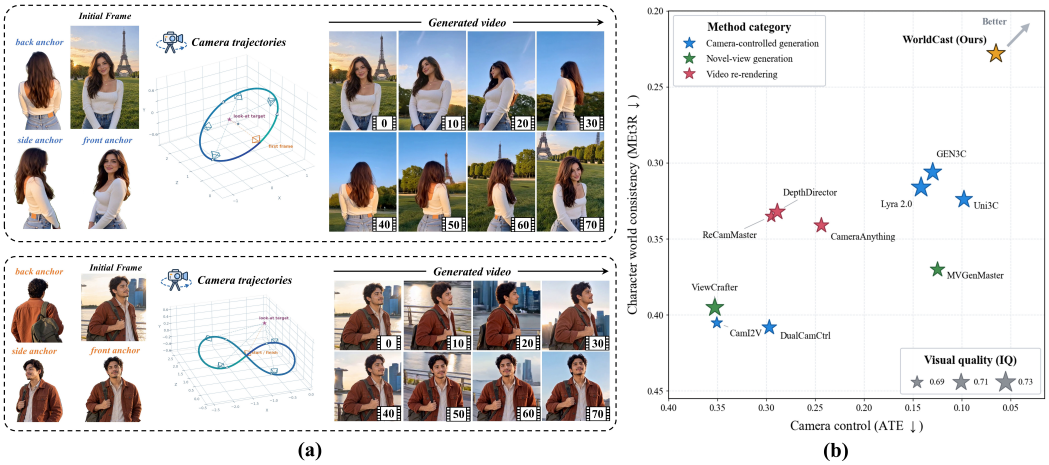
WORLDCAST: WORLD-CONSISTENT CHARACTER VIDEO GENERATION WITH LARGE-ANGLE CAMERA CONTROL

Anonymous authors
Paper under double-blind review

ABSTRACT

Precise camera control is essential for cinematic character video generation, yet existing methods struggle to follow large-angle trajectories that combine rotation, translation, and radius variation, with arbitrary initial azimuths. Such viewpoint changes also reveal character regions unseen in the initial frame, often leading to geometric distortions, appearance degradation, and identity drift. To systematically study these coupled challenges, we introduce ***CLACT***, *a large-scale **C**haracter-centric dataset and benchmark covering diverse **L**arge-**A**ngle **C**amera **T**rajectories*, comprising 80,000 synthetic training samples and 42 synthetic and real-world evaluation scenes, with accurate camera geometry and multi-view character observations. We further propose ***WorldCast***, *a framework that couples scale-aligned geometric guidance with geometry-aligned multi-view character memory*. Specifically, we back-project the initial view into a 3D point cloud and render it along the target trajectory to provide trajectory-aligned geometric guidance and coverage masks. To complement character information missing from the initial view, we introduce Geometry-Aligned Memory Attention (GAMA), which aligns multi-view anchor geometry with target-view geometry in a shared coordinate frame, enabling geometry-aware retrieval and injection of view-relevant character features. Experiments across diverse large-angle trajectories demonstrate improved camera-control accuracy and character consistency over the compared methods while maintaining competitive visual quality. *https://worldcast-video.github.io/worldcast/*



(a) (b)

Figure 1: ***WorldCast* enables accurate large-angle camera control while preserving character consistency. (a)** Generated sequences along tilted circular and lemniscate camera trajectories. **(b)** Quantitative comparison of camera-control accuracy (ATE ↓), cross-view geometric consistency (MEt3R ↓), and image quality (IQ ↑). Each star denotes a method; colors distinguish method categories, while star size indicates image quality. Both axes are reversed, such that methods with lower ATE and MEt3R appear toward the upper right.

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

Recent video diffusion models have achieved remarkable visual fidelity in text- and image-conditioned video generation (Wan et al., 2025; Yang et al., 2025). For cinematic content creation, however, visual quality alone is insufficient: creators also require precise control over camera motion (Courant et al., 2024). In character-centric videos, camera motion can reveal a subject from different viewpoints, establish spatial relationships, and direct visual attention. Realizing such shots therefore requires not only accurate trajectory following, but also consistent preservation of character geometry, appearance, and identity across substantial viewpoint changes (Yang et al., 2026b).

Camera-controlled video generation has progressed from camera-pose and ray-based conditioning to explicit geometric guidance derived from 3D renderings (Wang et al., 2024; Zheng et al., 2024; He et al., 2025; Ren et al., 2025; Cao et al., 2025b). Recent methods further explore joint appearance-depth modeling and more expressive camera representations to support complex camera motions (Zhang et al., 2026a; 2025). Meanwhile, multi-view generation and video re-rendering provide complementary means of improving cross-view consistency and synthesizing novel viewpoints (Zhou et al., 2025; Cao et al., 2025a; Bai et al., 2025; Li et al., 2026). Despite this progress, accurate trajectory following and consistent character generation remain difficult under large viewpoint changes.

Large-angle camera motion introduces two closely coupled challenges. First, trajectories involving substantial rotation together with translation, radius variation, and arbitrary initial azimuths place significantly stronger demands on camera control. Limited trajectory coverage or inaccurate geometric conditioning can result in under-rotation, incorrect translation, and drift from the requested path (Yang et al., 2026a). Second, the same viewpoint changes progressively expose character regions that are unobserved in the initial frame, leaving their geometry, appearance, and identity weakly constrained (Yu et al., 2024; Wu et al., 2026). These challenges are inherently linked: the target camera trajectory determines both the geometry that the generated video must follow and the previously unseen character content that must be synthesized at each viewpoint. Explicit geometric guidance can constrain the former, while multi-view character observations provide complementary evidence for the latter. Effectively combining them therefore requires relating both sources of information to the target viewpoints within a shared geometric reference frame.

Existing data remain insufficient for systematically studying and evaluating large-angle character-centric camera control. Current camera-motion datasets provide limited coverage of complex trajectories involving substantial rotation, translation, radius variation, and diverse initial azimuths (Li et al., 2025; Bai et al., 2025). Moreover, camera-motion sequences are not consistently paired with complementary views of the same character, limiting the joint study of trajectory accuracy and cross-view character consistency (Wu et al., 2026; Wang et al., 2026a). To address these limitations, we construct ***CLACT***, a character-centric dataset and benchmark designed for diverse large-angle camera motion. Its training corpus contains $50{,}000$ camera-motion sequences rendered in Unreal Engine 5 and $30{,}000$ Blender-rendered samples with multi-view character anchors, while the evaluation set covers 42 synthetic and real-world character scenes under multiple trajectory families. Together, these data support both training and systematic evaluation of trajectory accuracy, character consistency, and visual quality across diverse camera motions and substantial viewpoint changes.

Building on this benchmark, we propose ***WorldCast***, a framework that couples scale-aligned geometric guidance with geometry-aligned multi-view character memory. We align the relative scale of initial-view depth and camera translation, then back-project the initial frame into a 3D point cloud and render it along the target trajectory to obtain geometric guidance and coverage masks. To complement character information missing from the initial view, we introduce Geometry-Aligned Memory Attention (GAMA). Using camera poses jointly estimated from the initial frame and character anchors, GAMA aligns anchor and target-view geometry in a shared coordinate frame, enabling geometry-aware retrieval of view-relevant character features and their injection into the video tokens. As shown in Fig. 1, the framework improves camera-control accuracy and cross-view character consistency while maintaining competitive visual quality.

Our contributions are threefold:

- We introduce ***CLACT***, a character-centric dataset and benchmark combining diverse large-angle camera trajectories and multi-view character observations to support training and evaluation of camera control and character consistency.

- We propose ***WorldCast***, a video generation framework that couples scale-aligned geometric guidance with multi-view character memory for accurate large-angle camera control and consistent character generation.
- We introduce Geometry-Aligned Memory Attention (GAMA), which aligns multi-view anchor geometry with the target camera trajectory in a shared coordinate frame, enabling geometry-aware retrieval and injection of view-relevant character features.

2 Related Work

Camera-controllable video generation. Camera-controlled video generation incorporates explicit motion conditions into video diffusion models. MotionCtrl disentangles camera and object motion, while CameraCtrl introduces camera conditioning (Wang et al., 2024; He et al., 2025). CamCo and CamI2V incorporate camera geometry and epipolar constraints into cross-frame modeling (Xu et al., 2024; Zheng et al., 2024). VD3D and AC3D extend camera conditioning to video diffusion transformers (Bahmani et al., 2025b;a), while UCPE unifies camera pose, intrinsics, distortion, and orientation in a common representation (Zhang et al., 2026a). DualCamCtrl couples RGB and depth generation to strengthen adherence to camera conditions (Zhang et al., 2025). For video re-rendering, ReCamMaster, DepthDirector, and CameraAnything generate new camera trajectories while preserving source content and dynamics (Bai et al., 2025; Chen et al., 2026; Li et al., 2026).

Geometry-aware and multi-view generation. Geometric representations and complementary observations provide spatial constraints for viewpoint synthesis. 3DiM, GeNVS, and CAT3D study novel-view synthesis from one or a few observations (Watson et al., 2023; Chan et al., 2023; Gao et al., 2024), while ViewCrafter, MVGenMaster, and SEVA leverage geometry or multiple input views to synthesize target viewpoints (Yu et al., 2024; Cao et al., 2025a; Zhou et al., 2025). For camera-controlled video generation, RealCam-I2V, Uni3C, and GEN3C introduce explicit scene geometry or trajectory-aligned 3D guidance (Li et al., 2025; Cao et al., 2025b; Ren et al., 2025). ConfCtrl improves robustness to unreliable projections (Yang et al., 2026a), while Lyra 2.0 and WorldStereo further explore persistent geometry and geometry-constrained memory retrieval (Shen et al., 2026; Zhang et al., 2026b). Our work instead relates multi-view character-anchor geometry to the target trajectory, enabling view-relevant character evidence to guide generation.

Identity-consistent video generation. Subject-conditioned video generation seeks to preserve appearance and identity across generated frames. VideoBooth and DreamVideo introduce subject-specific conditioning (Jiang et al., 2024; Wei et al., 2024), while ConsisID and Magic-Me develop mechanisms for identity preservation (Yuan et al., 2025; Ma et al., 2024). ConsID-Gen uses complementary views to improve cross-view consistency, and Gloria represents character appearance through content anchors (Wu et al., 2026; Yang et al., 2026b). Virtually Being combines identity preservation across viewpoints with camera control through customization based on performance captures (Xu et al., 2025). We jointly address character consistency and large-angle camera control through geometric guidance and view-aware retrieval from multi-view character memory.

3 Method

3.1 Preliminaries: Video Diffusion Models

Latent video diffusion models (Blattmann et al., 2023) operate in the compressed space of a pre-trained variational autoencoder (VAE). Given a video $\boldsymbol{x}$, the encoder $\mathcal{E}$ produces a latent representation $\boldsymbol{z}_0 = \mathcal{E}(\boldsymbol{x}) \in \mathbb{R}^d$. Following the flow matching formulation (Lipman et al., 2022), we independently sample $\boldsymbol{z}_1 \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_d)$ and define the linear interpolation:

$$\boldsymbol{z}_\tau = (1-\tau)\boldsymbol{z}_0 + \tau \boldsymbol{z}_1, \qquad \tau \in [0,1]. \tag{1}$$

Conditioned on text and initial-frame features $c = (c_{\text{txt}}, c_{\text{img}})$, a diffusion transformer v_θ learns the conditional velocity field with target velocity $\boldsymbol{u}_\tau = \frac{\mathrm{d}\boldsymbol{z}_\tau}{\mathrm{d}\tau} = \boldsymbol{z}_1 - \boldsymbol{z}_0$ by minimizing:

$$\mathcal{L}_{\text{FM}}(\theta) = \mathbb{E}_{(\boldsymbol{z}_0, c)\sim p_{\text{data}},\, \tau\sim p_\tau,\, \boldsymbol{z}_1\sim\mathcal{N}(\mathbf{0},\boldsymbol{I}_d)} \left[\| v_\theta(\boldsymbol{z}_\tau, \tau, c) - \boldsymbol{u}_\tau \|_2^2 \right]. \tag{2}$$

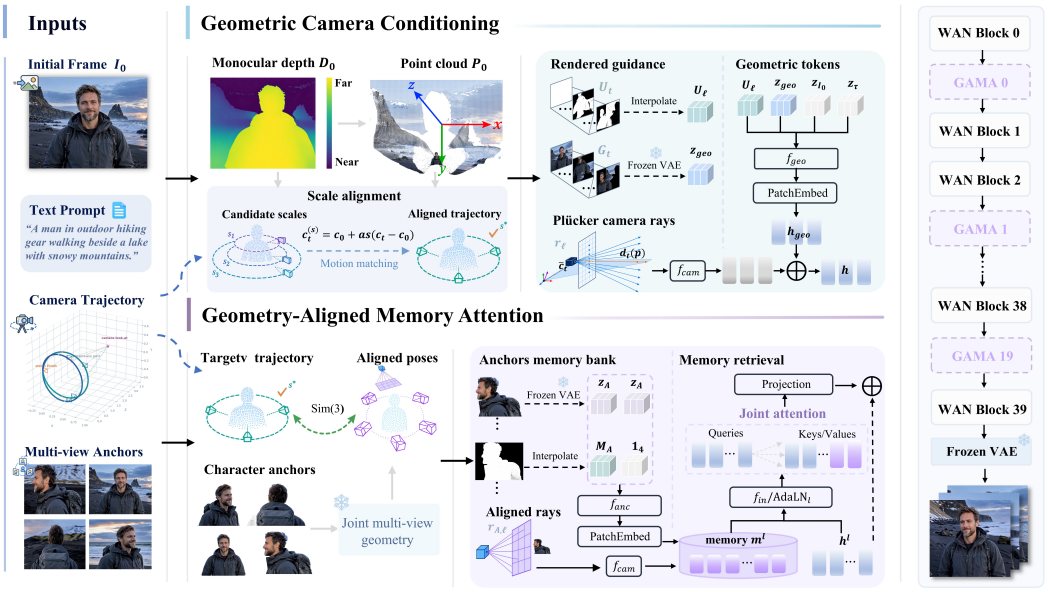


Figure 2: Overview of ***WorldCast***. The framework combines trajectory-aligned geometric guidance with geometry-aligned multi-view character memory.

The population-optimal velocity field is $v^*(\boldsymbol{z}_\tau, \tau, c) = \mathbb{E}[\boldsymbol{u}_\tau \mid \boldsymbol{z}_\tau, \tau, c]$. At inference, the ODE $\mathrm{d}\boldsymbol{z}_\tau/\mathrm{d}\tau = v_\theta(\boldsymbol{z}_\tau, \tau, c)$ is integrated from $\tau = 1$ to $\tau = 0$, and the resulting latent is decoded by the VAE. We adopt Wan2.1-I2V-14B (Wan et al., 2025), which combines a causal 3D VAE with a conditional diffusion transformer, as our generative backbone.

3.2 Problem Formulation

We formulate character-centric video generation under prescribed camera motion as a conditional generation problem (He et al., 2025). The inputs comprise an initial character image $I_0 \in \mathbb{R}^{H\times W\times 3}$ captured from an arbitrary viewpoint, a text prompt y, and a target camera sequence $\mathcal{C} = \{(\boldsymbol{K}_t, \boldsymbol{T}_t)\}_{t=0}^{N-1}$, where $\boldsymbol{K}_t \in \mathbb{R}^{3\times 3}$ denotes the camera intrinsics and $\boldsymbol{T}_t \in \mathrm{SE}(3)$ the camera-to-world pose. The goal is to synthesize future frames $\mathcal{X} = (I_1, \ldots, I_{N-1})$ that follow the prescribed trajectory while preserving character geometry, appearance, and identity and satisfying the text prompt.

Large-angle camera motion reveals character surfaces unobserved in I_0, coupling trajectory fidelity with the completion of missing character information (Li et al., 2025). To provide complementary evidence, we additionally condition generation on M views of the same character, $\mathcal{A} = \{A_j\}_{j=1}^{M}$, and define the available observations as $\mathcal{O} = \{I_0\} \cup \mathcal{A}$. The conditional generation problem is:

$$\widehat{\mathcal{X}} \sim p_\theta(\mathcal{X} \mid I_0, y, \mathcal{C}, \mathcal{A}), \qquad \widehat{\mathcal{X}} = \{\widehat{I}_t\}_{t=1}^{N-1}. \tag{3}$$

As illustrated in Fig. 2, we combine geometric camera conditioning (Sec. 3.3) with geometry-aligned character memory (Sec. 3.4) to achieve accurate trajectory following and consistent character geometry, appearance, and identity across viewpoints.

3.3 Geometric Camera Conditioning

Depth–trajectory scale alignment. We estimate the initial-view depth D_0 using a pretrained depth estimator (Bochkovskiy et al., 2025) and calibrate the relative scale between the predicted depth and camera translation using the training reference video (Cao et al., 2025a).

Let $\boldsymbol{c}_t \in \mathbb{R}^3$ denote the camera center at frame t. For each candidate scale $s \in \mathcal{S}$, we rescale the camera translation relative to $\boldsymbol{c}_0$ while keeping the rotations and D_0 fixed:

$$\boldsymbol{c}_t^{(s)} = \boldsymbol{c}_0 + as\big(\boldsymbol{c}_t - \boldsymbol{c}_0\big), \qquad s \in \mathcal{S}. \tag{4}$$

Here, a is the fixed scale factor provided by the camera metadata, and $\boldsymbol{c}_t^{(s)}$ denotes the rescaled camera center. For each candidate scale, we back-project the initial view into a colored point cloud and render it along the corresponding trajectory. We select the scale whose induced image motion best matches that of the reference video:

$$s^* = \arg\min_{s\in\mathcal{S}} \operatorname*{median}_{i\in\mathcal{V}_s} \left| \log \frac{m_i^{\mathrm{render}}(s)}{m_i^{\mathrm{ref}}} \right|. \tag{5}$$

Here, $m_i^{\mathrm{render}}(s)$ and m_i^{ref} denote the motion magnitudes of the i-th sampled frame pair in the rendered and reference sequences, respectively, and $\mathcal{V}_s$ is the set of valid frame pairs. Detailed projection derivations and motion-estimation procedures are provided in Appendix A.

Projected geometric guidance. We back-project the initial image into a colored point cloud (Ren et al., 2025). Let $\boldsymbol{R}_0$ and $\boldsymbol{c}_0$ denote the rotation and camera center of $\boldsymbol{T}_0$, respectively. For each pixel $\boldsymbol{p} = [u, v]^\top \in \Omega_0$ with valid depth, let $\widetilde{\boldsymbol{p}} = [u, v, 1]^\top$ denote its homogeneous coordinate. Its world-space position is:

$$\boldsymbol{X}(\boldsymbol{p}) = \boldsymbol{R}_0 \left(D_0(\boldsymbol{p}) \boldsymbol{K}_0^{-1} \widetilde{\boldsymbol{p}} \right) + \boldsymbol{c}_0, \qquad \mathcal{P}_0 = \left\{ \left(\boldsymbol{X}(\boldsymbol{p}), I_0(\boldsymbol{p}) \right) \,\middle|\, \boldsymbol{p} \in \Omega_0 \right\}, \tag{6}$$

where Ω_0 denotes the set of pixels with valid depth. We then render $\mathcal{P}_0$ along the calibrated trajectory using point splatting and depth buffering:

$$\left(G_t,\, U_t\right) = \mathcal{R}\left(\mathcal{P}_0,\, \boldsymbol{K}_t,\, \boldsymbol{T}_t^{(s^*)}\right), \qquad t = 0, \ldots, N-1, \tag{7}$$

where $\boldsymbol{T}_t^{(s^*)}$ retains the rotation of $\boldsymbol{T}_t$ and uses the calibrated camera center from Eq. (4). Here, G_t provides RGB guidance aligned with the target viewpoint, while $U_t \in \{0,1\}^{H\times W}$ indicates pixels not supported by the projected point cloud and therefore left for the generator to complete.

Geometric and camera conditioning. We encode the projected sequence $\mathcal{G} = (G_t)_{t=0}^{N-1}$ with the pretrained VAE to obtain $\boldsymbol{z}_{\mathrm{geo}}$ and resize the coverage masks to the latent grid as $\boldsymbol{U}_\ell$ (Cao et al., 2025b). Before patch embedding, we concatenate these conditions with the noisy video latent $\boldsymbol{z}_\tau$ and the native initial-frame condition $\boldsymbol{z}_{I_0}$:

$$h_{\mathrm{geo}} = \mathrm{PatchEmbed}(f_{\mathrm{geo}}([\boldsymbol{z}_\tau, \boldsymbol{z}_{I_0}, \boldsymbol{z}_{\mathrm{geo}}, \boldsymbol{U}_\ell]_{\mathrm{ch}})), \tag{8}$$

where $\boldsymbol{z}_{I_0}$ contains the native image-conditioning latent and known-frame mask, $[\cdot]_{\mathrm{ch}}$ denotes channel-wise concatenation, and f_{geo} is a $1\times1\times1$ convolution that projects the fused features to the input dimension of the pretrained patch embedding.

We further encode the target camera geometry using Plücker rays (Sitzmann et al., 2021) in a shared normalized coordinate frame:

$$\boldsymbol{r}_t(\boldsymbol{p}) = \left[\boldsymbol{d}_t(\boldsymbol{p}), \bar{\boldsymbol{c}}_t \times \boldsymbol{d}_t(\boldsymbol{p})\right], \qquad h = h_{\mathrm{geo}} + f_{\mathrm{cam}}(\boldsymbol{r}_\ell), \tag{9}$$

where $\bar{\boldsymbol{c}}_t$ and $\boldsymbol{d}_t(\boldsymbol{p})$ denote the camera center and unit ray direction in the normalized frame, respectively, and $\boldsymbol{r}_\ell$ denotes the ray grids resized to the latent resolution. The projection f_{cam} maps the ray features to the same token grid and feature dimension as h_{geo} for element-wise addition.

3.4 Geometry-Aligned Memory Attention

Anchor–trajectory geometric alignment. We extract character foreground masks for the initial frame I_0 and multi-view anchors $\mathcal{A} = \{A_j\}_{j=1}^M$ using a pretrained segmentation model (Carion et al., 2026), and suppress backgrounds before geometric estimation. A pretrained multi-view geometry model (Wang et al., 2026c) then jointly estimates their camera poses, recovering the relative geometry in a shared coordinate frame. To align the anchor geometry with the target trajectory, we establish camera-center correspondences between observations and target views, with I_0 as the shared initial view.

Let $\widehat{\boldsymbol{R}}_j$ and $\widehat{\boldsymbol{c}}_j$ denote the estimated camera-to-world rotation and camera center of observation j, where $j = 0$ corresponds to I_0. For each observation used in alignment, $\kappa(j)$ denotes the corresponding target frame, with $\kappa(0) = 0$. The remaining correspondences are obtained from source-frame metadata when available, or by nearest-azimuth matching. Given the valid correspondence set $\mathcal{J}$, we estimate a similarity transform using Umeyama alignment (Umeyama, 1991):

$$(\rho^*, \boldsymbol{Q}^*, \boldsymbol{b}^*) = \arg\min_{\rho>0,\, \boldsymbol{Q}\in\mathrm{SO}(3),\, \boldsymbol{b}\in\mathbb{R}^3} \sum_{j\in\mathcal{J}} \left\| \rho \boldsymbol{Q} \widehat{\boldsymbol{c}}_j + \boldsymbol{b} - \boldsymbol{c}_{\kappa(j)} \right\|_2^2. \tag{10}$$

After reconciling camera conventions, the estimated transform maps the anchor geometry into the target-trajectory coordinate frame:

$$\boldsymbol{c}_j^{\mathrm{aln}} = \rho^* \boldsymbol{Q}^* \widehat{\boldsymbol{c}}_j + \boldsymbol{b}^*, \qquad \boldsymbol{R}_j^{\mathrm{aln}} = \boldsymbol{Q}^* \widehat{\boldsymbol{R}}_j. \tag{11}$$

We then apply the same camera normalization as for the target trajectory, so that the aligned anchor geometry and target-view geometry are expressed in a common normalized frame.

Geometry-aligned memory construction. For each observation $A \in \mathcal{O}$, we suppress its background using the character mask and encode the masked observation with the pretrained VAE to obtain $\boldsymbol{z}_A = \mathcal{E}(A)$. We further construct a Plücker ray grid from its aligned camera geometry and form the anchor representation as

$$h_A = \mathrm{PatchEmbed}(f_{\mathrm{anc}}([\boldsymbol{z}_A, \mathbf{1}_4, \boldsymbol{z}_A, \boldsymbol{M}_A]_{\mathrm{ch}})) + f_{\mathrm{cam}}(\boldsymbol{r}_{A,\ell}), \tag{12}$$

where $\mathbf{1}_4$ is a four-channel all-ones mask, $\boldsymbol{M}_A$ is the character mask at latent resolution, and $\boldsymbol{r}_{A,\ell}$ is the resized Plücker ray grid. The latent $\boldsymbol{z}_A$ occupies both latent slots of the pretrained I2V input layout, while f_{anc} projects the fused features to the patch-embedding input dimension. The camera projection f_{cam} is shared with the target-video branch in Sec. 3.3. Finally, the resulting tokens from all observations are concatenated to form the geometry-aligned character memory.

Geometry-aware memory retrieval. We insert 20 memory-attention blocks after backbone blocks $0, 2, \ldots, 38$. At insertion l, let $m^{(l)}$ and $h^{(l)}$ denote the memory and video tokens. Their concatenation is projected from 5120 to 1024 dimensions and timestep-modulated:

$$x^{(l)} = f_{\mathrm{in}}\left([m^{(l)}, h^{(l)}]_{\mathrm{seq}}\right), \qquad \widetilde{x}^{(l)} = \mathrm{AdaLN}_l\left(x^{(l)}, \tau\right), \tag{13}$$

where $[\cdot]_{\mathrm{seq}}$ denotes token concatenation. Each block performs joint self-attention with eight heads ($d_h = 128$). Queries, keys, and values are projected from $\widetilde{x}^{(l)}$, with RMS normalization and rotary positional encoding applied to queries and keys. The video-token output is:

$$o_{\mathrm{vid}}^{(l)} = \mathrm{softmax}\left(\frac{Q_{\mathrm{vid}}^{(l)}\left([K_{\mathrm{mem}}^{(l)}, K_{\mathrm{vid}}^{(l)}]_{\mathrm{seq}}\right)^\top}{\sqrt{d_h}}\right)[V_{\mathrm{mem}}^{(l)}, V_{\mathrm{vid}}^{(l)}]_{\mathrm{seq}}. \tag{14}$$

Video queries retrieve complementary character evidence from the geometry-aligned memory while retaining video context, with shared camera encoding making the retrieval view-aware.

Residual memory injection. Let F_l denote the complete memory block, including the feed-forward refinement and timestep-conditioned modulation. Its residual update is projected to the backbone dimension and injected only into the video tokens:

$$\Delta x^{(l)} = F_l(x^{(l)}, \tau) - x^{(l)}, \qquad h^{(l)} \leftarrow h^{(l)} + f_{\mathrm{out}}^{(l)}\left([\Delta x^{(l)}]_{\mathrm{vid}}\right). \tag{15}$$

Where $[\cdot]_{\mathrm{vid}}$ selects the video-token portion of the joint sequence. Each $f_{\mathrm{out}}^{(l)}$ is independently zero-initialized, preserving the pretrained backbone at the start of adaptation. The updated tokens then continue through the backbone with progressively integrated geometry-aligned character evidence.

3.5 Training Data Construction

We construct an 80,000-sample character-centric training corpus, comprising 50,000 UE5 camera-motion sequences and 30,000 Blender-rendered samples with multi-view character anchors, as summarized in Fig. 3. Both collections cover 99 characters and 33 environments. Using assets from Fab[1], UE5 combines diverse scenes and character actions with arc, large-angle orbit, orbit-and-dolly, and free-form camera paths. Using assets from Sketchfab[2], our configurable Blender pipeline automates character placement, camera configuration, preview, and batch rendering across varied initial viewpoints, angular spans, translations, orbit radii, and indoor and outdoor scenes. Each sample provides a reference video, initial RGB-D, camera intrinsics and extrinsics, rendered RGB guidance, coverage masks, and complementary front, back, and side observations for the geometric conditioning and character memory in Sec. 3.3 and Sec. 3.4. Further construction details and trajectory examples are provided in Appendix B.

[1] https://www.fab.com/listings/2e872a96-686e-4955-8256-82fa19904f92.
[2] https://sketchfab.com/.

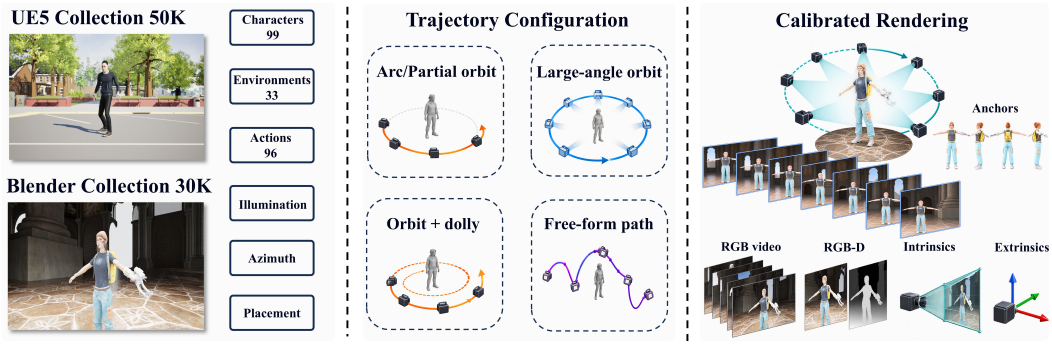


Figure 3: Overview of our training-data construction pipeline.

4 Experiments

4.1 Experimental Setup

Implementation details. We train our framework in two stages based on Wan2.1-I2V-14B. Stage 1 combines 76,698 rendered SpatialVID sequences (Wang et al., 2026b) with our 50,000 UE5 sequences for camera-control training. We update f_{geo}, the pretrained patch embedding, backbone self-attention layers, and Plücker camera projection f_{cam}, while freezing the remaining parameters. Stage 2 initializes from the Stage 1 checkpoint and trains GAMA on our 30,000 Blender samples with multi-view character anchors, optimizing only the memory-attention bank and anchor adapter.

Both stages follow the flow-matching objective in Eq. (2) with their respective conditioning inputs. We use AdamW with a constant learning rate of 10^{-4} and zero weight decay. Training uses 81-frame clips at 832×480 resolution. Stage 1 is trained for 5 epochs with a batch size of 64, and Stage 2 for 8 epochs with a batch size of 32.

Evaluation benchmark. We evaluate on ***CLACT***, comprising 42 character-centric scenes: 17 synthetic and 25 real-world cases. Each case is tested under six camera trajectories spanning large-angle orbits, translations, and combined motions, with different initial azimuths. The benchmark evaluates camera-control accuracy, cross-view character consistency, and visual quality under viewpoint changes.

Compared methods. We compare against three groups of methods. Direct camera-controlled video generation baselines include CamI2V (Zheng et al., 2024), Uni3C (Cao et al., 2025b), GEN3C (Ren et al., 2025), Lyra 2.0 (Shen et al., 2026), and DualCamCtrl (Zhang et al., 2025). We further compare with 3D-aware novel-view generation methods, including ViewCrafter (Yu et al., 2024) and MVGenMaster (Cao et al., 2025a), as well as video re-rendering methods ReCamMaster (Bai et al., 2025), DepthDirector (Chen et al., 2026), and CameraAnything (Li et al., 2026).

Evaluation metrics. We evaluate camera control using Absolute Trajectory Error (ATE), Relative Translation Error (RTE), and Relative Rotation Error (RRE) (Sturm et al., 2012). Character consistency is assessed using MEt3R for cross-view geometric consistency (Asim et al., 2025), ArcFace identity similarity (Deng et al., 2019), and CLIP and DINO similarities (Radford et al., 2021; Oquab et al., 2023). We report subject consistency (Subj.) and image quality (IQ) to assess visual fidelity (Huang et al., 2024).

Detailed benchmark configurations, baseline evaluation protocols, and metric definitions are provided in Appendix C.

4.2 Comparison with State-of-the-Art Methods

Quantitative comparison. As shown in Table 1, ***WorldCast*** achieves superior camera control and character consistency while maintaining competitive visual quality. The gains span trajectory accuracy, cross-view geometry, and identity and appearance preservation. Additional reference views yield limited or inconsistent improvements for ViewCrafter and MVGenMaster, reflecting the dif-

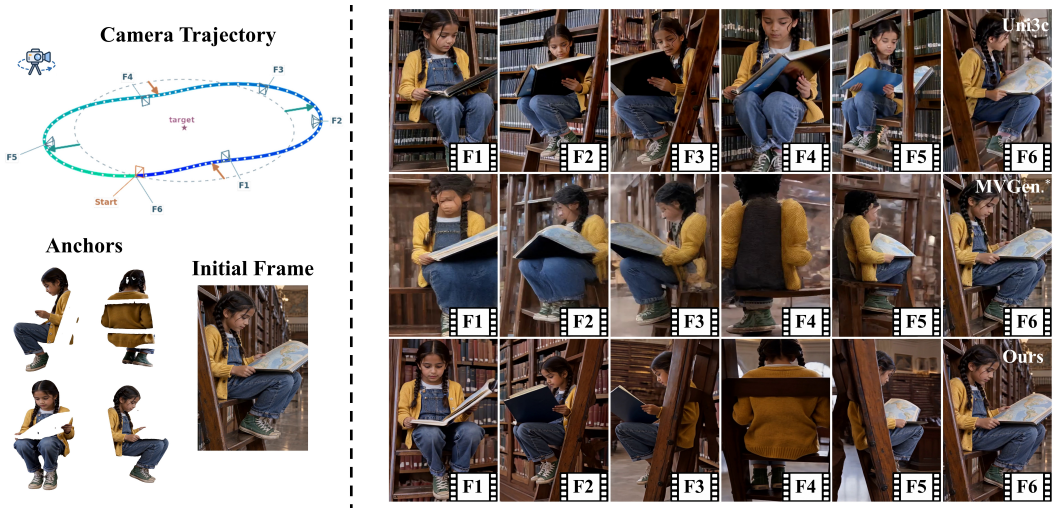


Figure 4: Qualitative comparison on large-angle camera-controlled video generation.

Table 1: **Quantitative comparison on our evaluation benchmark.** ↓ and ↑ denote lower- and higher-is-better metrics, respectively. Anchors denotes the use of additional character views.

Method	Anchors	Camera Control			Character Consistency				Image Quality	
		ATE ↓	RTE ↓	RRE ↓	MEt3R ↓	Arcf. ↑	CLIP ↑	DINO ↑	Subj. ↑	IQ ↑
Camera-controlled generation										
CamI2V (Zheng et al., 2024)	✗	0.351	0.281	11.490	0.405	0.147	0.804	0.675	0.874	0.686
DualCamCtrl (Zhang et al., 2025)	✗	0.297	0.082	3.607	0.408	0.185	0.796	0.661	0.904	0.699
GEN3C (Ren et al., 2025)	✗	0.130	0.038	1.255	0.306	0.433	0.819	0.696	0.915	0.716
Uni3C (Cao et al., 2025b)	✗	0.098	0.033	1.069	0.324	0.374	0.823	0.691	0.911	0.711
Lyra 2.0 (Shen et al., 2026)	✗	0.142	0.034	1.741	0.316	0.394	0.817	0.691	0.916	**0.725**
Novel-view generation										
ViewCrafter (Yu et al., 2024)	✗	0.353	0.203	9.198	0.395	0.223	0.810	0.683	0.888	0.715
ViewCrafter* (Yu et al., 2024)	✓	0.361	0.196	8.902	0.419	0.195	0.795	0.655	0.869	0.715
MVGenMaster (Cao et al., 2025a)	✗	0.125	0.099	2.929	0.370	0.227	0.796	0.655	0.869	0.699
MVGenMaster* (Cao et al., 2025a)	✓	0.218	0.152	6.938	0.301	0.299	0.821	0.707	0.862	0.691
Video re-rendering										
ReCamMaster (Bai et al., 2025)	✗	0.295	0.067	2.600	0.335	0.363	0.812	0.678	0.904	0.689
DepthDirector (Chen et al., 2026)	✗	0.289	0.066	1.732	0.332	0.423	0.826	0.686	**0.919**	0.705
CameraAnything (Li et al., 2026)	✗	0.244	0.078	4.879	0.341	0.358	0.807	0.648	0.898	0.700
WorldCast (Ours)	✓	**0.065**	**0.020**	**0.896**	**0.228**	**0.555**	**0.830**	**0.719**	0.917	0.719

ficulty of reliably aligning sparse cross-view geometry under large viewpoint changes. In contrast, our framework combines trajectory-aligned geometric guidance with geometry-aligned character memory, enabling more reliable large-angle camera control and character preservation.

Qualitative comparison. As shown in Fig. 4, we evaluate a challenging trajectory combining a large-angle orbit with dolly motion. We compare against Uni3C and MVGenMaster. Uni3C exhibits severe character distortion under large viewpoint changes, particularly in facial regions. MVGenMaster uses the initial frame together with multi-view references containing both character and scene content, yet still loses fine character details and appearance cues. In contrast, our approach follows the prescribed camera motion more reliably and better preserves character identity, 3D structure, and appearance throughout the trajectory.

4.3 Ablations

Effect of individual components. As shown in Table 2, the full configuration achieves the best performance across all metrics. Removing GAMA substantially degrades both camera control and character consistency, confirming the importance of multi-view character memory. Without anchor geometry alignment, rotation error increases markedly while identity similarity remains relatively

Table 2: **Ablation study of the proposed components.** We remove the entire GAMA branch, anchor geometry alignment, or anchor foreground masking while keeping the remaining settings unchanged.

Variant	Camera Control			Character Consistency				Visual Quality	
	ATE ↓	RTE ↓	RRE ↓	MEt3R ↓	Arcf. ↑	CLIP ↑	DINO ↑	Subj. ↑	IQ ↑
w/o GAMA	0.100	0.031	1.149	0.333	0.370	0.813	0.668	0.899	0.699
w/o Anchor Geometry Alignment	0.071	0.027	1.374	0.239	0.547	0.826	0.708	0.909	0.716
w/o Anchor Masking	0.078	0.025	1.110	0.247	0.546	0.827	0.710	0.908	0.709
WorldCast (Ours)	**0.065**	**0.020**	**0.896**	**0.228**	**0.555**	**0.830**	**0.719**	**0.917**	**0.719**

high, highlighting the importance of relating anchor observations to target viewpoints. Removing foreground masking also degrades camera control, character consistency, and visual quality, demonstrating the benefit of isolating character information from scene backgrounds.

Effect of UE5 data and camera-ray conditioning. We further examine the contributions of UE5 training data and Plücker ray conditioning to camera control. These ablations are conducted on the first-stage model before GAMA adaptation. As shown in Table 3, excluding the UE5 collection increases all three camera errors, demonstrating the benefit of large-angle training sequences. Removing Plücker conditioning also degrades camera accuracy, showing that explicit camera-ray information effectively complements the projected geometric guidance under large-angle camera motion.

Table 3: **First-stage camera-control ablation.**

Variant	Camera Control		
	ATE ↓	RTE ↓	RRE ↓
w/o UE5 data	0.359	0.156	4.295
w/o Plücker conditioning	0.275	0.056	2.113
Full Stage 1	**0.100**	**0.031**	**1.149**

Effect of anchor count. As shown in Fig. 5, increasing the number of character anchors from one to four improves cross-view geometry, identity preservation, and appearance similarity under large viewpoint changes. This indicates that complementary observations help recover character information missing from the initial view and provide richer cross-view evidence. The smaller gain from three to four anchors suggests diminishing returns as additional views contribute less new information.

Additional discussions, supplementary experiments, and more qualitative visualizations are provided in Appendix D.

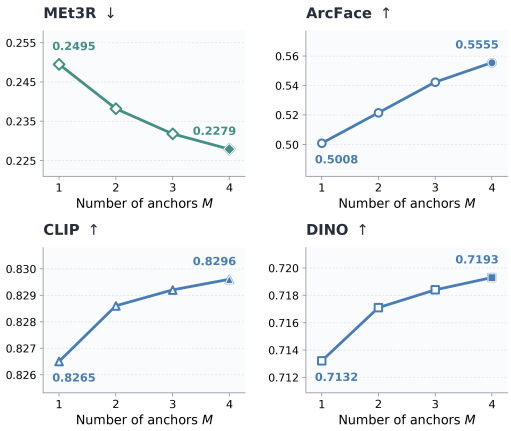


Figure 5: Effect of the number of character anchors on character consistency.

5 Conclusion

We presented ***CLACT***, a character-centric dataset and benchmark for large-angle camera-controlled video generation, together with ***WorldCast***, a framework that jointly improves trajectory accuracy and character preservation under substantial viewpoint changes. The key idea is to combine scale-aligned geometric guidance with Geometry-Aligned Memory Attention (GAMA), which aligns multi-view character observations with the target camera geometry and retrieves view-relevant character features during generation. Experiments on both synthetic and real-world scenes demonstrate more accurate camera control and stronger preservation of character geometry, appearance, and identity, while maintaining competitive visual quality. Ablation studies further confirm the importance of large-angle training data, explicit camera-ray conditioning, geometric anchor alignment, and multi-view character memory. Future work will explore more flexible anchor acquisition, more complex dynamic scenes, and longer or interactive camera trajectories.

AI USE STATEMENT

Generative AI tools were used to assist with language polishing, translation, and the revision of scientific writing, including improving clarity, grammar, and presentation. All AI-assisted content was reviewed and revised by the authors. Technical claims, mathematical formulations, experimental settings, citations, and implementation details were verified against the corresponding source materials and our implementation. Generative AI was not used to autonomously conduct experiments, determine experimental results, or make scientific conclusions. The authors take full responsibility for the final content of this work, including all text, claims, code, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies controllable character video generation and therefore has potential applications in visual-content creation as well as potential risks associated with synthetic human imagery. Our training benchmark is constructed primarily from controlled simulations using licensed or publicly available 3D character and environment assets. Any released real-world evaluation data will exclude personally identifiable private individuals and will follow the applicable data licenses and usage terms. Although our method is designed for research on camera control and character consistency, character-generation technology could be misused to create misleading or impersonating content. We therefore discourage the synthesis of deceptive depictions of real individuals without their knowledge or consent and will provide provenance and documentation for released benchmark assets and generated examples where applicable.

REPRODUCIBILITY STATEMENT

We provide the architectural design, conditioning formulation, training strategy, optimization settings, and evaluation protocol in the main paper and appendices. The supplementary material further documents the construction of the UE5 and Blender datasets, camera and depth alignment procedures, baseline input adaptations, evaluation metrics, and additional ablations. The benchmark generator exports camera intrinsics, camera-to-world poses, depth maps, multi-view character anchors, reference videos, point-cloud renders, and coverage masks. Upon publication, we plan to release the benchmark generation and evaluation code, pose-alignment and identity-evaluation procedures, trajectory configurations, training and inference configurations, and the settings required to reproduce results with the released checkpoint.

REFERENCES

Mohammad Asim, Christopher Wewer, Thomas Wimmer, Bernt Schiele, and Jan Eric Lenssen. Met3r: Measuring multi-view consistency in generated images. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6034–6044. IEEE, 2025.

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, 2025a.

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*, volume 2025, pp. 66712–66737, 2025b.

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 *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14834–14844. IEEE, 2025.

Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion

models. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22563–22575. IEEE, 2023.

Alexey Bochkovskiy, Amaël Delaunoy, Hugo Germain, Marcel Santos, Yichao Zhou, Stephan Richter, and Vladlen Koltun. Depth pro: Sharp monocular metric depth in less than a second. In *International Conference on Learning Representations*, volume 2025, pp. 75602–75637, 2025.

Chenjie Cao, Chaohui Yu, Shang Liu, Fan Wang, Xiangyang Xue, and Yanwei Fu. Mvgenmaster: Scaling multi-view generation from any image via 3d priors enhanced diffusion model. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6045–6056. IEEE, 2025a.

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

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris Coll-Vinent, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. In *International conference on learning representations*, volume 2026, pp. 138846–138923, 2026.

Eric R Chan, Koki Nagano, Matthew A Chan, Alexander W Bergman, Jeong Joon Park, Axel Levy, Miika Aittala, Shalini De Mello, Tero Karras, and Gordon Wetzstein. Generative novel view synthesis with 3d-aware diffusion models. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4194–4206. IEEE, 2023.

Dong-Yu Chen, Yixin Guo, Shuojin Yang, Tai-Jiang Mu, and Shi-Min Hu. Beyond inpainting: Unleash 3d understanding for precise camera-controlled video generation. *arXiv preprint arXiv:2601.10214*, 2026.

Robin Courant, Nicolas Dufour, Xi Wang, Marc Christie, and Vicky Kalogeiton. Et the exceptional trajectories: Text-to-camera-trajectory generation with character awareness. In *European conference on computer vision*, pp. 464–480. Springer, 2024.

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*, pp. 4690–4699, 2019.

Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24 (6):381–395, 1981.

Ruiqi Gao, Aleksander Holynski, Philipp Henzler, Arthur Brussee, Ricardo Martin-Brualla, Pratul Srinivasan, Jonathan T Barron, and Ben Poole. Cat3d: Create anything in 3d with multi-view diffusion models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 75468–75494, 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*, volume 2025, pp. 100433–100464, 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.

Yuming Jiang, Tianxing Wu, Shuai Yang, Chenyang Si, Dahua Lin, Yu Qiao, Chen Change Loy, and Ziwei Liu. Videobooth: Diffusion-based video generation with image prompts. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6689–6700. IEEE, 2024.

Teng Li, Guangcong Zheng, Rui Jiang, Shuigen Zhan, Tao Wu, Yehao Lu, Yining Lin, Chuanyun Deng, Yepan Xiong, Min Chen, et al. Realcam-i2v: Real-world image-to-video generation with interactive complex camera control. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 28785–28796. IEEE, 2025.

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. *arXiv preprint arXiv:2607.24591*, 2026.

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

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

Bruce D Lucas and Takeo Kanade. An iterative image registration technique with an application to stereo vision. In *IJCAI'81: 7th international joint conference on Artificial intelligence*, volume 2, pp. 674–679, 1981.

Ze Ma, Daquan Zhou, Xue-She Wang, Chun-Hsiao Yeh, Xiuyu Li, Huanrui Yang, Zhen Dong, Kurt Keutzer, and Jiashi Feng. Magic-me: Identity-specific video customized diffusion. In *European Conference on Computer Vision*, pp. 19–37. Springer, 2024.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

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

Tianchang Shen, Sherwin Bahmani, Kai He, Sangeetha Grama Srinivasan, Tianshi Cao, Jiawei Ren, Ruilong Li, Zian Wang, Nicholas Sharp, Zan Gojcic, et al. Lyra 2.0: Explorable generative 3d worlds. *arXiv preprint arXiv:2604.13036*, 2026.

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.

Jürgen Sturm, Wolfram Burgard, and Daniel Cremers. Evaluating egomotion and structure-from-motion approaches using the tum rgb-d benchmark. In *Proc. of the Workshop on Color-Depth Camera Fusion in Robotics at the IEEE/RJS International Conference on Intelligent Robot Systems (IROS)*, volume 13, pp. 6, 2012.

Shinji Umeyama. Least-squares estimation of transformation parameters between two point patterns. *IEEE Transactions on pattern analysis and machine intelligence*, 13(4):376–380, 1991.

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.

Cong Wang, Zhentao Yu, Hongmei Wang, Weicong Liang, Zixiang Zhou, Zilin Yang, Jiarong Ou, Rui Chen, Yuan Zhou, and Qinglin Lu. Harmoview: Harmonizing multi-view constraints for identity-consistent video generation. *arXiv preprint arXiv:2606.10839*, 2026a.

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, 2026b.

Jianyuan Wang, Minghao Chen, Shangzhan Zhang, Nikita Karaev, Johannes Lutz Schönberger, Patrick Labatut, Piotr Bojanowski, David Novotny, Andrea Vedaldi, and Christian Rupprecht. VGGT-Ω. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21486–21499, June 2026c.

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, 2024.

Daniel Watson, William Chan, Ricardo Martin-Brualla, Jonathan Ho, Andrea Tagliasacchi, and Mohammad Norouzi. Novel view synthesis with diffusion models. In *International Conference on Learning Representations*, 2023.

Yujie Wei, Shiwei Zhang, Zhiwu Qing, Hangjie Yuan, Zhiheng Liu, Yu Liu, Yingya Zhang, Jingren Zhou, and Hongming Shan. Dream video: Composing your dream videos with customized subject and motion. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6537–6549. IEEE, 2024.

Mingyang Wu, Ashirbad Mishra, Soumik Dey, Shuo Xing, Naveen Ravipati, Hansi Wu, Binbin Li, and Zhengzhong Tu. Consid-gen: View-consistent and identity-preserving image-to-video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1853–1863, June 2026.

Dejia Xu, Weili Nie, Chao Liu, Sifei Liu, Jan Kautz, Zhangyang Wang, and Arash Vahdat. Camco: Camera-controllable 3d-consistent image-to-video generation. *arXiv preprint arXiv:2406.02509*, 2024.

Yuancheng Xu, Wenqi Xian, Li Ma, Julien Philip, Ahmet Taşel, Yiwei Zhao, Ryan Burgert, Mingming He, Oliver Hermann, Oliver Pilarski, et al. Virtually being: Customizing camera-controllable video diffusion models with volumetric performance captures. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, pp. 1–12, 2025.

Liudi Yang, George Eskandar, Fengyi Shen, Mohammad Altillawi, Yang Bai, Chi Zhang, Ziyuan Liu, and Abhinav Valada. Confctrl: Enabling precise camera control in video diffusion via confidence-aware interpolation. *arXiv preprint arXiv:2603.09819*, 2026a.

Yuhang Yang, Fan Zhang, Huaijin Pi, Ailing Zeng, Shuai Guo, Guowei Xu, Wei Zhai, Yang Cao, and Zheng-Jun Zha. Gloria: Consistent character video generation via content anchors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 36724–36735, 2026b.

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

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

Shenghai Yuan, Jinfa Huang, Xianyi He, Yunyang Ge, Yujun Shi, Liuhan Chen, Jiebo Luo, and Li Yuan. Identity-preserving text-to-video generation by frequency decomposition. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12978–12988. IEEE, 2025.

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, 2026a.

Hongfei Zhang, Kanghao Chen, Zixin Zhang, Harold Haodong Chen, Yuanhuiyi Lyu, Yuqi Zhang, Shuai Yang, Kun Zhou, and Yingcong Chen. Dualcamctrl: Dual-branch diffusion model for geometry-aware camera-controlled video generation. *arXiv preprint arXiv:2511.23127*, 2025.

Yisu Zhang, Chenjie Cao, Tengfei Wang, Xuhui Zuo, Junta Wu, Jianke Zhu, and Chunchao Guo. Worldstereo: Bridging camera-guided video generation and scene reconstruction via 3d geometric memories. *arXiv preprint arXiv:2603.02049*, 2026b.

Guangcong Zheng, Teng Li, Rui Jiang, Yehao Lu, Tao Wu, and Xi Li. Cami2v: Camera-controlled image-to-video diffusion model. *arXiv preprint arXiv:2410.15957*, 2024.

Jensen Zhou, Hang Gao, Vikram Voleti, Aaryaman Vasishta, Chun-Han Yao, Mark Boss, Philip Torr, Christian Rupprecht, and Varun Jampani. Stable virtual camera: Generative view synthesis with diffusion models. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12405–12414. IEEE, 2025.

A Depth–Trajectory Scale Calibration

This section details the reprojection geometry and motion-estimation procedure for the scale calibration in Sec. 3.3. Predicted depth and camera translation may follow different scale conventions. We calibrate their relative scale by matching the motion induced by point-cloud rendering to that of the training reference video.

Reprojection geometry. Let D_0 denote the predicted depth of the initial frame, and let $\boldsymbol{R}_t$ and $\boldsymbol{c}_t$ denote the camera-to-world rotation and camera center at frame t, respectively. For an initial-view pixel $\boldsymbol{p} = [u, v]^\top$, we define its homogeneous coordinate as $\widetilde{\boldsymbol{p}} = [u, v, 1]^\top$. The corresponding 3D point in the initial camera frame is

$$\boldsymbol{q}_0(\boldsymbol{p}) = D_0(\boldsymbol{p})\boldsymbol{K}_0^{-1}\widetilde{\boldsymbol{p}}. \tag{A.1}$$

Following Eq. (4), the camera translation is rescaled while the rotations and predicted depth remain fixed. The point is transformed from the initial camera frame to the target camera frame as

$$\begin{aligned}\boldsymbol{q}_t^{(s)}(\boldsymbol{p}) &= \boldsymbol{R}_t^\top \left(\boldsymbol{R}_0\boldsymbol{q}_0(\boldsymbol{p}) + \boldsymbol{c}_0 - \boldsymbol{c}_t^{(s)}\right) \\ &= \boldsymbol{R}_t^\top \boldsymbol{R}_0\boldsymbol{q}_0(\boldsymbol{p}) + as\boldsymbol{R}_t^\top(\boldsymbol{c}_0 - \boldsymbol{c}_t).\end{aligned} \tag{A.2}$$

Using perspective projection $\pi([x, y, z]^\top) = [x/z, y/z]^\top$, the corresponding target pixel is

$$\boldsymbol{p}_t^{(s)} = \pi\left(\boldsymbol{K}_t\boldsymbol{q}_t^{(s)}(\boldsymbol{p})\right). \tag{A.3}$$

Substituting Eq. (A.1) and exploiting the scale invariance of perspective projection gives

$$\boldsymbol{p}_t^{(s)} = \pi\left(\boldsymbol{K}_t\left[\boldsymbol{R}_t^\top\boldsymbol{R}_0\boldsymbol{K}_0^{-1}\widetilde{\boldsymbol{p}} + \frac{as}{D_0(\boldsymbol{p})}\boldsymbol{R}_t^\top(\boldsymbol{c}_0 - \boldsymbol{c}_t)\right]\right). \tag{A.4}$$

This expression makes the relevant scale ambiguity explicit: for fixed rotations and intrinsics, the translational component of image motion depends on camera displacement relative to scene depth. Varying s therefore changes the induced parallax without altering the predicted depth structure or camera rotations.

Motion estimation. We back-project the initial frame into a colored point cloud. For each candidate scale $s \in \mathcal{S}$, we render this point cloud along the rescaled trajectory, obtaining $\mathcal{G}^{(s)} = (G_t^{(s)})_{t=0}^{N-1}$. Point splatting with depth buffering handles visibility. We then estimate image motion independently from the rendered sequence $\mathcal{G}^{(s)}$ and the training reference sequence $\mathcal{I}^{\mathrm{ref}} = (I_t^{\mathrm{ref}})_{t=0}^{N-1}$ at the same preview resolution.

We sample frame pairs $(t_i, t_i + \Delta)$, where $t_i = i\Delta$, $t_i + \Delta < N$, and $\Delta = 4$ by default. For each pair, up to 1200 corners are detected in the first frame and tracked to the second using pyramidal Lucas–Kanade optical flow (Lucas & Kanade, 1981). Pairs with fewer than 16 detected corners or valid tracks are discarded. To suppress unreliable correspondences, we fit a partial affine model with RANSAC (Fischler & Bolles, 1981) using a 3-pixel reprojection threshold. RANSAC inliers are retained when at least 12 are available; otherwise, all valid tracks are used.

For a sequence V, let $\boldsymbol{p}_{i,k}(V)$ and $\boldsymbol{p}'_{i,k}(V)$ denote the coordinates of retained track k in the two frames of pair i, and let $\mathcal{J}_i(V)$ denote the corresponding track-index set. We define the motion magnitude of the pair as

$$m_i(V) = \operatorname*{median}_{k\in\mathcal{J}_i(V)} \left\|\boldsymbol{p}'_{i,k}(V) - \boldsymbol{p}_{i,k}(V)\right\|_2. \tag{A.5}$$

Applying the same estimator to both sequences yields $m_i^{\mathrm{render}}(s) = m_i(\mathcal{G}^{(s)})$ and $m_i^{\mathrm{ref}} = m_i(\mathcal{I}^{\mathrm{ref}})$.

Scale selection. For each candidate s, we retain frame pairs whose motion estimates are finite and larger than $\epsilon = 10^{-3}$ pixels:

$$\mathcal{V}_s = \left\{i \;\middle|\; m_i^{\mathrm{render}}(s) \in (\epsilon, \infty),\ m_i^{\mathrm{ref}} \in (\epsilon, \infty)\right\}. \tag{A.6}$$

We evaluate Eq. (5) over the candidate set

$$\mathcal{S} = \{0.05, 0.075, 0.10, 0.15, 0.20, 0.25, 0.33, 0.40, 0.50\}, \tag{A.7}$$

and assign an infinite score to candidates with fewer than four valid frame pairs. The absolute log ratio measures multiplicative motion mismatch symmetrically, while the median reduces sensitivity to atypical frame pairs.

The selected scale s^* is substituted into Eq. (4) to construct the geometric conditions for the training clip. During training, inconsistent depth and translation scales cause the rendered parallax to deviate from that of the ground-truth video, weakening the alignment between geometric guidance and supervision. Calibration reduces this discrepancy by matching their image-motion magnitudes. At inference, camera translation is interpreted relative to the estimated scene depth, so the initial-view point cloud and requested trajectory jointly define the target parallax. No reference-based scale calibration is required, as there is no independently observed video motion to match.

B Training Data Construction

We construct two complementary training collections: UE5 sequences for large-angle camera-control learning and Blender samples with multi-view anchors for character-memory learning. Table B.1 summarizes their training roles and camera configurations, while Fig. B.1 shows their scale, asset statistics, and representative trajectory settings.

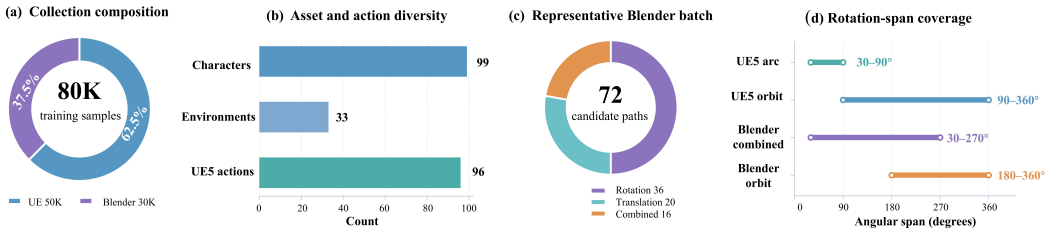


Figure B.1: **Training-data composition and camera configurations.** (a) Sizes and training roles of the UE5 and Blender collections. (b) Character and environment counts shared by both collections, together with the number of UE5 character actions. (c) A representative Blender batch containing 36 rotational, 20 translational, and 16 combined trajectories. (d) Angular-span configurations of the camera trajectories.

B.1 UE5 Character-Centric Camera-Motion Data

Scenes and character actions. We render 50,000 character-centric sequences using 99 character models, 33 environments, and 96 character actions. The environments span indoor and outdoor settings, including city streets, shopping malls, cafes, offices, and rural scenes, while the actions include motions such as waving, dancing, and cheering. Each sequence contains a single character, with the character, action, environment, and camera configuration varied across the collection. For each character–environment configuration, we additionally render front, back, left, and right views as complementary observations.

Camera trajectories. We design four trajectory families: arcs spanning 30°–90°, large-angle orbits spanning 90°–360°, orbit-and-dolly paths combining 90°–360° sweeps with camera-distance variation, and free-form trajectories generated by B-spline interpolation between multiple keyframes. Both constant- and variable-speed camera motions are included. The configurations of the UE5 and Blender collections are summarized in Table B.1, and representative angular spans are shown in Fig. B.1(d).

B.2 Blender Data with Multi-View Character Anchors

Asset preparation and scene placement. We generate 30,000 samples using an automated Blender pipeline with Cycles rendering. The assets include indoor and outdoor environments as well

Table B.1: Comparison of the UE5 and Blender training collections.

Configuration	UE5	Blender
Training samples	50,000	30,000
Training stage	Large-angle camera control	Multi-view memory injection
Anchor conditioning	Not used	Front, back, left, and right views
Camera paths	Arcs, orbits, orbit-and-dolly, and free-form paths	Orbits, translations, and combined rotation–translation paths
Motion variation	Angular span, camera distance, and speed	Initial azimuth, angular span, translation, and orbit radius

Both collections use 99 character models and 33 indoor and outdoor environments.

as static, rigged, and animated characters. The pipeline computes scene bounds, rescales characters, and evaluates candidate placements based on ground height, visibility, and clearance. Character placement and camera framing are adjusted to maintain ground contact and keep the subject within view. Orbit centers are derived from the placed character geometry, while camera radius and focal length are selected according to character size and scene constraints.

Trajectory generation. Camera trajectories are parameterized by initial azimuth, angular span, orbit radius, and translation direction and distance. The pipeline supports 180°–360° orbits, partial orbits from different initial viewpoints, pure translations, combined rotation–translation paths, and smooth radius variation. A representative batch contains 36 rotational, 20 translational, and 16 combined paths, yielding 72 candidate trajectories per character–environment pair, as illustrated in Fig. B.1(c).

Multi-view anchors and geometric outputs. We render front, back, left, and right observations of each character as multi-view anchors, together with the reference video, initial RGB image, initial-view depth, and per-frame camera intrinsics and extrinsics. The initial RGB-D observation is back-projected into a colored point cloud and rendered along the recorded trajectory to produce projected RGB guidance and masks indicating regions unsupported by the initial view. The initial image and multi-view anchors are further processed by the camera-estimation and geometric-alignment procedure in Sec. 3.4 to construct the geometry-aligned character memory.

Preview and batch rendering. A configurable web interface supports asset selection, camera adjustment, four-view preview, and batch rendering. The previews enable inspection of framing, occlusion, and ground contact before full-sequence rendering. Persistent Blender workers reuse loaded assets across tasks, while asset identifiers, camera parameters, and rendering settings are recorded for reproducibility.

C Benchmark and Evaluation Details

This section provides additional details of the evaluation benchmark, baseline input configurations, and metric protocol used in Sec. 4.

C.1 Benchmark Construction

Evaluation scenes. The evaluation set contains 42 character-centric scenes, including 17 synthetic scenes rendered in UE5 and 25 collected from real-world imagery. The two subsets provide complementary evaluation settings: synthetic cases offer controlled character and scene configurations, while real-world cases assess generalization to natural images and diverse visual content. All evaluation cases are disjoint from the training collections.

Camera trajectories. Each synthetic or real-world scene is evaluated under six prescribed camera trajectories covering large-angle orbits, pure translations, and combined rotation–translation motions. These trajectories span substantial viewpoint changes and diverse initial azimuths, varying the

character content observed in the initial view and enabling joint evaluation of trajectory adherence and character consistency as previously unseen regions become visible.

Evaluation inputs. Each evaluation case provides an initial character image and a prescribed camera trajectory. For methods that support or require multiple reference views, we additionally provide multi-view character observations. Our method takes the initial image together with the corresponding multi-view character anchors. All compared methods follow the same target trajectories, with camera parameters converted to their native representations when necessary. Generation resolution and sequence length follow the native settings of each method.

C.2 Baseline Evaluation Protocol

The compared methods span three generation paradigms. We adapt the benchmark inputs to their native formulations while preserving the prescribed target camera motion.

Direct camera-controlled generation. Camera-controlled image-to-video methods are provided with the initial image and prescribed camera trajectory following their native input formats.

Novel-view generation. For novel-view synthesis methods, we evaluate both their native single-image setting and, where supported, a multi-image setting with character anchors. For the latter, we follow each method's native multi-view formulation and provide the initial image together with the character anchors as reference views. Target views are then synthesized along the prescribed camera trajectory. This setting allows us to separately examine the effect of additional reference observations when comparing with our multi-view character conditioning.

Video re-rendering. Video re-rendering methods take a source video rather than a single initial image as input. We therefore construct a static source video by repeating the initial image along the temporal dimension and apply the same prescribed camera trajectory. This setup isolates camera-induced changes from source-video motion and provides a common static-input setting for comparison with image-conditioned methods.

Generation settings. Each method is evaluated using the generation resolution and sequence length supported by its native implementation. All generated outputs are then evaluated using the same quantitative evaluation pipeline.

C.3 Evaluation Metrics

We evaluate each method along three complementary dimensions: camera-control accuracy, character consistency, and visual quality.

Camera-control accuracy. We recover the camera trajectory of each generated video using Depth Anything 3 (Lin et al., 2025). To remove the global coordinate-frame and scale ambiguity of monocular pose estimation, we align each predicted trajectory to its target trajectory using a Sim(3) Umeyama transformation. Following standard trajectory-evaluation practice (Sturm et al., 2012), we compute Absolute Trajectory Error (ATE), Relative Translation Error (RTE), and Relative Rotation Error (RRE). ATE measures the RMSE between aligned and target camera centers, while RTE and RRE measure the translational and rotational errors of consecutive relative poses, respectively. Lower values indicate better camera adherence.

Character consistency. We use MEt3R (Asim et al., 2025) to measure cross-view geometric consistency using all generated frames together with the four observation anchors. We further evaluate identity and appearance consistency using ArcFace (Deng et al., 2019), CLIP (Radford et al., 2021), and DINO (Oquab et al., 2023) similarities. To account for large viewpoint changes, for each generated frame, we select the observation anchor with the closest camera pose and compute its similarity to that anchor. The resulting similarities are averaged over all frames and videos. Higher ArcFace, CLIP, and DINO similarities indicate stronger consistency, while lower MEt3R indicates better geometric consistency.

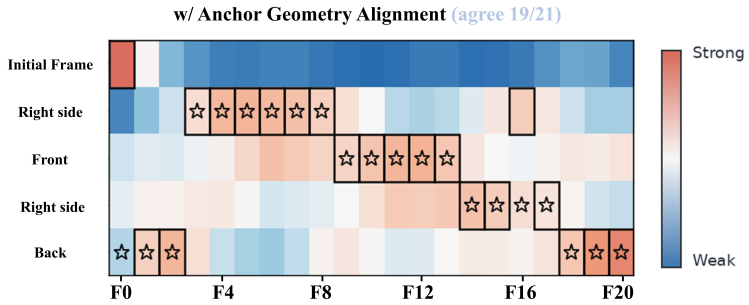


Figure D.1: **View-aware anchor routing under a 360° camera orbit.** Attention responses are visualized for the 21 temporal latent positions corresponding to an 81-frame generated video. Stars denote the geometrically expected observations with the closest camera poses to the target viewpoints, while outlined cells indicate the observations receiving the strongest attention responses. The predicted routing agrees with the expected routing for $19/21$ latent positions.

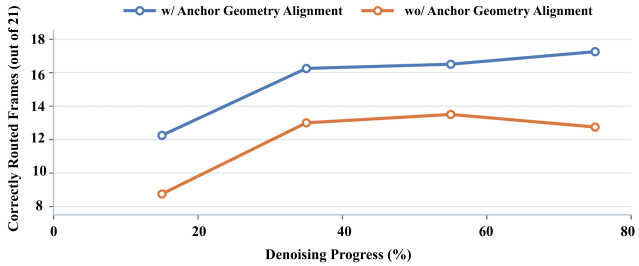


Figure D.2: **Effect of Anchor Geometry Alignment on attention routing.** Routing accuracy during training with and without anchor geometry alignment. Geometry alignment enables faster learning of view-aware anchor retrieval and achieves higher routing accuracy after convergence.

Visual quality. We additionally report subject consistency (Subj.) and image quality (IQ) following their default evaluation settings (Huang et al., 2024) to assess the overall visual fidelity of generated videos.

D Additional Analyses and Visualizations

This section provides additional experimental analyses and qualitative visualizations to further evaluate the proposed framework.

D.1 Analysis of Geometry-Aligned Anchor Routing

View-aware anchor routing. We analyze how the generated video features retrieve character information from different observations under a 360° camera orbit. As shown in Fig. D.1, the 81 generated frames correspond to 21 temporal latent positions, each attending to the initial frame and multi-view character anchors. For each latent position, the star marks the expected anchor whose camera pose is closest to the target viewpoint, while the outlined cell denotes the observation receiving the strongest attention response. We observe agreement for 19 of the 21 latent positions. Moreover, as the camera moves around the character, the dominant attention response progressively shifts across anchors with different viewpoints, showing that the memory branch retrieves character information according to the evolving target camera geometry.

Effect of anchor geometry alignment on routing. We further compare the learning dynamics with and without Anchor Geometry Alignment. As shown in Fig. D.2, geometry alignment leads to faster improvement in routing accuracy during training and achieves a higher final accuracy. This suggests that explicitly relating anchor viewpoints to the target camera geometry provides a stronger geometric prior for learning view-aware memory retrieval.

D.2 Additional Qualitative Results

Comparison across diverse trajectories. Fig. D.3–Fig. D.7 present additional comparisons across diverse camera trajectories, including large-angle orbits, orbit-and-dolly motions, spiral and hyperbolic paths, as well as conventional camera motions. The examples also cover different initial azimuths, further evaluating camera control and character preservation under substantial viewpoint changes.

Generalization across camera trajectories. We further evaluate trajectory generalization while keeping the same set of character anchors fixed. As shown in Fig. D.8 and Fig. D.9, the same multi-view observations support diverse target camera motions, including orbits, translations, and combined motions. This shows that the character anchors are not tied to a particular camera path and can be reused across different target trajectories.

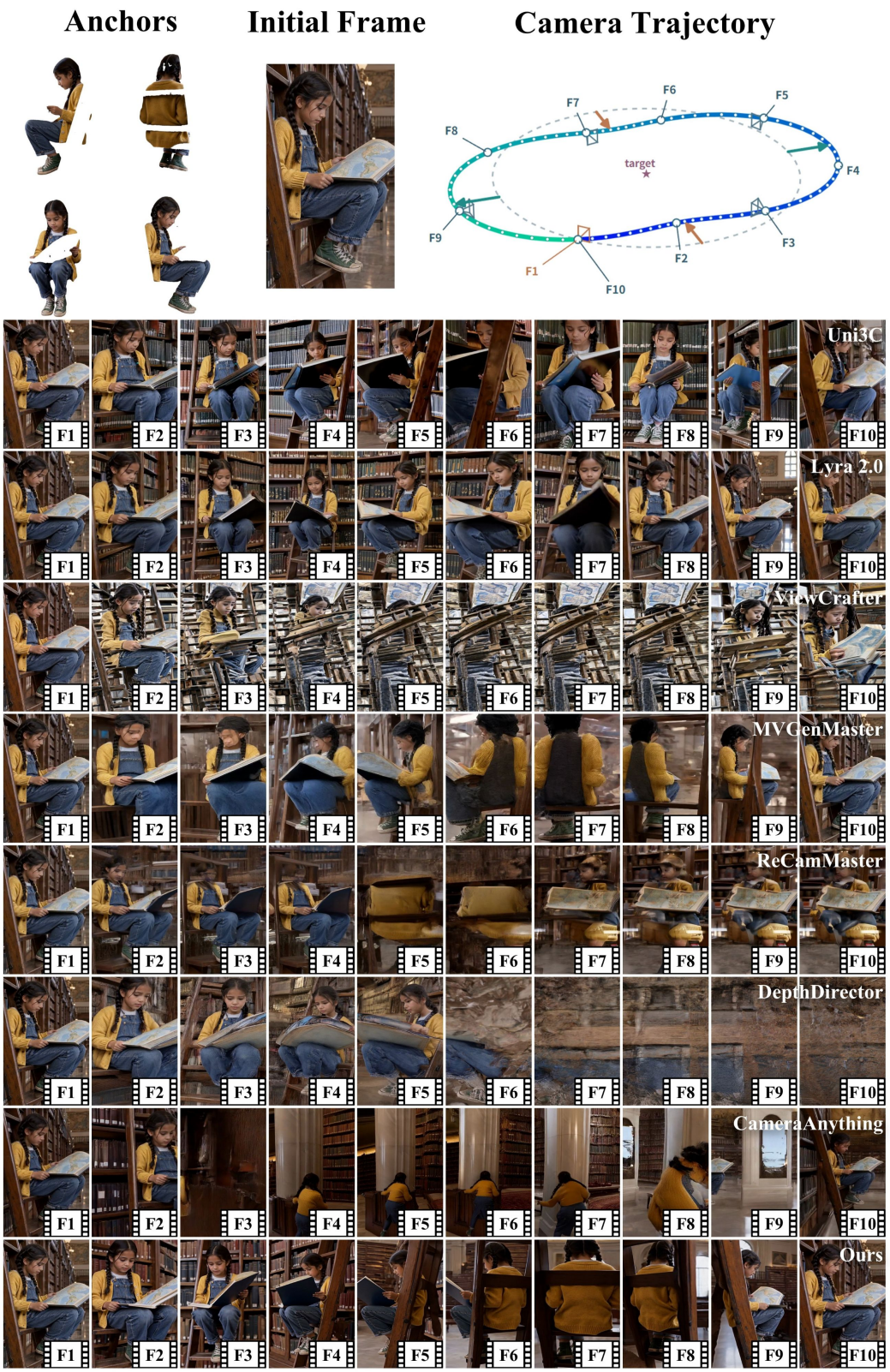


Figure D.3: Additional qualitative comparison with state-of-the-art methods under diverse large-angle camera motions.

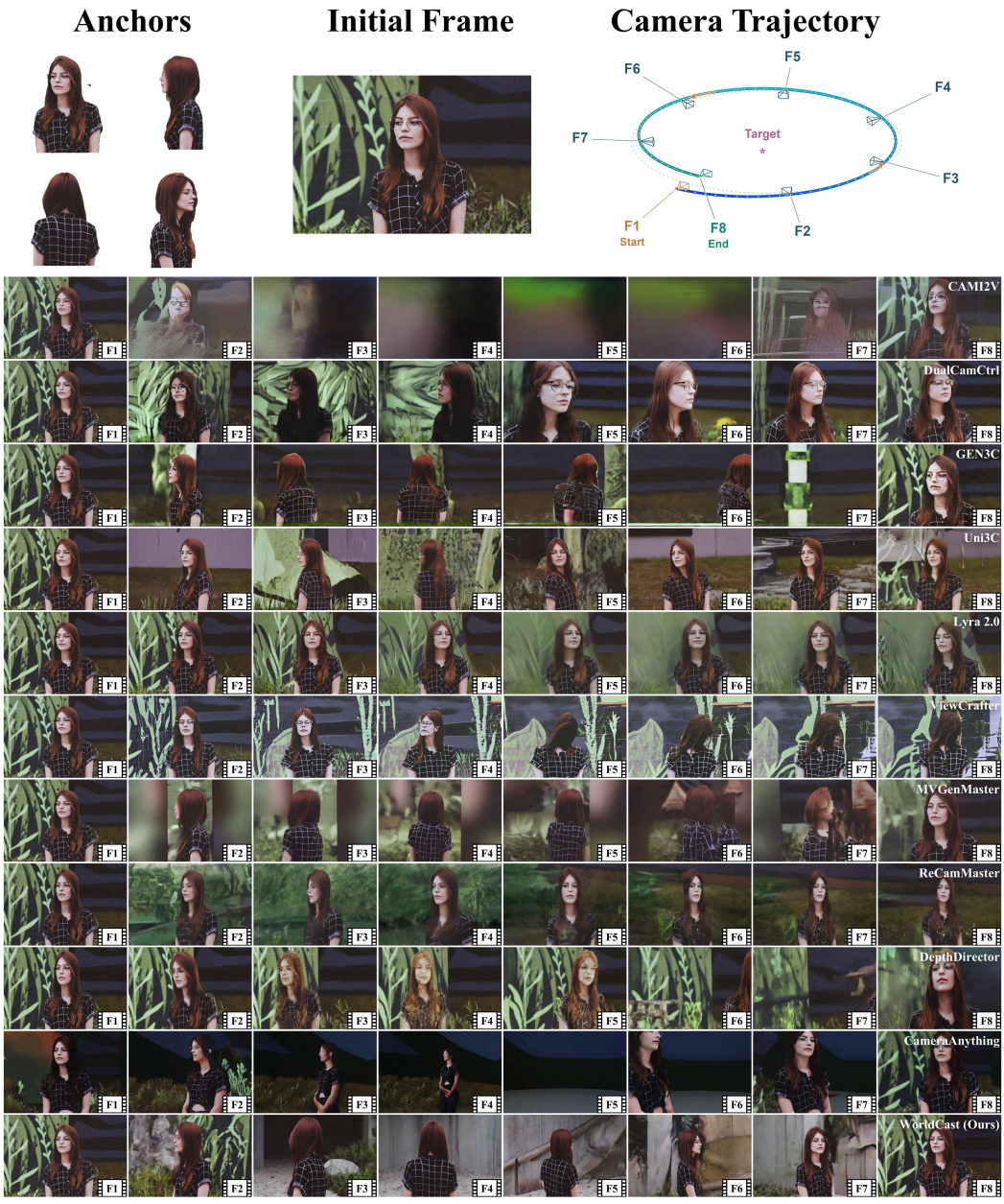


Figure D.4: Additional qualitative comparison under large-angle camera motion with different initial viewpoints.

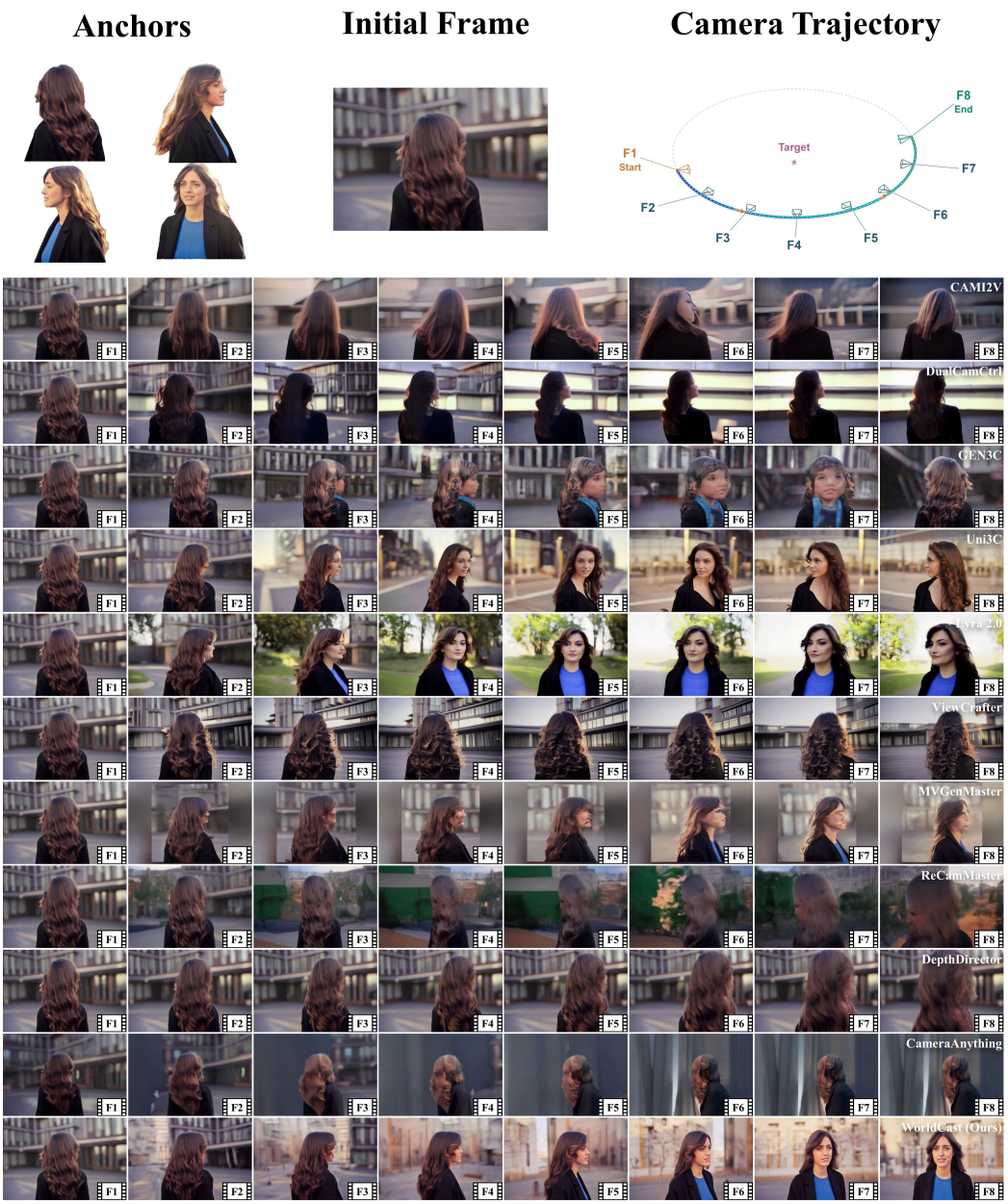


Figure D.5: Additional qualitative comparison on combined rotation and translation trajectories.

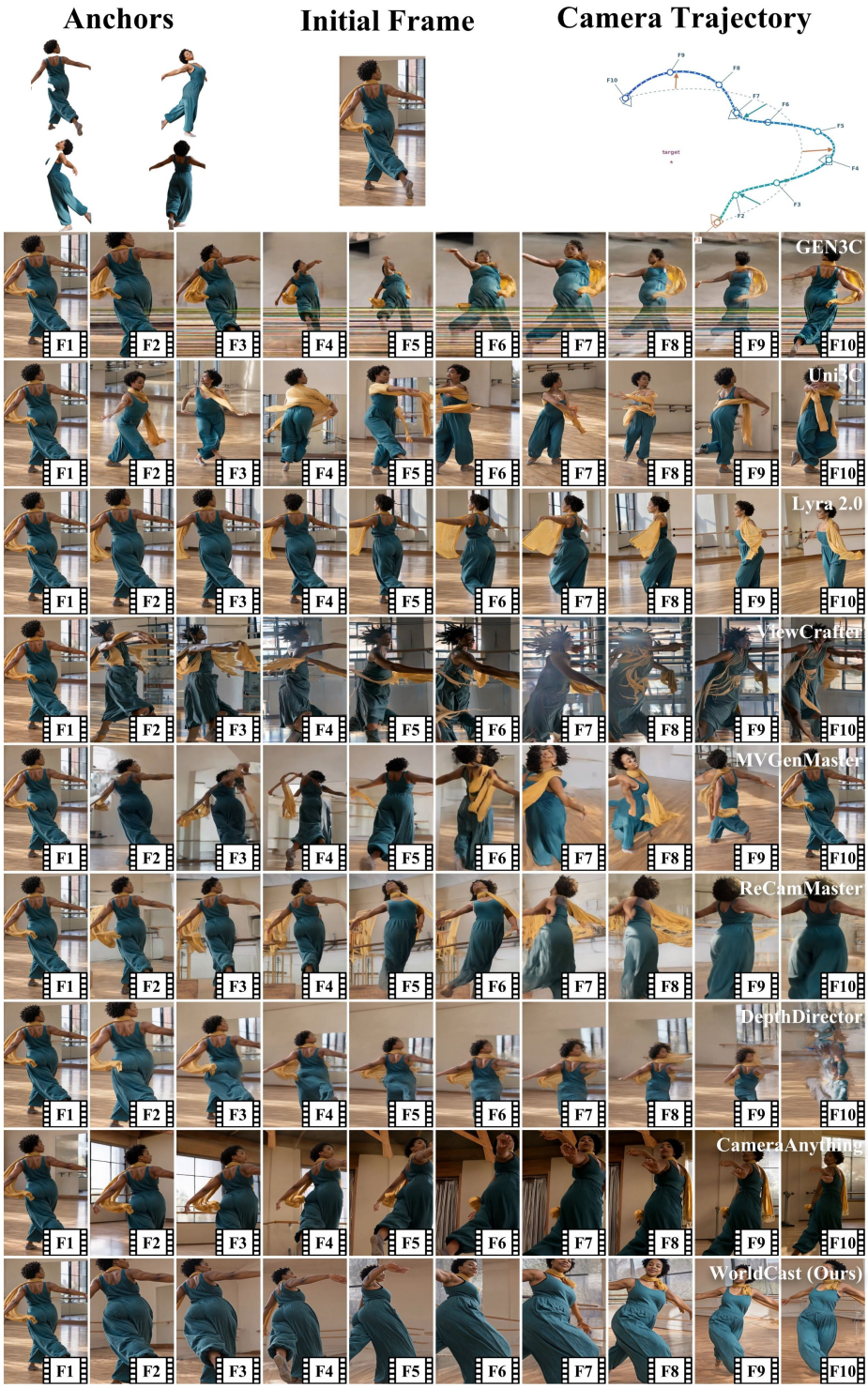


Figure D.6: Additional qualitative comparison on complex large-angle camera trajectories.

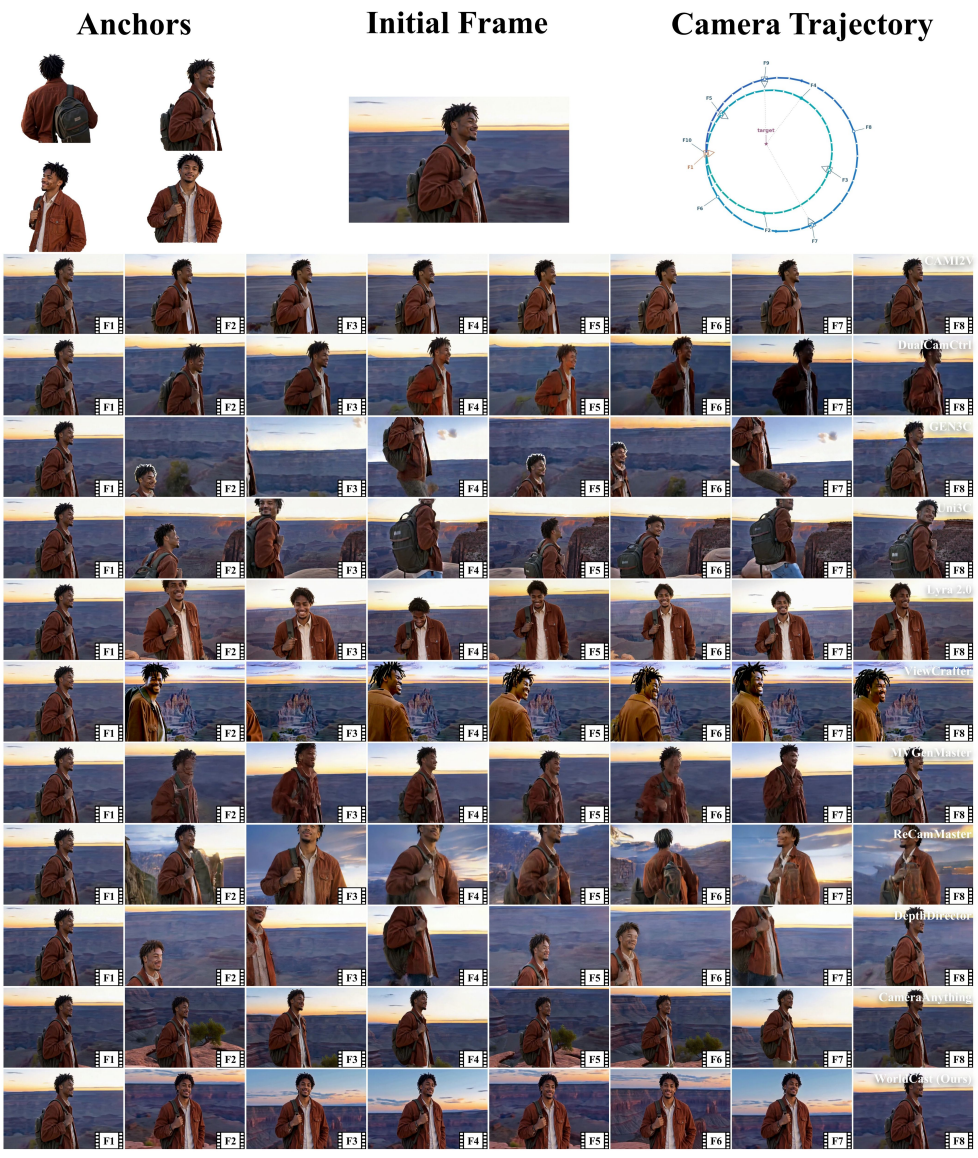


Figure D.7: Additional qualitative comparison across diverse camera motions and initial azimuths.

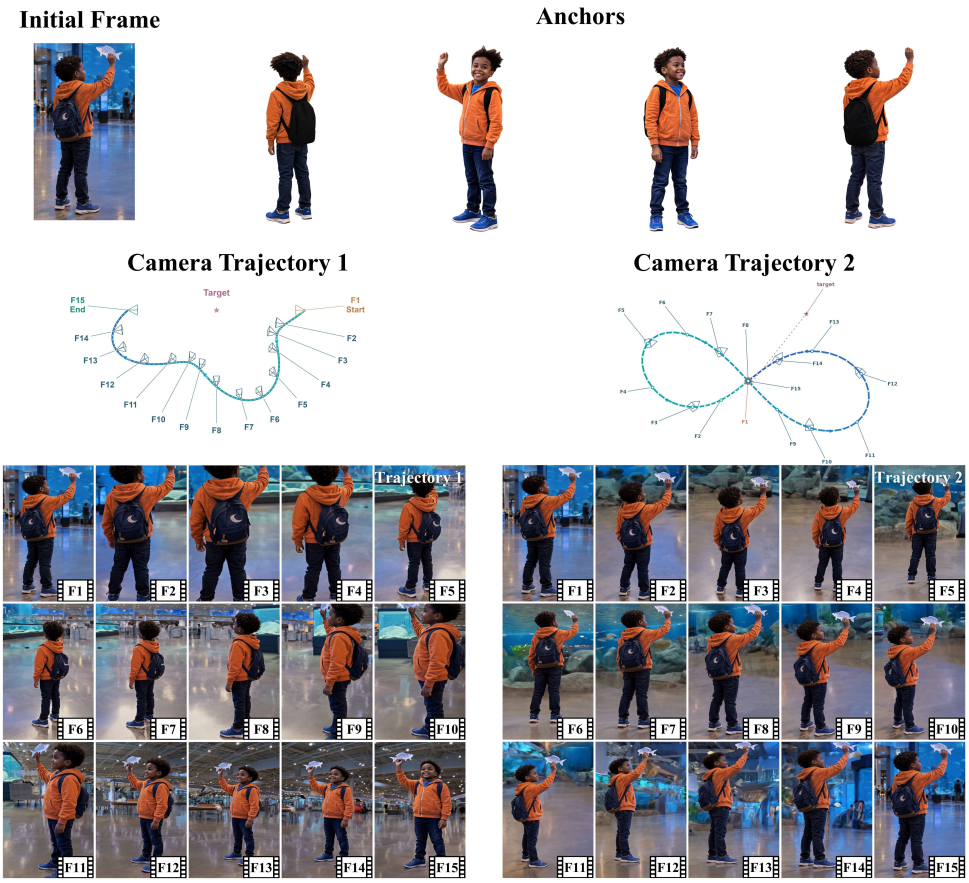


Figure D.8: Trajectory generalization using the same set of character anchors across diverse target camera motions.

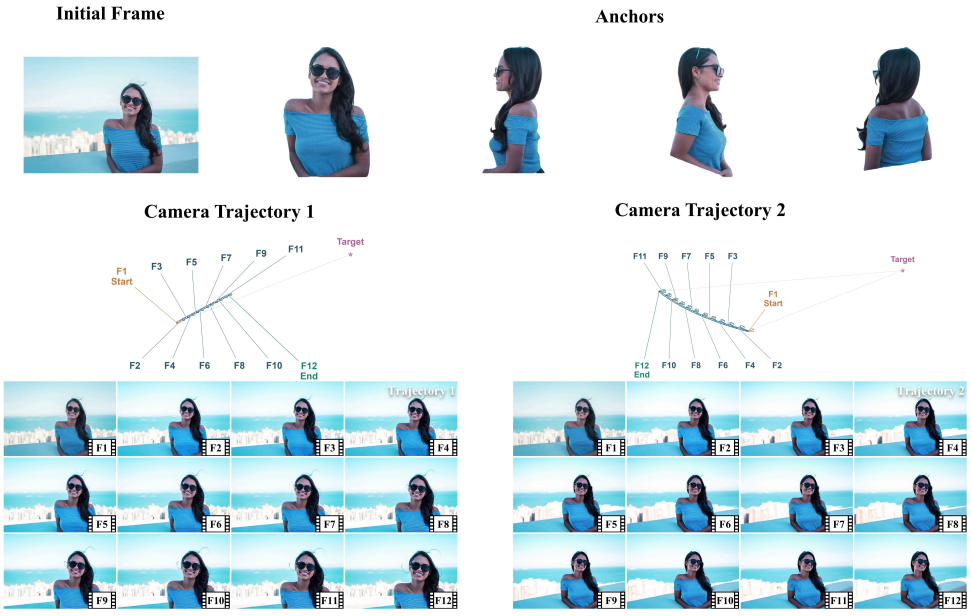


Figure D.9: Trajectory generalization using the same set of character anchors across diverse target camera motions.