TTA4D: OBJECT-CENTRIC FREE-VIEWPOINT VIDEO GENERATION VIA ONE-SHOT TEST-TIME ADAPTATION

Anonymous authors
Paper under double-blind review

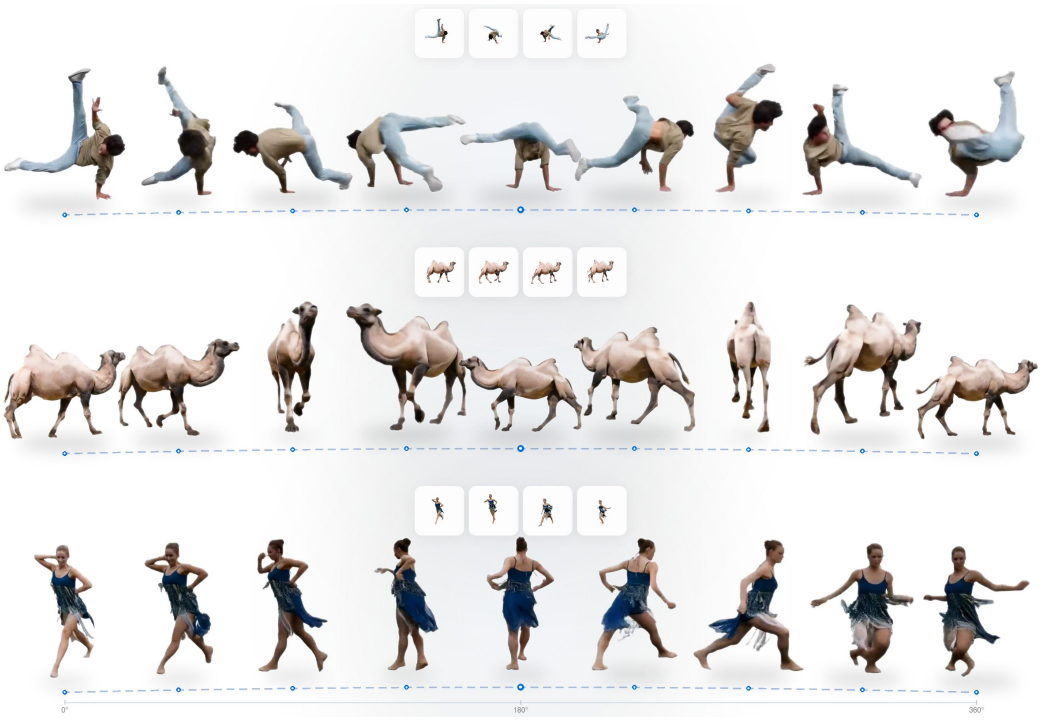

Figure 1: **Free-viewpoint video from a single monocular video on real-world DAVIS.** Each row shows input frames (thumbnails, top) and the video TTA4D generates along a full 360° object-centric orbit (bottom). TTA4D adapts a plain image-to-video model (Wan2.2-I2V, with SS4D's coarse voxels as geometry) to each object in approximately 10–15 minutes on a single GPU and preserves the input's appearance and motion from viewpoints the input never observed. Comparisons with baselines are given in Fig. 3 and Table 2.

ABSTRACT

Generating free-viewpoint videos of a dynamic object from a single monocular video requires temporal coherence and view-consistent geometry, yet explicit 4D methods often distort geometry on in-the-wild inputs, and controllable video models drift from the input's appearance under large viewpoint changes. We present **TTA4D**, a one-shot test-time adaptation framework that converts a pre-trained image-to-video diffusion model into an object-specific free-viewpoint video generator in approximately 10–15 minutes on a single GPU. Our key observation is that the missing supervision, a moving object seen from a moving camera, factorizes into two readily available signals: a *dynamic-object fixed-camera* signal from the input video, which teaches the model to follow geometric conditioning as the object moves, and a *static-object dynamic-camera* signal rendered from a single-image 3D reconstruction, which teaches it to follow conditioning as the viewpoint changes. After LoRA adaptation on both, the model composes them along user-

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.

specified object-centric trajectories while preserving the input's appearance and motion. Because TTA4D adds only LoRA weights, without architectural changes or paired control data, it is directly applicable to newly released video foundation models. With this adaptation, a plain image-to-video model outperforms the depth-controllable VACE given the same geometry, and adapting VACE itself yields further improvement. On Consistent4D, TTA4D with the VACE backbone outperforms six baselines on all four metrics for both seen and unseen trajectories, and its outputs can directly supervise a 4D Gaussian Splatting reconstruction. On 22 real-world DAVIS videos, even its weakest configuration, a plain image-to-video model with coarse voxel geometry, receives a mean rating of $4.56/5$ in a blinded user study, compared with 3.26 for the strongest baseline.

1 INTRODUCTION

Generating free-viewpoint videos of a dynamic object from a single monocular video has been approached along two main lines. The first constructs an explicit 4D representation, either by distilling diffusion priors or by feed-forward prediction learned from curated 4D data (Jiang et al., 2024; Ren et al., 2024; Li et al., 2025c); such representations are view-consistent by construction, but frequently exhibit geometric distortion and loss of appearance detail on in-the-wild inputs. The second re-renders the input with camera- or depth-controlled video models (Yu et al., 2025b; Ren et al., 2025; Jiang et al., 2025), which synthesize realistic frames but, being trained for zero-shot control on large paired datasets, tend to deviate from the appearance of the specific object and to violate its geometry under large viewpoint changes. Consequently, neither line fully meets the requirements of applications such as VR/AR (Azuma, 1997), content creation, and world simulation (NVIDIA et al., 2025; Cao et al., 2025), which call for videos that remain faithful to the appearance and motion of a given object from arbitrary viewpoints. What is missing is *object-specific geometry following*: a video model that follows arbitrary geometric conditioning while remaining faithful to a particular object.

Video foundation models (Ho et al., 2022; OpenAI, 2024; Google DeepMind, 2024; Kong et al., 2024; Team Wan et al., 2025) are a natural starting point for this capability, as they already encode rich priors on object motion and view-dependent appearance. The central difficulty lies in supervision: the target configuration, a *moving* object viewed from a *moving* camera, is never observed. Our key observation is that this supervision factorizes into two components, each of which is readily available. The input video, together with depth and masks estimated from it, characterizes how the geometric conditioning evolves as the object moves under a fixed camera. Renderings of a single-image 3D reconstruction (Xiang et al., 2025a) characterize how it evolves as the camera moves around a static object, including surfaces that are not visible in the input.

Building on this observation, we introduce TTA4D, a one-shot test-time adaptation framework (Fig. 2). Given a monocular video, TTA4D constructs a *dynamic-object fixed-camera* (DOFC) signal from the input and a *static-object dynamic-camera* (SODC) signal from the 3D renderings, both expressed as a reference image with per-frame depth and masks, and adapts a pre-trained video diffusion model on them with LoRA (Hu et al., 2022) in approximately 10–15 minutes. Although neither signal contains a moving object viewed from a moving camera, the adapted model composes the two: at inference, it follows depth and masks rendered along a user-specified trajectory from off-the-shelf dynamic geometry, such as the coarse voxels of SS4D (Li et al., 2025c) or the more accurate World Tracing (Zhang et al., 2026a). On a plain image-to-video backbone, adaptation on the input video alone leaves PSNR on novel trajectories nearly unchanged, whereas incorporating SODC improves it by 4.5 dB (Section 4.4).

TTA4D requires only a standard image-to-video model. Following LoRA-Edit (Gao et al., 2026), it reuses the built-in image conditioning of the model and introduces only LoRA weights; newly released video foundation models can therefore be adopted directly, without architectural modifications or a controllable variant trained on paired data. Moreover, accurate geometry alone is insufficient: given identical World Tracing geometry and reference image, the depth-controllable VACE (Jiang et al., 2025) follows the silhouette of the object but not its appearance (LPIPS 0.129, compared with 0.091 after adaptation), and adapting such a model, when available, yields the best

results. Finally, the generated views are sufficiently consistent to directly supervise a 4D Gaussian Splatting reconstruction (Section 4.5).

We evaluate TTA4D on Consistent4D, which provides ground-truth novel views, and on 22 real-world DAVIS videos through a blinded user study. Our contributions are as follows:

1. TTA4D, a one-shot test-time adaptation framework that converts a pre-trained image-to-video model into an object-specific, geometry-conditioned free-viewpoint video generator in approximately 10–15 minutes, without multi-view supervision, architectural modifications, or paired control data.
2. A factorized dual-signal supervision in which DOFC and SODC each provide one factor of the unobservable dynamic-object, dynamic-camera supervision, and which the adapted model composes to follow trajectories not observed during adaptation.
3. A comprehensive evaluation demonstrating that (i) a plain image-to-video model adapted with TTA4D outperforms a depth-controllable model given the same geometry; (ii) TTA4D with a VACE backbone outperforms six baselines on all four metrics for both seen and unseen trajectories on Consistent4D, and its outputs can directly supervise a 4DGS reconstruction; and (iii) on DAVIS, its weakest configuration receives a mean rating of $4.56/5$, compared with 3.26 for the strongest baseline.

2 Related Work

3D and 4D generation. 3D generation ranges from score distillation (Poole et al., 2023; Lin et al., 2023) to feed-forward predictors (Liu et al., 2023; Shi et al., 2023; Voleti et al., 2024; Xu et al., 2024; Tang et al., 2024; Xiang et al., 2025b;a), which are now sufficiently mature to provide our SODC signal. Monocular dynamic reconstruction (Gao et al., 2021; Tian et al., 2023; 2024) and 4D shape reconstruction (Cao et al., 2024) recover the observed dynamics but do not generate unobserved appearance. Optimization-based 4D generation (Singer et al., 2023; Bahmani et al., 2024b;a; Jiang et al., 2024; Ren et al., 2023; Zhao et al., 2023) distills diffusion priors into dynamic NeRF or Gaussian representations (Mildenhall et al., 2020; Wu et al., 2024); it is computationally expensive and inherits the mode-seeking behavior of score distillation. Feed-forward and multi-view-diffusion 4D methods (Ren et al., 2024; Zhang et al., 2025a; Litman et al., 2026; Xie et al., 2025; Yao et al., 2025; Wu et al., 2025; Zhang et al., 2025c; 2026b; Hu et al., 2025; Fan et al., 2025; Song et al., 2026) are faster, but most depend on curated multi-view or 4D data. SV4D (Xie et al., 2025) outputs a fixed set of viewpoints rather than continuous trajectories; we therefore do not include it as a baseline. TTA4D's inference geometry can be obtained from any 3D or 4D estimator (Wang et al., 2025; Zhou et al., 2026); we use SS4D's coarse voxels (Li et al., 2025c) and World Tracing (Zhang et al., 2026a), which also predicts surfaces occluded in the input view.

Camera- and geometry-controlled video generation. Video diffusion models (Blattmann et al., 2023; Yang et al., 2025; Team Wan et al., 2025) have been extended with camera control (He et al., 2025; Wang et al., 2024; Bahmani et al., 2025; Hou & Chen, 2025; Li et al., 2025b; Kuang et al., 2024; Li et al., 2025a; Luo et al., 2025), with 3D-informed rendering from depth or point clouds (Ren et al., 2025; Yu et al., 2025b;c; Bai et al., 2025; Van Hoorick et al., 2024; Zhang et al., 2025d; Liu et al., 2026; Tian et al., 2025; Yu et al., 2024; 2025a), and with versatile control from depth, masks, and reference images (Jiang et al., 2025). Trained for zero-shot control on large paired corpora, they can produce inconsistent geometry under large viewpoint changes and drift toward the appearance style of their training data. We compare with TrajectoryCrafter and GEN3C.

Per-instance adaptation of generative models. LoRA (Hu et al., 2022) is widely used to personalize generative models, *e.g.*, DreamBooth (Ruiz et al., 2023) for subject identity, Tune-A-Video (Wu et al., 2023) for one-shot text-to-video generation, and LoRA-Edit (Gao et al., 2026) for first-frame-guided video editing. Most closely related to our work, ReCapture (Zhang et al., 2025b) and Reangle-A-Video (Jeong et al., 2025) fine-tune a video model on one input video to re-render it along new camera paths, supervised mainly by anchor or warped videos from point-cloud reprojection with losses restricted to valid pixels (ReCapture additionally uses multi-view diffusion for large rotations). They re-render the full scene, including the background, and Reangle-A-Video supports six predefined camera motions. TTA4D instead renders an isolated object along full object-centric

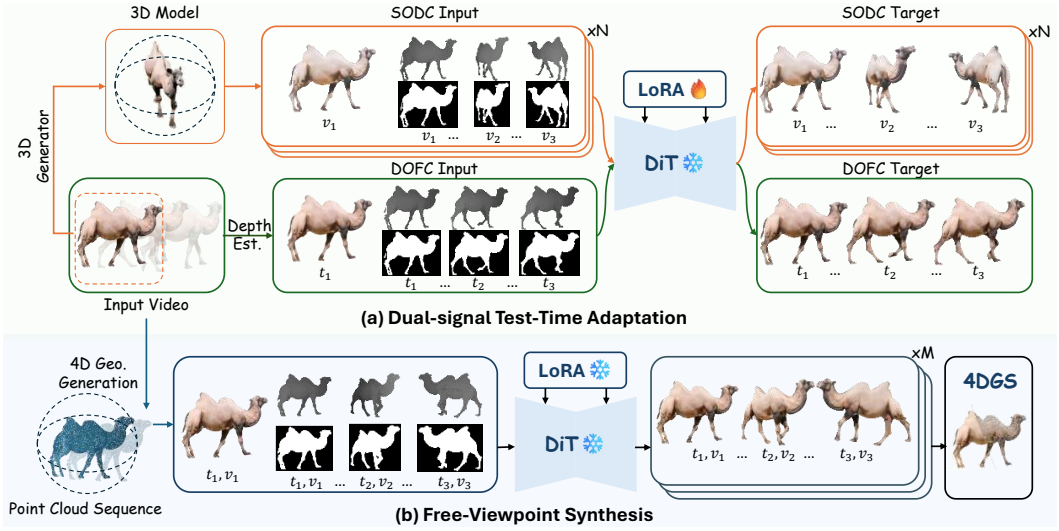

Figure 2: **Overview of TTA4D.** (a) From a single input video, we construct two training signals with a shared conditioning format: a reference image plus per-frame depth and masks. SODC (orange) renders a 3D model generated from the first frame along object-centric paths (stacked boxes; three paths in our experiments) and teaches the model to follow conditioning that changes with the viewpoint, including views of unobserved sides such as the back of the object; DOFC (green) pairs the input video with estimated depth and masks and teaches it to follow conditioning that changes with object motion. The reference image is the first input frame for DOFC and a rendering of the 3D model for SODC. Only LoRA weights of the frozen video diffusion transformer (DiT) are trained. (b) Depth and masks rendered from dynamic geometry (*e.g.*, World Tracing or SS4D's voxels) along any user-specified trajectory condition the adapted model, with the first input frame as the reference. One adapted model serves any number of trajectories (stacked boxes), and its outputs can optionally supervise a 4DGS reconstruction.

orbits, conditions on complete depth-and-mask sequences rather than partially valid warped pixels, and supervises unseen surfaces through SODC; we therefore do not include them as baselines. We call this *one-shot test-time adaptation*: as in one-shot tuning (Wu et al., 2023) and test-time training (Sun et al., 2020), the model is adapted on the single test input at inference time, using only supervision derived from it and never ground-truth target views; unlike entropy-based adaptation under distribution shift (Wang et al., 2021), it also learns a new conditioning task.

3 Method

Given a monocular video $\mathbf{V} = \{I_t\}_{t=1}^{T}$ of a dynamic object, TTA4D generates videos $\hat{\mathbf{V}}^{\Pi}$ along user-specified object-centric camera trajectories $\Pi = \{\pi_t\}_{t=1}^{T}$ (Fig. 2). It extracts geometric conditioning (Section 3.1), builds two training signals from it (Section 3.2), and adapts a pre-trained video diffusion model with LoRA (Section 3.3).

3.1 Geometric Conditioning

All signals share one conditioning format: per-frame depth and object masks. For **DOFC**, we estimate depth with Depth Anything 3 (Lin et al., 2025) and propagate masks with SAM 2 (Ravi et al., 2025), giving $\{D_t^{\text{obs}}, M_t^{\text{obs}}\}_{t=1}^{T}$. For **SODC**, we reconstruct a static 3D model $\mathcal{G}$ from I_1 with TRELLIS.2 (Xiang et al., 2025a), which can be rendered from any object-centric pose:

$$I^{\pi}_{\text{static}},\ D^{\pi}_{\text{static}},\ M^{\pi}_{\text{static}} = \text{Render}(\mathcal{G}, \pi). \tag{1}$$

Because $\mathcal{G}$ is reconstructed from a single image, it synthesizes plausible appearance for unobserved parts, which enables SODC to supervise, and TTA4D to generate, trajectories far beyond the input's viewpoint coverage, including full 360° orbits.

For **inference**, any dynamic geometry that can be rendered into depth and masks can be used, including the coarse first-stage voxels of SS4D (Li et al., 2025c); more accurate geometry yields higher quality. For the Consistent4D results we use World Tracing (Zhang et al., 2026a), whose layered per-frame geometry includes surfaces occluded in the input view; we fuse its valid points into point clouds $\{\mathcal{O}_t\}_{t=1}^{T}$ in a shared object-centric frame and render them by point splatting. The DAVIS evaluation uses SS4D's voxels, rendered into the same format. For any pose π,

$$(D_t^{\pi}, M_t^{\pi}) = \mathrm{Render}(\mathcal{O}_t, \pi), \quad t = 1, \ldots, T. \tag{2}$$

This geometry is used only as inference-time conditioning, never as training ground truth. All depth sequences of an object share one fixed encoding, so that depth from different sources, frames, and viewpoints is directly comparable. We first align the depth sources with fixed affine mappings fitted to the 5th and 95th foreground depth percentiles of their first frames, taking the input-view depth as the reference. We then determine a shared range covering the per-frame 0.1%–99.9% foreground quantiles across the object's depth sequences, with a 10% margin on either side. Depths are normalized to this range, clipped, and passed through a fixed monotone contrast curve, with nearer surfaces encoded as brighter values; we write $\bar{D}$ for encoded depth. Masks are supplied as separate binary conditions.

3.2 Dual-Signal Supervision

DOFC uses the input video as supervision, and SODC renders $\mathcal{G}$ along K object-centric trajectories $\{\Pi^k\}_{k=1}^{K}$, with $\mathbf{V}_{\text{static}}^k = \{I_{\text{static}}^{\pi_t^k}\}_{t=1}^{T}$ and its depth and masks from Eq. (1):

$$\mathcal{S}_{\text{DOFC}} = \left(I_1,\ \{\bar{D}_t^{\text{obs}}, M_t^{\text{obs}}\}_{t=1}^{T}\right) \xrightarrow{\text{supervise}} \mathbf{V}, \tag{3}$$

$$\mathcal{S}_{\text{SODC}}^k = \left(I_{\text{static}}^{\pi_1^k},\ \{\bar{D}_{\text{static}}^{\pi_t^k}, M_{\text{static}}^{\pi_t^k}\}_{t=1}^{T}\right) \xrightarrow{\text{supervise}} \mathbf{V}_{\text{static}}^k, \quad k = 1, \ldots, K. \tag{4}$$

DOFC teaches the model to follow conditioning that changes with the object's motion under a fixed camera, while generating the object's appearance, but it provides no viewpoint variation. SODC supplies this variation: although $\mathbf{V}_{\text{static}}^k$ shows a static object, its frame-to-frame viewpoint change emulates camera motion. Each SODC sequence uses its first rendered frame $I_{\text{static}}^{\pi_1^k}$ as the reference, so that reference and supervision come from the same 3D model; DOFC and inference use the real first frame I_1, which anchors the output to the input identity and style.

Why the signals compose. At inference, the rendered conditioning changes with both object motion and camera motion. DOFC teaches the model to follow changes caused by object motion, and SODC those caused by camera motion. Neither signal contains both, but the two share one conditioning interface, which allows the adapted model to follow both simultaneously. We attribute this compositional behavior to the pre-trained video prior, which models object motion and camera motion jointly in natural videos; the ablation in Section 4.4 provides empirical support for this interpretation.

3.3 Test-Time Adaptation via LoRA

TTA4D requires only a pre-trained latent image-to-video diffusion model (Rombach et al., 2022; Team Wan et al., 2025); a depth-controllable model such as VACE (Jiang et al., 2025) can be used but is not required. For the j-th training sequence, with target latent $\mathbf{z}_0^j = \mathcal{E}(\mathbf{V}_{\text{target}}^j)$ and noised latent $\mathbf{z}_s^j = \alpha_s \mathbf{z}_0^j + \sigma_s \boldsymbol{\epsilon}$, $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, the backbone f_θ predicts its native target $\mathbf{y}_s^j$ (the flow-matching velocity for Wan) given the conditioning $\mathbf{c}_j = \{I_{\text{ref}}, \{\bar{D}_t, M_t\}_{t=1}^{T}\}$. For VACE, depth, masks, and the reference image enter through its native context inputs. For the image-to-video (I2V) backbone, which has no depth input, we follow LoRA-Edit (Gao et al., 2026) and reuse the built-in image conditioning without architectural changes: Wan-I2V conditions the denoiser on a pseudo-video $\mathbf{V}_{\text{cond}}$, whose first frame is the reference image and whose remaining frames are zero placeholders, together with a binary mask $\mathbf{M}_{\text{cond}}$ that marks the given frames. We replace the placeholders with the encoded depth maps $\{\bar{D}_t\}$ and supply the object masks through $\mathbf{M}_{\text{cond}}$; adaptation trains the model to interpret these frames as geometric guidance rather than as content to be reproduced. As in LoRA-Edit, the text prompt is a fixed special token followed by a caption of I_1 generated by CogVLM2 (Hong et al., 2024).

Table 1: **Comparison with state-of-the-art methods on Consistent4D.** Seen trajectories are used for SODC adaptation, and unseen trajectories are not (Section 4.1). TTA4D denotes the full configuration, with World Tracing geometry and a VACE backbone (TTA4D + WT + VACE); other configurations are analyzed in Table 3. The indented +4DGS row renders a 4D Gaussian representation optimized on TTA4D videos, whose supervision views include the four fixed azimuths; no ground truth is used (Section 4.5). Cell colors mark the top three per column; tied values share a color.

Method	Seen Trajectories				Unseen Trajectories			
	PSNR↑	SSIM↑	LPIPS↓	IoU↑	PSNR↑	SSIM↑	LPIPS↓	IoU↑
GVFDiffusion	16.866	0.914	0.112	0.498	15.522	0.902	0.132	0.429
L4GM	19.011	0.913	0.109	0.658	19.536	0.914	0.107	0.696
Lift4D	18.232	0.892	0.116	0.645	18.687	0.891	0.114	0.679
SS4D	18.351	0.904	0.105	0.641	17.589	0.896	0.113	0.626
TrajectoryCrafter	17.688	0.893	0.142	0.615	17.153	0.877	0.150	0.625
GEN3C	18.199	0.895	0.213	0.659	16.354	0.858	0.254	0.610
TTA4D	20.842	0.926	**0.084**	**0.747**	20.482	0.916	0.098	0.740
+ 4DGS	**20.901**	**0.932**	**0.084**	0.744	**20.571**	**0.924**	**0.096**	**0.750**

We freeze all pre-trained weights θ and add trainable low-rank updates (Hu et al., 2022) to the self- and cross-attention projections: for each $W \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$,

$$W' = W + \frac{\alpha}{r} BA, \qquad B \in \mathbb{R}^{d_{\text{out}} \times r},\ A \in \mathbb{R}^{r \times d_{\text{in}}},\ r \ll \min(d_{\text{in}}, d_{\text{out}}), \tag{5}$$

with scaling factor α. We optimize only $\phi = \{A_i, B_i\}_i$, where $\theta(\phi)$ denotes the backbone with these updates applied, and weight the DOFC sequence and each SODC sequence equally:

$$\mathcal{L} = \frac{1}{K+1} \sum_{j=0}^{K} \mathcal{L}_j, \qquad \mathcal{L}_j = \mathbb{E}_{\boldsymbol{\epsilon},\, s} \left[\left\| \mathbf{y}_s^j - f_{\theta(\phi)}(\mathbf{z}_s^j, s, \mathbf{c}_j) \right\|_2^2 \right], \tag{6}$$

where $j{=}0$ denotes DOFC and $j{=}k$ the k-th SODC sequence. At inference, with reference $I_{\text{ref}} = I_1$ and depth and masks rendered along the target trajectory via Eq. (2), the adapted model generates $\hat{\mathbf{V}}^{\Pi} = \text{Denoise}_{\theta(\phi^\star)}(I_1, \{\bar{D}_t^{\pi_t}, M_t^{\pi_t}\}_{t=1}^{T})$. One adapted model serves any number of trajectories for the same video; each additional trajectory requires only rendering its conditioning and a single sampling run.

4 Experiments

4.1 Experimental Setup

Implementation. Following the LoRA protocol of LoRA-Edit (Gao et al., 2026), we use Wan2.2-I2V-A14B (Team Wan et al., 2025) (denoted I2V) and Wan2.2-VACE-Fun-A14B, a VACE-style control model built on Wan2.2 (Jiang et al., 2025; Alibaba PAI, 2025). Adaptation uses one DOFC and $K{=}3$ SODC sequences (`spin360`, `arc left`, `arc right`) of 24 frames, with 100 steps per sequence (400 in total), LoRA rank 32, AdamW, and learning rate 10^{-4} (full settings in Section A), which takes approximately 10–15 minutes per object on one NVIDIA RTX 6000 Ada GPU. On DAVIS, we deliberately use the weakest configuration, Wan2.2-I2V with SS4D's coarse voxels and neither World Tracing nor VACE, so that the results isolate the contribution of test-time adaptation. On Consistent4D, we call the full configuration (World Tracing geometry, VACE backbone) TTA4D and name other configurations by geometry and backbone, *e.g.*, TTA4D (SS4D + I2V); on DAVIS, TTA4D refers to TTA4D (SS4D + I2V).

Data, metrics, and baselines. On Consistent4D (Jiang et al., 2024), whose evaluation set provides ground-truth novel views, we report PSNR, SSIM, LPIPS, and foreground IoU with masks extracted by BiRefNet (Zheng et al., 2024); the ablations use the same objects. The 22 real-world DAVIS sequences (Perazzi et al., 2016; Pont-Tuset et al., 2017), selected to contain a dominant foreground object without large occlusions, have no ground-truth novel views; on them, we conduct qualitative comparisons and a blinded user study. We compare with GVFDiffusion (Zhang et al.,

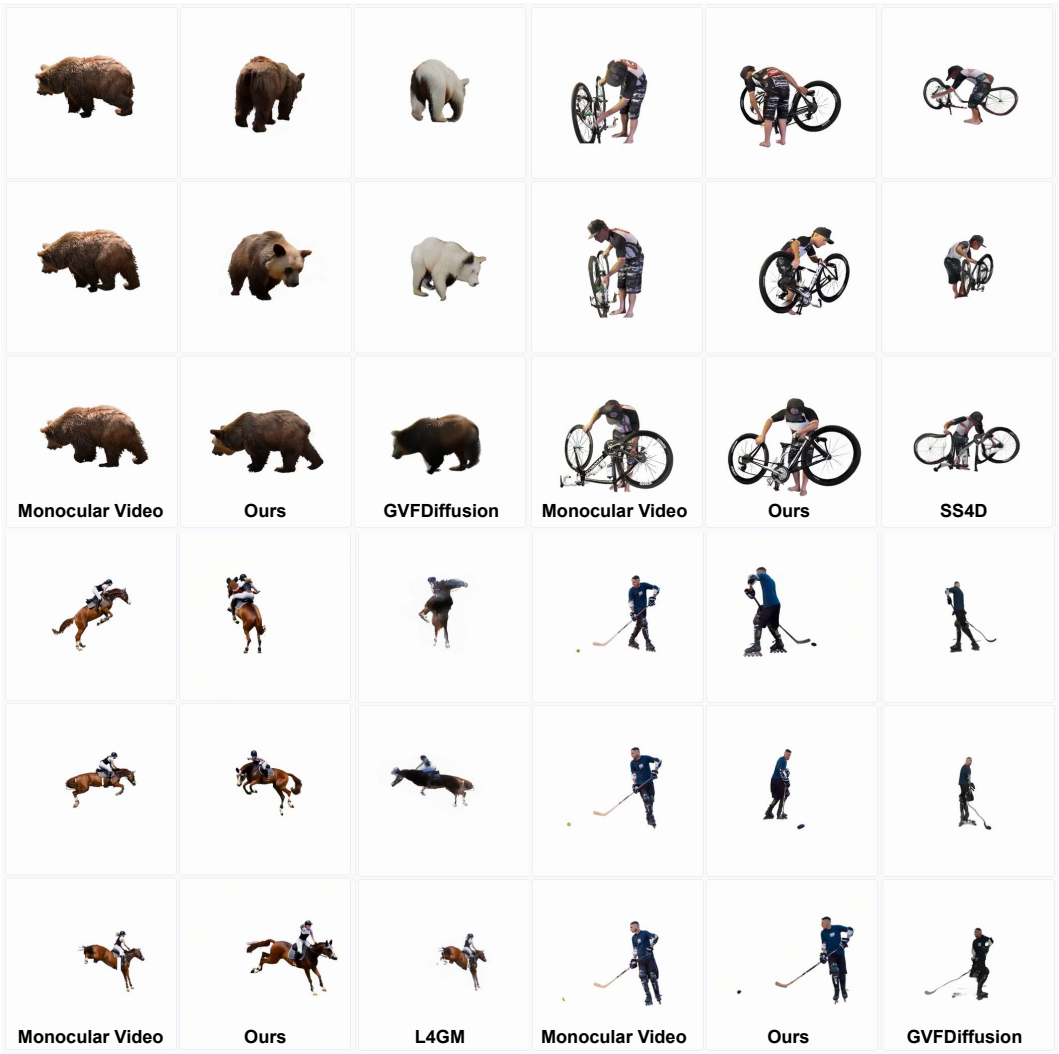

Figure 3: **Qualitative comparison on DAVIS.** `bear` (top left), `bike-packing` (top right), `horsejump-high` (bottom left), and `hockey` (bottom right). Each panel shows the input video, TTA4D (ours), and one baseline (GVFDiffusion, SS4D, L4GM, and GVFDiffusion, respectively) at frames 0, 8, and 23 (top to bottom) along an approximate `spin360` trajectory. TTA4D uses its weakest configuration (Wan2.2-I2V with SS4D's coarse voxels). For compactness, each sequence is paired with one baseline; comparisons against all six baselines on 16 DAVIS sequences are provided in Section F. Camera poses and timestamps are not exactly matched across methods.

2025a) and L4GM (Ren et al., 2024) (feed-forward 4D), SS4D (Li et al., 2025c) (native 4D generation), Lift4D (Litman et al., 2026) (in-the-wild 4D reconstruction), and TrajectoryCrafter (Yu et al., 2025b) and GEN3C (Ren et al., 2025) (camera-controlled video, given the same upstream geometry as TTA4D), using official implementations on the same inputs and trajectories. The three *seen* trajectories are those used for SODC; the five *unseen* ones are `spin180` and fixed azimuths of $-75°$, $15°$, $105°$, and $195°$. For each object, one adapted model is evaluated on all eight trajectories; even on the seen trajectories, SODC supervises only the static 3D model and never the moving object, so every evaluation trajectory requires generating the moving object from unsupervised viewpoints.

4.2 Quantitative Comparison on Consistent4D

TTA4D outperforms all six baselines on all four metrics for both trajectory groups (Table 1). Its PSNR exceeds that of the strongest baseline, L4GM, by 1.831 dB on seen and 0.946 dB on un-

seen trajectories. Seen and unseen scores are close, suggesting that adaptation does not overfit to the SODC trajectories. GEN3C degrades most from seen to unseen trajectories (18.199 to 16.354 PSNR), whereas TTA4D decreases by only 0.36 dB.

4.3 Qualitative Comparison and User Study

Figure 3 shows four DAVIS sequences. TTA4D preserves the bear's fur and body shape, the person–bicycle structure in `bike-packing`, the horse–rider configuration in `horsejump-high`, and the player in `hockey`, whereas the baselines exhibit appearance changes (*e.g.*, GVFDiffusion's pale, desaturated bear) or structural distortions under viewpoint change; Section F compares all six baselines on 16 DAVIS sequences, including these four. In a blinded user study, 20 participants each rated every video, seeing the input alongside all methods' anonymized, randomly ordered outputs for the same trajectory, on a 1–5 scale for overall quality, smoothness, style consistency (faithfulness to the input), and object consistency (stability of identity and structure); the full protocol is in Section E. TTA4D is rated highest on every axis (Table 2), exceeding the best baseline by more than one point on each; the margins are largest for style and object consistency, consistent with the drift and distortions visible in Fig. 3. The camera-controlled video models are rated lowest (*e.g.*, 2.15 overall for TrajectoryCrafter). All DAVIS comparisons use TTA4D's weakest configuration, which uses neither World Tracing nor a controllable backbone, so these margins are attributable to test-time adaptation.

4.4 Ablation Study

The two signals compose. On Wan2.2-I2V (Table 3), DOFC-only adaptation improves LPIPS (0.380 to 0.195) but leaves PSNR nearly unchanged (15.147 to 15.289): following conditioning under the input camera does not transfer to viewpoint changes. Adding SODC raises PSNR to 19.767 and IoU from 0.673 to 0.739, and lowers LPIPS to 0.109. Since neither signal contains a moving object seen from a moving camera, this gain arises from composing what the model learns from each signal: it is SODC that converts a plain image-to-video model into a geometry-following free-viewpoint generator. On VACE, DOFC-only adaptation already yields improvements (19.585 PSNR, 0.111 LPIPS), and adding SODC improves both further (20.662, 0.091).

Geometry alone is insufficient, and a controllable model is not required. Each group in Table 3 holds the inference geometry fixed. With SS4D's coarse voxels, TTA4D (SS4D + I2V) replaces the second stage of the full SS4D pipeline and outperforms it in PSNR (19.018 vs. 17.970), SSIM (0.912 vs. 0.900), and IoU (0.709 vs. 0.634), at slightly higher LPIPS (0.115 vs. 0.109). With the same backbone, replacing these voxels with World Tracing geometry raises PSNR to 19.767 and IoU to 0.739 and lowers LPIPS to 0.109, at comparable SSIM (0.911 vs. 0.912). With World Tracing geometry, zero-shot VACE receives the same conditioning and reference image as TTA4D (WT + VACE) and closely follows the silhouette (IoU 0.723 vs. 0.744), but its appearance error is substantially higher (LPIPS 0.129 vs. 0.091, a 29% reduction after adaptation, with PSNR rising by 1.5 dB). DOFC-only adaptation recovers part of this gain by anchoring the model to the input's

Table 2: **User study on DAVIS.** Mean 1–5 ratings over 20 participants (higher is better). TTA4D uses its weakest configuration.

Method	Overall	Smooth.	Style	Object
GVFDiffusion	3.05	3.12	2.98	3.04
L4GM	3.18	3.05	3.24	3.16
SS4D	3.26	3.31	3.08	3.22
Lift4D	3.08	3.15	2.95	3.02
TrajectoryCrafter	2.15	2.30	2.05	1.95
GEN3C	2.95	3.02	2.85	2.92
TTA4D (Ours)	**4.56**	**4.48**	**4.61**	**4.67**

Table 3: **Ablation on Consistent4D.** Rows within a group share the same geometry source (SS4D voxels or World Tracing, WT). DOFC-only: adaptation on the input video only. Values average the seen and unseen group means (per-group values in Section D).

Method	PSNR ↑	SSIM ↑	LPIPS ↓	IoU ↑
SS4D	17.970	0.900	**0.109**	0.634
TTA4D (SS4D + I2V)	**19.018**	**0.912**	0.115	**0.709**
Zero-shot VACE	19.168	0.903	0.129	0.723
DOFC-only VACE	19.585	0.909	0.111	0.730
TTA4D (WT + VACE)	**20.662**	**0.921**	**0.091**	**0.744**
Zero-shot I2V	15.147	0.847	0.380	0.642
DOFC-only I2V	15.289	0.852	0.195	0.673
TTA4D (WT + I2V)	**19.767**	**0.911**	**0.109**	**0.739**

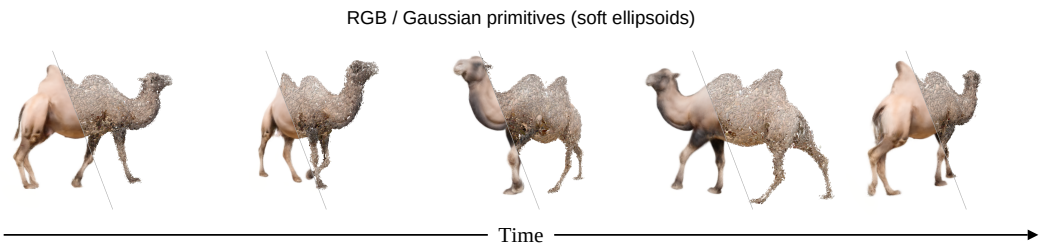

Figure 4: **Downstream 4DGS.** `spin360` renderings of a 4DGS of DAVIS `camel` fitted only to TTA4D (WT + VACE) videos; each frame shows RGB (left) and Gaussians as soft ellipsoids (right).

appearance, and SODC adds the rest by teaching how that appearance changes with viewpoint: zero-shot control provides geometry following but not fidelity to the specific object. Qualitatively, zero-shot VACE renders textures in the style of its training distribution rather than that of the input video (Section C). Wan2.2-I2V, never trained to follow depth or masks, outperforms zero-shot VACE on all four metrics after adaptation, and also outperforms DOFC-only VACE (Table 3). Per-object adaptation thus outperforms the zero-shot use of a model trained on large-scale paired control data, and the two approaches are complementary.

4.5 Downstream 4D Gaussian Reconstruction

Because TTA4D outputs, unlike explicit 4D representations, are not constrained to be mutually consistent, we assess their consistency by fitting a standard 4D Gaussian Splatting model (Wu et al., 2024) to TTA4D videos generated at 24 evenly spaced azimuths at zero elevation, after a static 3D Gaussian initialization of the first frame. These supervision views include the four fixed unseen azimuths but consist only of TTA4D outputs; ground truth is used only for evaluation. Fusing the views preserves their quality (+4DGS in Table 1): scores on seen trajectories are on par (*e.g.*, 20.901 vs. 20.842 PSNR), and all four metrics improve on unseen trajectories (*e.g.*, 20.571 vs. 20.482 PSNR). Because mutually inconsistent views would conflict during fusion and degrade these scores, this result indicates that the generated views are largely consistent; the resulting 4DGS also supports real-time novel-view rendering. Figure 4 applies this procedure to DAVIS `camel`, the only DAVIS result using TTA4D (WT + VACE).

5 Conclusion

This work shows that the missing supervision for free-viewpoint video of a dynamic object, a moving object seen from a moving camera, factorizes into a dynamic-object fixed-camera signal from the input video and a static-object dynamic-camera signal from a single-image 3D reconstruction, which a LoRA-adapted video diffusion model composes after approximately 10–15 minutes of adaptation per object. A plain image-to-video model adapted in this manner outperforms a depth-controllable model given the same geometry, adapting the controllable model yields the best results, and new video foundation models can be adopted without training a controllable variant. TTA4D outperforms six baselines on Consistent4D, produces views sufficiently consistent to supervise 4DGS, and, even in its weakest configuration, is strongly preferred on real-world DAVIS videos. Per-object test-time adaptation of video foundation models is thus a practical alternative to collecting multi-view or 4D training data. More broadly, factorizing supervision that cannot be observed into components that can may benefit other generation tasks for which paired training data are scarce.

Limitations. Adaptation requires minutes per object, compared with seconds for feed-forward methods. Quality depends on the inference geometry and on the first-frame 3D reconstruction behind SODC, which can bias appearance when the first frame is unrepresentative (atypical pose, transient occluders, or later topology change); because this model is generated from a single image, its unseen surfaces are plausible rather than guaranteed to be accurate. We use 24-frame object-centric clips; quantitative evaluation is limited to Consistent4D, and real-world results rely on a user study because no ground-truth novel views exist. Future work includes amortized adaptation, SODC built from multiple frames, longer videos, and full dynamic scenes.

Reproducibility Statement

Section 3 and Section A describe the full pipeline, including the conditioning extraction, the construction of both training signals, the loss, and the main adaptation hyperparameters in Section 4.1 and Section A. All components (Depth Anything 3, SAM 2, TRELLIS.2, SS4D, World Tracing, Wan2.2, VACE, and 4DGS) have publicly released code and weights. We will release our code upon acceptance.

AI Use Statement

In this work, we used large language models to edit the text for clarity and to help verify bibliographic entries. We did not use generative AI tools to develop research ideas, design or run experiments, produce results, or create figures. The authors reviewed all AI-assisted text, and every bibliographic entry was checked against arXiv and the official proceedings. We take responsibility for the final content of this work.

References

Alibaba PAI. VideoX-Fun: A more flexible framework that can generate videos at any resolution and creates videos from images. https://github.com/aigc-apps/VideoX-Fun, 2025.

Ronald T. Azuma. A survey of augmented reality. *Presence: Teleoperators & Virtual Environments*, 6(4):355–385, 1997.

Sherwin Bahmani, Xian Liu, Wang Yifan, Ivan Skorokhodov, Victor Rong, Ziwei Liu, Xihui Liu, Jeong Joon Park, Sergey Tulyakov, Gordon Wetzstein, Andrea Tagliasacchi, and David B. Lindell. TC4D: Trajectory-conditioned text-to-4D generation. In *European Conference on Computer Vision (ECCV)*, 2024a.

Sherwin Bahmani, Ivan Skorokhodov, Victor Rong, Gordon Wetzstein, Leonidas Guibas, Peter Wonka, Sergey Tulyakov, Jeong Joon Park, Andrea Tagliasacchi, and David B. Lindell. 4D-fy: Text-to-4D generation using hybrid score distillation sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024b.

Sherwin Bahmani, Ivan Skorokhodov, Aliaksandr Siarohin, Willi Menapace, Guocheng Qian, Michael Vasilkovsky, Hsin-Ying Lee, Chaoyang Wang, Jiaxu Zou, Andrea Tagliasacchi, et al. VD3D: Taming large video diffusion transformers for 3D camera control. In *International Conference on Learning Representations (ICLR)*, 2025.

Jianhong Bai, Menghan Xia, Xiao Fu, Xintao Wang, Lianrui Mu, Jinwen Cao, Zuozhu Liu, Haoji Hu, Xiang Bai, Pengfei Wan, et al. ReCamMaster: Camera-controlled generative rendering from a single video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

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

Wei Cao, Chang Luo, Biao Zhang, Matthias Nießner, and Jiapeng Tang. Motion2VecSets: 4D latent vector set diffusion for non-rigid shape reconstruction and tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Wei Cao, Marcel Hallgarten, Tianyu Li, Daniel Dauner, Xunjiang Gu, Caojun Wang, Yakov Miron, Marco Aiello, Hongyang Li, Igor Gilitschenski, Boris Ivanovic, Marco Pavone, Andreas Geiger, and Kashyap Chitta. Pseudo-simulation for autonomous driving. In *Conference on Robot Learning (CoRL)*, 2025.

Xiang Fan, Sharath Girish, Vivek Ramanujan, Chaoyang Wang, Ashkan Mirzaei, Petr Sushko, Aliaksandr Siarohin, Sergey Tulyakov, and Ranjay Krishna. OmniView: An all-seeing diffusion model for 3D and 4D view synthesis. *arXiv preprint arXiv:2512.10940*, 2025.

Chen Gao, Ayush Saraf, Johannes Kopf, and Jia-Bin Huang. Dynamic view synthesis from dynamic monocular video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

Chenjian Gao, Lihe Ding, Xin Cai, Zhanpeng Huang, Zibin Wang, and Tianfan Xue. LoRA-Edit: Controllable first-frame-guided video editing via mask-aware LoRA fine-tuning. In *International Conference on Learning Representations (ICLR)*, 2026.

Google DeepMind. Veo. https://deepmind.google/models/veo/, 2024.

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 (ICLR)*, 2025.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Wenyi Hong, Weihan Wang, Ming Ding, Wenmeng Yu, Qingsong Lv, Yan Wang, Yean Cheng, Shiyu Huang, Junhui Ji, Zhao Xue, et al. CogVLM2: Visual language models for image and video understanding. *arXiv preprint arXiv:2408.16500*, 2024.

Chen Hou and Zhibo Chen. Training-free camera control for video generation. In *International Conference on Learning Representations (ICLR)*, 2025.

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

Tao Hu, Haoyang Peng, Xiao Liu, and Yuewen Ma. EX-4D: EXtreme viewpoint 4D video synthesis via depth watertight mesh. *arXiv preprint arXiv:2506.05554*, 2025.

Hyeonho Jeong, Suhyeon Lee, and Jong Chul Ye. Reangle-A-Video: 4D video generation as video-to-video translation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11164–11175, 2025.

Yanqin Jiang, Li Zhang, Jin Gao, Weimin Hu, and Yao Yao. Consistent4D: Consistent 360° dynamic object generation from monocular video. In *International Conference on Learning Representations (ICLR)*, 2024.

Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. VACE: All-in-one video creation and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3D Gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), 2023.

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

Zhengfei Kuang, Shengqu Cai, Hao He, Yinghao Xu, Hongsheng Li, Leonidas J. Guibas, and Gordon Wetzstein. Collaborative video diffusion: Consistent multi-video generation with camera control. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

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

Teng Li, Guangcong Zheng, Rui Jiang, Shuigen Zhan, Tao Wu, Yehao Lu, Yining Lin, Chuanyun Deng, Yepan Xiong, Min Chen, Lin Cheng, and Xi Li. RealCam-I2V: Real-world image-to-video generation with interactive complex camera control. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 28785–28796, 2025b.

Zhibing Li, Mengchen Zhang, Tong Wu, Jing Tan, Jiaqi Wang, and Dahua Lin. SS4D: Native 4D generative model via structured spacetime latents. *ACM Transactions on Graphics*, 44(6):244:1–244:12, 2025c. doi: 10.1145/3763302.

Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3D: High-resolution text-to-3D content creation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

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

Yehonathan Litman, Xiaoxuan Ma, Manan Shah, Nicolás Ugrinovic, Kris Kitani, Fernando De la Torre, and Shubham Tulsiani. Lift4D: Harmonizing single-view 3D estimation for 4D reconstruction in-the-wild. *arXiv preprint arXiv:2606.23688*, 2026.

Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3D object. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Tianqi Liu, Zhaoxi Chen, Zihao Huang, Shaocong Xu, Saining Zhang, Chongjie Ye, Bohan Li, Zhiguo Cao, Wei Li, Hao Zhao, et al. Light-X: Generative 4D video rendering with camera and illumination control. In *International Conference on Learning Representations (ICLR)*, 2026.

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 *ACM SIGGRAPH Asia Conference Papers*, 2025.

Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. In *European Conference on Computer Vision (ECCV)*, 2020.

NVIDIA, Arslan Ali, Junjie Bai, Maciej Bala, Yogesh Balaji, Aaron Blakeman, Tiffany Cai, Jiaxin Cao, Tianshi Cao, Elizabeth Cha, et al. World simulation with video foundation models for physical AI. *arXiv preprint arXiv:2511.00062*, 2025.

OpenAI. Sora: Creating video from text. https://openai.com/index/sora/, 2024.

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

Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 DAVIS challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017.

Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. DreamFusion: Text-to-3D using 2D diffusion. In *International Conference on Learning Representations (ICLR)*, 2023.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. SAM 2: Segment anything in images and videos. In *International Conference on Learning Representations (ICLR)*, 2025.

Jiawei Ren, Liang Pan, Jiaxiang Tang, Chi Zhang, Ang Cao, Gang Zeng, and Ziwei Liu. DreamGaussian4D: Generative 4D Gaussian splatting. *arXiv preprint arXiv:2312.17142*, 2023.

Jiawei Ren, Kevin Xie, Ashkan Mirzaei, Hanxue Liang, Xiaohui Zeng, Karsten Kreis, Ziwei Liu, Antonio Torralba, Sanja Fidler, Seung Wook Kim, and Huan Ling. L4GM: Large 4D Gaussian reconstruction model. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

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 (CVPR)*, 2025.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. DreamBooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Ruoxi Shi, Hansheng Chen, Zhuoyang Zhang, Minghua Liu, Chao Xu, Xinyue Wei, Linghao Chen, Chong Zeng, and Hao Su. Zero123++: a single image to consistent multi-view diffusion base model. *arXiv preprint arXiv:2310.15110*, 2023.

Uriel Singer, Shelly Sheynin, Adam Polyak, Oron Ashual, Iurii Makarov, Filippos Kokkinos, Naman Goyal, Andrea Vedaldi, Devi Parikh, Justin Johnson, and Yaniv Taigman. Text-to-4D dynamic scene generation. In *International Conference on Machine Learning (ICML)*, 2023.

Chenxi Song, Yanming Yang, Tong Zhao, Ruibo Li, and Chi Zhang. Taming video models for 3D and 4D generation via zero-shot camera control. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *International Conference on Machine Learning (ICML)*, pp. 9229–9248, 2020.

Jiaxiang Tang, Zhaoxi Chen, Xiaokang Chen, Tengfei Wang, Gang Zeng, and Ziwei Liu. LGM: Large multi-view Gaussian model for high-resolution 3D content creation. In *European Conference on Computer Vision (ECCV)*, 2024.

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

Fengrui Tian, Shaoyi Du, and Yueqi Duan. MonoNeRF: Learning a generalizable dynamic radiance field from monocular videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Fengrui Tian, Yueqi Duan, Angtian Wang, Jianfei Guo, and Shaoyi Du. Semantic Flow: Learning semantic fields of dynamic scenes from monocular videos. In *International Conference on Learning Representations (ICLR)*, 2024.

Fengrui Tian, Tianjiao Ding, Jinqi Luo, Hancheng Min, and René Vidal. Voyaging into perpetual dynamic scenes from a single view. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

Basile Van Hoorick, Rundi Wu, Ege Ozguroglu, Kyle Sargent, Ruoshi Liu, Pavel Tokmakov, Achal Dave, Changxi Zheng, and Carl Vondrick. Generative camera dolly: Extreme monocular dynamic novel view synthesis. In *European Conference on Computer Vision (ECCV)*, 2024.

Vikram Voleti, Chun-Han Yao, Mark Boss, Adam Letts, David Pankratz, Dmitry Tochilkin, Christian Laforte, Robin Rombach, and Varun Jampani. SV3D: Novel multi-view synthesis and 3D generation from a single image using latent video diffusion. In *European Conference on Computer Vision (ECCV)*, 2024.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations (ICLR)*, 2021.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual geometry grounded transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

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 Conference Papers*, 2024.

Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, and Xinggang Wang. 4D Gaussian splatting for real-time dynamic scene rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20310–20320, 2024.

Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-A-Video: One-shot tuning of image diffusion models for text-to-video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7623–7633, 2023.

Rundi Wu, Ruiqi Gao, Ben Poole, Alex Trevithick, Changxi Zheng, Jonathan T. Barron, and Aleksander Holynski. CAT4D: Create anything in 4D with multi-view video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26057–26068, 2025.

Jianfeng Xiang, Xiaoxue Chen, Sicheng Xu, Ruicheng Wang, Zelong Lv, Yu Deng, Hongyuan Zhu, Yue Dong, Hao Zhao, Nicholas Jing Yuan, and Jiaolong Yang. Native and compact structured latents for 3D generation. *arXiv preprint arXiv:2512.14692*, 2025a.

Jianfeng Xiang, Zelong Lv, Sicheng Xu, Yu Deng, Ruicheng Wang, Bowen Zhang, Dong Chen, Xin Tong, and Jiaolong Yang. Structured 3D latents for scalable and versatile 3D generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025b.

Yiming Xie, Chun-Han Yao, Vikram Voleti, Huaizu Jiang, and Varun Jampani. SV4D: Dynamic 3D content generation with multi-frame and multi-view consistency. In *International Conference on Learning Representations (ICLR)*, 2025.

Jiale Xu, Weihao Cheng, Yiming Gao, Xintao Wang, Shenghua Gao, and Ying Shan. InstantMesh: Efficient 3D mesh generation from a single image with sparse-view large reconstruction models. *arXiv preprint arXiv:2404.07191*, 2024.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. CogVideoX: Text-to-video diffusion models with an expert transformer. In *International Conference on Learning Representations (ICLR)*, 2025.

Chun-Han Yao, Yiming Xie, Vikram Voleti, Huaizu Jiang, and Varun Jampani. SV4D 2.0: Enhancing spatio-temporal consistency in multi-view video diffusion for high-quality 4D generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

Hong-Xing Yu, Haoyi Duan, Junhwa Hur, Kyle Sargent, Michael Rubinstein, William T. Freeman, Forrester Cole, Deqing Sun, Noah Snavely, Jiajun Wu, et al. WonderJourney: Going from anywhere to everywhere. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Hong-Xing Yu, Haoyi Duan, Charles Herrmann, William T. Freeman, and Jiajun Wu. WonderWorld: Interactive 3D scene generation from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025a.

Mark Yu, Wenbo Hu, Jinbo Xing, and Ying Shan. TrajectoryCrafter: Redirecting camera trajectory for monocular videos via diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025b.

Wangbo Yu, Jinbo Xing, Li Yuan, Wenbo Hu, Xiaoyu Li, Zhipeng Huang, Xiangjun Gao, Tien-Tsin Wong, Ying Shan, and Yonghong Tian. ViewCrafter: Taming video diffusion models for high-fidelity novel view synthesis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025c. doi: 10.1109/TPAMI.2025.3613256.

Bowen Zhang, Sicheng Xu, Chuxin Wang, Jiaolong Yang, Feng Zhao, Dong Chen, and Baining Guo. Gaussian variation field diffusion for high-fidelity video-to-4D synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025a.

David Junhao Zhang, Roni Paiss, Shiran Zada, Nikhil Karnad, David E. Jacobs, Yael Pritch, Inbar Mosseri, Mike Zheng Shou, Neal Wadhwa, and Nataniel Ruiz. ReCapture: Generative video camera controls for user-provided videos using masked video fine-tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2050–2062, 2025b.

Hao Zhang, Chun-Han Yao, Simon Donné, Narendra Ahuja, and Varun Jampani. Stable part diffusion 4D: Multi-view RGB and kinematic parts video generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025c.

Hao Zhang, Mohamed El Banani, Jen-Hao Cheng, Paul Zhang, Yi Hua, Ben Mildenhall, Christoph Lassner, Narendra Ahuja, and Gengshan Yang. World tracing: Generative pixel-aligned geometry beyond the visible. arXiv preprint arXiv:2606.13652, 2026a.

Songchun Zhang, Huiyao Xu, Sitong Guo, Zhongwei Xie, Hujun Bao, Weiwei Xu, and Changqing Zou. SpatialCrafter: Unleashing the imagination of video diffusion models for scene reconstruction from limited observations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025d.

Yanran Zhang, Ziyi Wang, Wenzhao Zheng, Zheng Zhu, Jie Zhou, and Jiwen Lu. Geometry-aware single-image 4D synthesis via dense trajectory generation. In *European Conference on Computer Vision (ECCV)*, 2026b.

Yuyang Zhao, Zhiwen Yan, Enze Xie, Lanqing Hong, Zhenguo Li, and Gim Hee Lee. Animate124: Animating one image to 4D dynamic scene. *arXiv preprint arXiv:2311.14603*, 2023.

Peng Zheng, Dehong Gao, Deng-Ping Fan, Li Liu, Jorma Laaksonen, Wanli Ouyang, and Nicu Sebe. Bilateral reference for high-resolution dichotomous image segmentation. *CAAI Artificial Intelligence Research*, 3:9150038, 2024.

Kaichen Zhou, Yuhan Wang, Grace Chen, Xinhai Chang, Gaspard Beaudouin, Fangneng Zhan, Paul Pu Liang, and Mengyu Wang. PAGE-4D: Disentangled pose and geometry estimation for VGGT-4D perception. In *International Conference on Learning Representations (ICLR)*, 2026.

A Implementation Details

Latent video diffusion. TTA4D builds on latent video diffusion models (Rombach et al., 2022; Team Wan et al., 2025). Given a target video $\mathbf{V}_{\text{target}}$, a pre-trained encoder $\mathcal{E}$ produces a latent $\mathbf{z}_0 = \mathcal{E}(\mathbf{V}_{\text{target}})$, and the forward process at noise level s is

$$\mathbf{z}_s = \alpha_s \mathbf{z}_0 + \sigma_s \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \tag{7}$$

We keep each backbone's native training target (the flow-matching velocity for Wan-based models) and its native noise schedule.

LoRA configuration. LoRA updates (Eq. (5)) with rank r=32 and scaling factor α=32 (*i.e.*, α/r=1) are attached to the self- and cross-attention projections of the backbone; all other weights remain frozen. We optimize the LoRA parameters with AdamW (β_1=0.9, β_2=0.999) at learning rate 10^{-4} and batch size 1, at a training resolution of 256. Adaptation uses 400 steps in total (100 per training sequence) and takes approximately 10–15 minutes per object.

Inference. We sample at a resolution of 512 with 20 denoising steps and a classifier-free guidance scale of 5.0. Sampling a 49-frame video takes approximately 4.7 minutes; all reported results use 24-frame clips.

Backbone compatibility. The conditioning extraction (Section 3.1) and dual-signal construction (Section 3.2) are independent of the backbone, and adaptation only updates LoRA parameters. We apply the same procedure to two backbones with different conditioning interfaces: a plain image-to-video model (Wan2.2-I2V) and a depth-controllable model (VACE). Because a plain image-to-video model suffices and its built-in image conditioning is reused without architectural changes, a newly released video foundation model requires only re-running the lightweight adaptation, with no full-model retraining and no new paired control-video dataset.

B Extended Method Description and Related Work

Pipeline summary. TTA4D has three stages: (1) geometric conditioning extraction (Section 3.1), which produces depth maps and object masks for training and inference; (2) dual-signal construction (Section 3.2), which assembles DOFC, covering object motion under the input camera, and SODC, covering viewpoint change under camera motion; and (3) test-time adaptation via LoRA (Section 3.3), which specializes a pre-trained video diffusion model to the target object in approximately 10–15 minutes. At inference, the adapted model takes a reference image and a depth-and-mask sequence rendered along the desired trajectory, and generates a video that follows this geometry while preserving the input's appearance and motion. The conditioning constrains the broad video diffusion prior toward the desired object shape and camera trajectory, while the adapted model supplies high-frequency appearance.

Figure 2 in detail. In panel (a), the stacked orange boxes denote the $K=3$ SODC trajectories (`spin360`, `arc left`, and `arc right`). The reference image is the first input frame for DOFC and at inference, and the first rendered frame of each trajectory for SODC. Because the 3D model is generated from a single image, it synthesizes plausible appearance for parts never observed in the input, such as the back of a forward-facing object; this property allows SODC to supervise full 360° orbits. In panel (b), the stacked boxes denote multiple user-specified trajectories served by the same adapted model.

Relation to two-stage 4D pipelines. Two-stage pipelines such as SS4D (Li et al., 2025c) first generate a coarse volumetric sequence and then convert it to 3D Gaussians (Kerbl et al., 2023; Xiang et al., 2025b). The second stage requires 4D supervision that is difficult to collect, and it generalizes poorly to real-world objects. TTA4D can reuse the first stage of such a pipeline as its inference geometry while replacing the second stage with test-time adaptation, which is the TTA4D (SS4D + I2V) configuration in Table 3 and the configuration used on DAVIS.

Relation to per-video camera redirection. ReCapture (Zhang et al., 2025b) and Reangle-A-Video (Jeong et al., 2025) supervise with anchor or warped videos in which only the pixels visible after reprojection are valid, and restrict their losses to those pixels. TTA4D instead conditions on complete depth-and-mask sequences rather than partially valid warped pixels, and its SODC signal provides explicit supervision for surfaces never seen in the input.

C Extended Analysis: VACE under Identical Geometric Conditioning

VACE (Jiang et al., 2025) is an informative reference because, unlike the other baselines, it can consume exactly the same depth-and-mask sequence and reference image as TTA4D. VACE is trained for zero-shot control on a large paired corpus of depth-, mask-, and reference-conditioned videos, whereas TTA4D (WT + VACE) starts from the same weights and additionally performs per-object adaptation. Holding the conditioning fixed therefore isolates the contribution of adaptation.

Geometry following versus appearance fidelity. Table 3 separates the two. Adaptation improves IoU only modestly (0.723 to 0.744), so zero-shot VACE already follows the conditioned silhouette.

The larger gains are in appearance: LPIPS drops by 29% (0.129 to 0.091) and PSNR rises by 1.5 dB (19.168 to 20.662). DOFC-only adaptation recovers part of this gain (LPIPS 0.111), indicating that anchoring the model to the input's appearance is a major source of improvement, and SODC adds the rest by teaching how that appearance should change with viewpoint. Qualitatively, zero-shot VACE renders textures in a style that reflects its training distribution rather than the style of the input video. This is expected: VACE's only information about the target is the reference image, so its prior over appearance under novel viewpoints comes from its training data, whereas the adapted model has been trained on the target object itself.

D Per-Group Results of TTA4D Configurations

Table 4 reports seen- and unseen-trajectory results on Consistent4D for the configurations averaged in Table 3, together with the full SS4D pipeline, whose first-stage voxels TTA4D (SS4D + I2V) uses as inference geometry. Per-group values for zero-shot and DOFC-only adaptation are not reported separately.

Table 4: **Per-group results of TTA4D configurations on Consistent4D.** Same metrics and trajectory groups as Table 1.

Method	Seen Trajectories				Unseen Trajectories			
	PSNR↑	SSIM↑	LPIPS↓	IoU↑	PSNR↑	SSIM↑	LPIPS↓	IoU↑
SS4D	18.351	0.904	0.105	0.641	17.589	0.896	0.113	0.626
TTA4D (SS4D + I2V)	19.152	0.914	0.110	0.712	18.884	0.910	0.119	0.706
TTA4D (WT + I2V)	19.876	0.914	0.106	0.744	19.657	0.907	0.112	0.734
TTA4D (WT + VACE)	20.842	0.926	0.084	0.747	20.482	0.916	0.098	0.740

E User Study Protocol

We conducted a blinded user study on the 22 DAVIS sequences with 20 participants, each of whom rated every video in the study. For each comparison, a participant sees the input video alongside all methods' outputs for the same camera trajectory; methods are shown under anonymous labels in an order randomized independently per comparison. Each result is rated on a 1–5 Likert scale along four axes: *overall quality*; *smoothness*, the temporal stability and motion fluency; *style consistency*, the faithfulness to the color, texture, and style of the input; and *object consistency*, the stability of object identity and structure across frames and viewpoints. All TTA4D results use the weakest configuration (Wan2.2-I2V with SS4D's coarse voxels). Table 2 reports mean ratings: TTA4D receives 4.56, 4.48, 4.61, and 4.67 for overall quality, smoothness, style consistency, and object consistency, against best-baseline scores of 3.26, 3.31, 3.24, and 3.22.

F Additional Visual Results on DAVIS

Figure 3 in the main paper shows four DAVIS sequences, each paired with a single baseline. Here we provide qualitative comparisons for 16 of the 22 DAVIS sequences used in the user study, including those four, across TTA4D and all six baselines: GVFDiffusion, L4GM, SS4D, Lift4D, TrajectoryCrafter, and GEN3C. The eight figures below each show two sequences side by side, with three samples per sequence (zero-based frame indices 0, 8, and 23). The top row contains input-video frames; the remaining rows show the methods along approximate `spin360` camera paths, with time increasing from left to right within each sequence. As in the main paper, TTA4D uses its weakest configuration (Wan2.2-I2V with SS4D's coarse voxels). For display, each method receives a fixed scale and translation estimated from its first frame, followed by a shared crop for the sequence and a common white background. Camera poses and timestamps are not exactly matched across all sources.



Figure 5: **Qualitative comparisons on DAVIS (part 1/8).** `bear` (left) and `horsejump-high` (right). Each sequence shows frames 0, 8, and 23; rows show the input video, six baselines, and TTA4D (ours).



Figure 6: **Qualitative comparisons on DAVIS (part 2/8).** `bike-packing` (left) and `blackswan` (right). Each sequence shows frames 0, 8, and 23; rows show the input video, six baselines, and TTA4D (ours).



Figure 7: **Qualitative comparisons on DAVIS (part 3/8).** `camel` (left) and `car-roundabout` (right). Each sequence shows frames 0, 8, and 23; rows show the input video, six baselines, and TTA4D (ours).



Figure 8: **Qualitative comparisons on DAVIS (part 4/8).** `car-shadow` (left) and `cows` (right). Each sequence shows frames 0, 8, and 23; rows show the input video, six baselines, and TTA4D (ours).



Figure 9: **Qualitative comparisons on DAVIS (part 5/8).** `dance-jump` (left) and `dance-twirl` (right). Each sequence shows frames 0, 8, and 23; rows show the input video, six baselines, and TTA4D (ours).



Figure 10: **Qualitative comparisons on DAVIS (part 6/8).** `gold-fish` (left) and `hike` (right). Each sequence shows frames 0, 8, and 23; rows show the input video, six baselines, and TTA4D (ours).



Figure 11: **Qualitative comparisons on DAVIS (part 7/8).** `hockey` (left) and `horsejump-low` (right). Each sequence shows frames 0, 8, and 23; rows show the input video, six baselines, and TTA4D (ours).



Figure 12: **Qualitative comparisons on DAVIS (part 8/8).** `breakdance` (left) and `dog` (right). Each sequence shows frames 0, 8, and 23; rows show the input video, six baselines, and TTA4D (ours).

F.1 Recurring Failure Modes of Existing Methods

Across Fig. 3 and Figs. 5 to 12 three failure modes recur. **(i) Geometric distortion.** Explicit 4D methods (SS4D, L4GM, GVFDiffusion, Lift4D) often fail to produce coherent shapes when their intermediate 3D estimate is inaccurate, resulting in broken parts, missing limbs, and deformed silhouettes that are most apparent under large viewpoint changes. **(ii) Temporal flicker.** Baseline renderings often fluctuate between adjacent frames, with unstable textures and boundaries, most noticeably on articulated subjects such as horses, dancers, and hockey players. **(iii) Appearance drift and reprojection artifacts.** Camera-controlled video models (TrajectoryCrafter, GEN3C) generate from reprojected geometry; when the target view departs far from the input view, large unobserved regions must be synthesized without knowledge of the object, and the outputs drift in color and texture or fragment (Figs. 5 to 12). TTA4D mitigates these failure modes by anchoring generation to the target object: DOFC ties the object's appearance to the input video while teaching the model to follow object motion in the conditioning, and SODC teaches it to follow viewpoint changes, including views of unobserved sides. TTA4D also has failure modes: when the first-frame 3D reconstruction is severely wrong or the object undergoes large topology changes, it can inherit distorted geometry or appearance from the SODC signal (Section 5).

G Broader Impacts and Responsible Use

TTA4D can support creative production, simulation, and accessibility by converting casual monocular videos into controllable object-centric renderings. The same capability could be misused to create misleading videos or unauthorized re-renderings of recognizable subjects. Responsible deployment should respect consent and licensing of input videos, disclose generated content, and use provenance or watermarking mechanisms when outputs are shared publicly.