings of the N_s reliable source faces. We construct

$$p = \text{normalize}\left(\frac{1}{N_s}\sum_{j=1}^{N_s} e_j^s\right). \tag{21}$$

For each generated frame I_t^g, we select the detected face whose embedding is most similar to p:

$$e_t^g \in \arg\max_{e \in \mathcal{F}(I_t^g)} p^\top e, \qquad s_t = p^\top e_t^g, \tag{22}$$

where $\mathcal{F}(I_t^g)$ is the set of detected face embeddings. For video n, let $\mathcal{V}_n$ contain the generated frames with a selected face. We take the median similarity within each video, then the median across valid videos $\mathcal{S}_{\text{ID}}$:

$$\text{FaceID}_n = \text{median}_{t \in \mathcal{V}_n}\, s_t, \qquad \text{FaceID} = \text{median}_{n \in \mathcal{S}_{\text{ID}}}\, \text{FaceID}_n. \tag{23}$$

Higher scores indicate stronger preservation of the source identity. Frames without a selected face are excluded, rather than assigned zero.

Face temporal consistency. F-Temp uses the same multi-frame source prototype and face-selection procedure as Face-ID. Let $\mathcal{P}_n$ contain indices t for which both original adjacent frames t and $t+1$ have a selected face. We compute

$$\text{F-Temp}_n = \frac{1}{|\mathcal{P}_n|} \sum_{t \in \mathcal{P}_n} (e_t^g)^\top e_{t+1}^g, \qquad \text{F-Temp} = \frac{1}{|\mathcal{S}_{\text{Temp}}|} \sum_{n \in \mathcal{S}_{\text{Temp}}} \text{F-Temp}_n, \tag{24}$$

where $\mathcal{S}_{\text{Temp}}$ contains videos with an eligible source prototype and at least one valid adjacent-frame pair. F-Temp has no 41-frame minimum on generated face detections, and each valid video receives equal weight in the method-level average.

Face inter-frame similarity. F-IFS uses the largest face in the first source frame as the identity reference and selects the most similar detected face in each generated frame. A generated video must have at least 41 frames with a selected face to be eligible. We compute cosine similarity over adjacent original frames with valid selections, using the same pairwise form as F-Temp. For the method-level F-IFS score, however, we pool all valid pairs from that method's eligible videos and take their mean. Thus, videos with more valid pairs receive more weight. The valid pairs are determined separately for each method, without taking an intersection across methods.

Subject motion preservation. We use CoTracker3 Karaev et al. (2025) trajectories and remove global background/camera motion with a RANSAC-estimated partial-affine transform. Subject points are selected from source-video residual motion, and the source-derived sampling mask is reused for the generated video. Let A_t be the estimated global transform, Q_t the visible subject points, and $x_{i,t}$ the position of point i at time t. The residual motion vector and amplitude are

$$\begin{aligned} r_{i,t} &= x_{i,t+1} - A_t(x_{i,t}), \\ v_t &= \frac{\text{median}_{i \in Q_t}\, r_{i,t}}{\sqrt{W^2 + H^2}}, \qquad a_t = \frac{\text{median}_{i \in Q_t}\, \|r_{i,t}\|_2}{\sqrt{W^2 + H^2}}, \end{aligned} \tag{25}$$

where W and H are the image width and height. The transform and residuals are computed separately for the source and generated videos. A valid tracking step requires at least 12 visible background points, three visible subject points, and a successful global-transform fit.

Trajectory and amplitude fidelity. Let $\mathcal{T}$ contain jointly valid source and generated tracking steps. We accumulate the residual motion vectors over these steps to obtain $P_t^s = \sum_{k \in \mathcal{T},\, k \le t} v_k^s$ and $P_t^g = \sum_{k \in \mathcal{T},\, k \le t} v_k^g$. Trajectory Fidelity compares these cumulative residual trajectories, while Amplitude Fidelity compares the mean residual motion magnitudes:

$$\begin{aligned} \text{Traj.} &= \exp\left(- \frac{\langle \|P_t^g - P_t^s\|_2 \rangle_{t \in \mathcal{T}}}{\max(\langle \|P_t^s\|_2 \rangle_{t \in \mathcal{T}}, 0.01)} \right), \\ \text{Amp.} &= \exp\left(- \left| \log \frac{\langle a_t^g \rangle_{t \in \mathcal{T}} + \epsilon}{\langle a_t^s \rangle_{t \in \mathcal{T}} + \epsilon} \right| \right), \end{aligned} \tag{26}$$

where angle brackets denote an arithmetic mean and $\epsilon = 10^{-6}$. Amplitude Fidelity penalizes both stronger and weaker motion than the source and equals one when the mean amplitudes match.

RAFT Flow Fidelity. The Flow metric reported in the tables uses RAFT-Large Teed & Deng (2020) on videos sampled every four frames. A source-derived subject bounding box defines the shared region of interest. Flow outside this region estimates global partial-affine motion; after compensation, flow inside the region yields median residual vectors u_t^s, u_t^g and median residual magnitudes b_t^s, b_t^g. We combine directional and magnitude agreement:

$$\begin{aligned} d_t &= \text{clip}\left(\frac{1}{2} \left[1 + \frac{(u_t^s)^\top u_t^g}{\|u_t^s\|_2 \|u_t^g\|_2 + \epsilon} \right], 0, 1 \right), \\ m_t &= \exp\left(- \left| \log \frac{b_t^g + \epsilon}{b_t^s + \epsilon} \right| \right), \qquad \text{Flow} = \frac{1}{|\mathcal{T}_f|} \sum_{t \in \mathcal{T}_f} d_t m_t, \end{aligned} \tag{27}$$

where $\mathcal{T}_f$ contains jointly valid RAFT flow steps and $\epsilon = 10^{-6}$.

Validity and aggregation. A step is jointly valid when its source and generated residual motion magnitudes are both finite. Trajectory and Amplitude Fidelity require at least 20 jointly valid tracking steps; Flow Fidelity requires at least eight jointly valid RAFT flow steps. Because RAFT uses

a four-frame sampling interval, these flow steps are not adjacent original-frame pairs. Instances below the corresponding threshold are excluded from that metric. Each method-level score is the arithmetic mean over its valid instances. These metrics measure agreement in camera-compensated image-plane motion; source-derived sampling locations do not establish cross-view physical point correspondences.

B.6 User Study Protocol

Cases and participants. We recruit 14 participants, each of whom evaluates four cases, yielding 56 participant–case evaluations. Cases are drawn from HumanVid, SpatialVID, and classic film sequences, without GT target depth. For each assigned case, the participant rates all four methods: GRAFT, ReCamMaster, ReDirector, and DepthDirector. This within-participant comparison helps reduce the influence of differences in individual rating scales.

Presentation. Figure 13 shows the interface, which presents the source video, a warped camera-motion reference, and four generated results. Methods are anonymized as A–D, with their order randomized for each evaluation. The method-to-label mapping is recorded privately and is accessible only to the organizers. The source video provides the appearance and subject-motion reference, while the warped video indicates the requested camera motion and approximate scene layout.

Rating criteria. Participants assign scores from 1 (poor) to 5 (excellent) for four criteria: **(i)** *Visual quality*: clarity, naturalness, and visible artifacts; **(ii)** *Camera-motion plausibility*: coherent viewpoint changes consistent with the warped reference; **(iii)** *Subject motion preservation*: retention of the source actions and their temporal progression; and **(iv)** *Scene structural integrity*: coherent scene and object structure without collapse, unintended disappearance, or duplication.

Aggregation. For each method and criterion, we compute the arithmetic mean and standard deviation directly over the 56 participant ratings. Table 4 reports the resulting scores.

B.7 Computational Efficiency

Training cost. In the short training-log segment reported for geometry training, ordinary steps take approximately 6–7 seconds and steps with addressing distillation take approximately 14 seconds. With one AD update in each five-step cycle, the implied average is

$$\frac{4(6\text{–}7) + 14}{5} = 7.6\text{–}8.4\ \mathrm{s/step}, \tag{28}$$

or approximately 8 seconds per step. This is a local timing observation, not a measurement of full-run wall-clock time or a controlled efficiency comparison with other methods.

Inference cost. GRAFT uses 30 denoising steps for target-depth generation and 30 denoising steps for RGB generation. The two stages run sequentially, with a total denoising time of 92 seconds per video. This timing measures the denoising stages rather than the complete end-to-end pipeline. AD is used only during training and incurs no inference-time distillation cost.

C Preliminaries

C.1 Flow Matching and Video Diffusion Transformers

From diffusion to flow matching. Denoising diffusion models Ho et al. (2020) learn to reverse a noise corruption process. Deterministic sampling is also possible through probability-flow ODEs Song et al. (2021b) and DDIM Song et al. (2021a). Flow matching Lipman et al. (2023) and rectified flow Liu et al. (2023) instead train a velocity field along prescribed interpolation paths. Here, we use $t = 0$ for noise and $t = 1$ for data. Let $\mathbf{x}_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ be noise and $\mathbf{x}_1$ a clean target latent. The linear interpolation and its conditional velocity are

$$\mathbf{x}_t = (1-t)\mathbf{x}_0 + t\mathbf{x}_1, \qquad \mathbf{u}_t = \mathbf{x}_1 - \mathbf{x}_0, \qquad t \in [0,1]. \tag{29}$$

Given conditioning inputs c, a velocity predictor $\mathbf{v}_\psi$ with parameters ψ is trained using

$$\mathcal{L}_{\mathrm{FM}}(\psi) = \mathbb{E}_{t,\mathbf{x}_0,\mathbf{x}_1,c} \left\| \mathbf{v}_\psi(\mathbf{x}_t, t, c) - (\mathbf{x}_1 - \mathbf{x}_0) \right\|_2^2 . \tag{30}$$

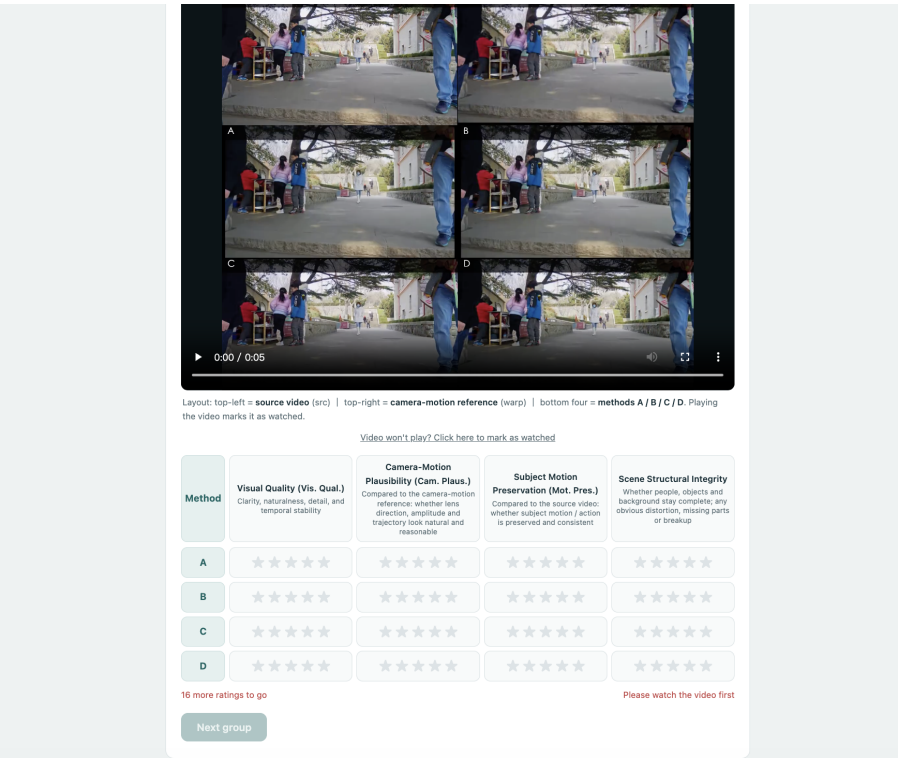


Figure 13: **User study interface.** The panel displays the source video at the upper left, the warped camera-motion reference at the upper right, and four anonymized generated results labeled A–D below, with method order randomized for each evaluation. Participants rate each result on visual quality, camera-motion plausibility, subject motion preservation, and scene structural integrity using a five-point scale. The warped reference conveys the requested camera change and is not ground-truth target video.

Sampling integrates $\mathrm{d}\mathbf{x}_t/\mathrm{d}t = \mathbf{v}_\psi(\mathbf{x}_t, t, c)$ from $t = 0$ to $t = 1$.

At the pure-noise endpoint, $\mathbf{x}_t = \mathbf{x}_0$ contains no residual clean target content. Two experts with matching target latent shapes can therefore receive exactly the same noise input despite predicting different quantities. Away from this endpoint, their interpolated inputs generally differ because their clean targets differ. This distinction underlies the probing condition used in Section 3.2.

Latent video DiTs. Latent diffusion Rombach et al. (2022) performs generation in an autoencoder's compressed representation rather than directly in pixel space. A video VAE Kingma & Welling (2014); Wan et al. (2025) compresses both spatial and temporal dimensions. The resulting latent is partitioned into tokens and processed by a Diffusion Transformer (DiT) Peebles & Xie (2023), whose blocks combine attention and feed-forward operations with timestep conditioning. GRAFT uses the same video VAE for RGB and color-encoded depth, giving the two prediction tasks matching token layouts. Each token represents a compressed spatiotemporal region rather than an individual pixel in an original video frame.

C.2 Video Retaking and Geometric Reprojection

Video retaking Bai et al. (2025); Yu et al. (2025); Chen et al. (2026); Park et al. (2026) changes the viewpoint of a source video while preserving its content and motion. Text-to-video camera control Wang et al. (2024b); He et al. (2025); Bahmani et al. (2025) specifies camera motion without requiring preservation of an observed source event; reference-based control can instead obtain camera motion from another video Luo et al. (2025). Geometric reprojection, used in rendering-guided approaches Ren et al. (2025); Yu et al. (2025); Chen et al. (2026), also provides the basis for GRAFT's first-frame depth reference.

For a pinhole camera, let $\mathbf{K}_s, \mathbf{K}_t$ denote source and target intrinsics, and let $(\mathbf{R}_s, \mathbf{t}_s)$ and $(\mathbf{R}_t, \mathbf{t}_t)$ be world-to-camera poses. Given a source pixel $\mathbf{u} = (u, v)$ with homogeneous coordinates $\tilde{\mathbf{u}} = (u, v, 1)^\top$ and optical-axis depth $D_s(\mathbf{u})$, its world-space point is

$$\mathbf{p}^{\mathrm{w}}(\mathbf{u}) = \mathbf{R}_s^\top \left(D_s(\mathbf{u}) \mathbf{K}_s^{-1} \tilde{\mathbf{u}} - \mathbf{t}_s \right). \tag{31}$$

Transforming this point into the target camera gives its projected pixel and target depth:

$$\mathbf{p}^{\mathrm{t}} = \mathbf{R}_t \mathbf{p}^{\mathrm{w}} + \mathbf{t}_t, \qquad \tilde{\mathbf{u}}' \propto \mathbf{K}_t \mathbf{p}^{\mathrm{t}}, \qquad z' = (\mathbf{p}^{\mathrm{t}})_3. \tag{32}$$

Here, $(\mathbf{p}^{\mathrm{t}})_3$ is the third coordinate, and homogeneous image coordinates are normalized by their third component. Rasterizing valid projections with visibility handling produces a partial target-view depth map and a validity mask. Newly visible regions without source observations remain unfilled. This operation describes geometric reference construction for the different-initial-viewpoint setting; it does not construct the addressing supervision. The experiment-specific reference inputs are discussed in Appendix B.2.

C.3 Camera Geometry and Unified Camera Positional Encoding

Viewing rays and relative encodings. For a pinhole camera with intrinsics $\mathbf{K}$ and world-to-camera pose $(\mathbf{R}, \mathbf{t})$, pixel $\mathbf{u}$ defines a ray with origin $\mathbf{o} = -\mathbf{R}^\top \mathbf{t}$ and unit direction $\mathbf{d} = \mathbf{R}^\top \mathbf{K}^{-1} \tilde{\mathbf{u}} / \|\mathbf{R}^\top \mathbf{K}^{-1} \tilde{\mathbf{u}}\|_2$. Its Plücker representation is

$$\boldsymbol{\pi} = (\mathbf{d}, \mathbf{o} \times \mathbf{d}) \in \mathbb{R}^6. \tag{33}$$

Light Field Networks Sitzmann et al. (2021) use this representation for ray-based scene modeling, and CameraCtrl He et al. (2025) uses it for camera conditioning. Plücker coordinates describe a 3D line; the camera model determines how pixels map to rays. Relative encodings instead incorporate geometric relationships within attention: GTA Miyato et al. (2024) uses geometric transformations, PRoPE Li et al. (2026a) includes camera intrinsics and extrinsics through projective transformations, and ReDirector's RoCE Park et al. (2026) applies camera-conditioned RoPE phase shifts.

Relative ray encoding. UCPE Zhang et al. (2026) assigns each token i a world-space ray origin $\mathbf{o}_i$ and unit direction $\mathbf{d}_i$. Rays can be obtained from different projection models, including the unified camera model Mei & Rives (2007). Let $\mathbf{b}_i$ be the corresponding camera's downward axis in world coordinates. For nonparallel $\mathbf{b}_i$ and $\mathbf{d}_i$, a normalized local basis is

$$\mathbf{e}_{z,i} = \mathbf{d}_i, \qquad \mathbf{e}_{x,i} = \frac{\mathbf{b}_i \times \mathbf{e}_{z,i}}{\|\mathbf{b}_i \times \mathbf{e}_{z,i}\|_2}, \qquad \mathbf{e}_{y,i} = \mathbf{e}_{z,i} \times \mathbf{e}_{x,i}. \tag{34}$$

The rotation $\mathbf{R}_i^{\mathrm{wr}} = [\mathbf{e}_{x,i}, \mathbf{e}_{y,i}, \mathbf{e}_{z,i}]$ and origin $\mathbf{o}_i$ define a ray-to-world transform $\mathbf{T}_i^{\mathrm{wr}}$. Its inverse $\mathbf{T}_i^{\mathrm{rw}}$ maps world coordinates into the local ray frame and supplies the geometric transformation used within attention.

Orientation and injection. UCPE also provides absolute orientation cues through a latitude-up map. Latitude describes ray elevation relative to the world horizontal plane; the up component describes the image-plane direction induced by perturbing the ray toward world up. These cues represent pitch and roll. UCPE combines ray transforms with RoPE in a spatial attention adapter. GRAFT uses target-only UCPE for the geometry expert's camera conditions $\mathcal{C}$; the appearance expert receives target depth rather than camera conditions directly.

C.4 Attention Transfer

Attention Transfer Zagoruyko & Komodakis (2017) supervises a student using spatial attention maps derived from a teacher's CNN activations. Knowledge distillation Hinton et al. (2015) transfers teacher output distributions. GRAFT uses target-to-source attention distributions as the transferable signal.

For one target query, let p_j and q_j denote the teacher and student probabilities over source token j, with the teacher held fixed. Their forward KL divergence is

$$D_{\mathrm{KL}}(p \| q) = \sum_{j=1}^{N_{\mathrm{src}}} p_j \log \frac{p_j}{q_j}. \tag{35}$$

Minimizing this loss is equivalent to minimizing the teacher-weighted cross-entropy $-\sum_j p_j \log q_j$, since $\sum_j p_j \log p_j$ is constant with respect to the student. The supervision can therefore retain weight on multiple source regions and frames without selecting a single correspondence. GRAFT applies this objective to head-averaged addressing distributions and aggregates it over target queries and selected blocks, as specified in Equation 3.

D AI USE STATEMENT

We used large language models to assist with manuscript editing and language polishing, including improving clarity, grammar, and organization. The models were not used for generating experimental results, designing algorithms, or producing scientific claims.

E ETHICAL STATEMENT

All datasets used in this work are publicly available and do not contain personally identifiable information collected by the authors. Our experiments are conducted on existing benchmarks and publicly released datasets. The human evaluation study, when applicable, was conducted with participants' informed consent, and all responses were collected anonymously. We follow responsible AI practices and acknowledge potential limitations of generative video models, including possible artifacts and biases inherited from training data.

F REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we provide detailed descriptions of the model architecture, training configuration, evaluation protocols, and implementation details. We report datasets, metrics, and experimental settings used in our evaluations. Code and pretrained models will be released upon publication to enable further research and verification.

G LIMITATIONS AND FUTURE WORK

Depth provides limited cues for dynamics with little geometric variation, such as motion on lake surfaces. Such scenes remain challenging for GRAFT and motivate richer intermediate representations. Joint prediction of depth and optical flow is a promising direction for capturing both scene geometry and fine-grained motion. Our training data are also limited in scale and scene diversity, and their synthetic origin introduces a domain gap to real-world videos. Expanding the training set and incorporating more diverse real-world observations may improve robustness across scenes and motion patterns.