ONE-SHOT: Compositional Human-Environment Video Synthesis via Spatial-Decoupled Motion Injection and Hybrid Context Integration

Fengyuan Yang[1,2] **Luying Huang**[2,†] **Jiazhi Guan**[2,*] **Quanwei Yang**[2] **Dongwei Pan**[2]
Jianglin Fu[2] **Haocheng Feng**[2] **Wei He**[2] **Kaisiyuan Wang**[2] **Hang Zhou**[2,*] **Angela Yao**[1]
[1]National University of Singapore [2]Baidu AMU
Project page: `https://martayang.github.io/ONE-SHOT/`

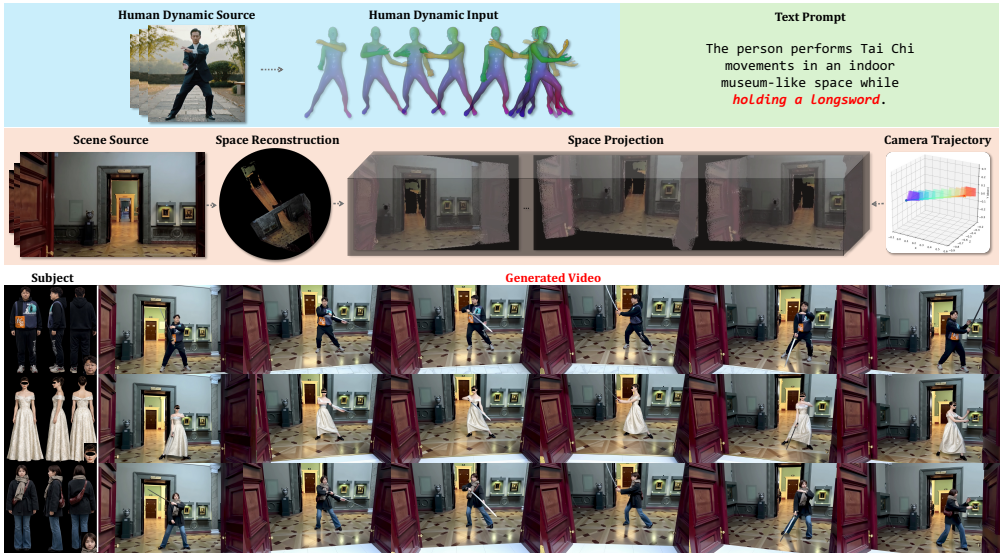


Figure 1: **ONE-SHOT is a novel compositional human-environment generation framework** that decouples human dynamics, text prompts and environment space. The generated videos faithfully follow the specified motion, appearance, and environment controls, while maintaining high consistency and incorporating text-specified semantic details such as "holding a longsword".

Abstract

Recent advances in Video Foundation Models (VFMs) have revolutionized human-centric video synthesis, yet fine-grained and independent control of subjects and scenes remains a critical challenge. Recent attempts to achieve joint human-environment control often rely on explicit human-scene 3D alignment, improving geometric control but sacrificing generative flexibility and long-horizon consistency while requiring heavy 3D pre-processing. We present ONE-SHOT, a parameter-efficient framework for compositional human-environment video generation. Our key insight is to factorize the generative process into disentangled signals, separating human dynamics from scene cues. Specifically, our canonical-space motion injection and Dynamic-Grounded-RoPE map canonical human motion to the target video region, enabling accurate motion and placement control without dense human-scene 3D alignment. Hybrid Context Integration further maintains subject and scene consistency for minute-level synthesis. Experiments show that ONE-SHOT consistently outperforms state-of-the-art methods in structural control, flexible composition, text-guided semantic control, and long-horizon consistency.

† Project lead. * Corresponding authors.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

1 Introduction

Imagine a video generating system that places a chosen human subject into a target 3D scene, drives the subject with a specified motion, and synthesizes a coherent video along a specified camera trajectory as if it were filmed in *one continuous shot*. As shown in Fig. 1, such compositional human-environment generation requires independent control over subject appearance, human motion, scene environment, camera trajectory, and text-specified semantic details. Recent advances in Video Foundation Models (VFMs), such as Wan [1] and HunyuanVideo [2], make this vision increasingly possible, by producing high-quality videos and supporting text-to-video generation, video editing, and human animation [3–6]. However, the setting above demands more than visual realism: the generated video should follow a specified body motion, preserve the reference identity, align with the 3D scene and camera trajectory, and retain text-level semantic controllability.

Subject- and video-editing methods [4, 7, 8, 5, 6] have expanded VFM-based generation toward subject-consistent generation, character animation, and video editing. These methods can edit or animate the given visual content, but their controls are usually tied to 2D observations, reference appearances, masks, or driving signals, where the human subject and the surrounding scene are already coupled in pixel space. As a result, they are not naturally designed for factorized combinations of human motion, scene structure, and camera trajectory. For compositional synthesis, we instead need factorized controls that specify human dynamics independently from scene structure, enabling more flexible composition of motions with different environments and camera trajectories.

Spatially grounded scene cues provide a natural way to support such factorization. Recent 3D-aware methods [9, 10] show that scene geometry (e.g., point clouds or RGB-D projections) together with camera trajectories can improve human placement, viewpoint control, and human-scene consistency. However, current 3D-aware systems still face several practical challenges: they often require heavy human-scene 3D alignment, dense geometric conditioning can over-constrain the generative and text-controllable capacity of pretrained VFMs, and most systems remain restricted to short-duration synthesis ($\sim$10s), making it difficult to maintain spatio-temporal consistency over long horizons.

To address these challenges, we propose a c**O**mpositional huma**N**-**E**nvironment video synthesis framework based on **S**patial-decoupled motion injection and **H**ybrid c**O**ntext in**T**egration, namely **ONE-SHOT**. Instead of constructing a fully aligned human-scene control signal before generation, ONE-SHOT deliberately keeps the scene and motion controls in decoupled domains. The environment control is rendered from the scene point cloud under the target camera trajectory, while human motion is kept in a canonical body-centered space and injected through a dedicated cross-attention pathway during generation. During motion injection, we only need a coarse human layout indicating where the subject should appear in the target video, rather than requiring complex human-scene alignment as 3D preprocessing [9, 10]. This design enables flexible composition of motion, scene, and camera trajectory, while avoiding overly dense controls that can over-constrain generation and weaken the pretrained VFM's expressiveness and text-level semantic controllability, as illustrated in Fig. 1.

Keeping motion in a scene-independent canonical space makes it reusable across different environments and camera trajectories, but it also removes the direct spatial correspondence between target video tokens and canonical motion tokens. We address this coordinate-grounding challenge with ***Dynamic-Grounded-RoPE***, a layout-conditioned positional grounding strategy for motion cross-attention. Given the coarse human layout, Dynamic-Grounded-RoPE normalizes and rescales the RoPE coordinates of target-region queries to match the coordinate scale of the canonical motion token map, while assigning background queries a shared background label. In this way, motion injection can associate the target human region with canonical pose tokens in attention space, without first explicitly aligning human motion with scene geometry in 3D [9, 10].

Beyond spatial grounding, compositional human-environment videos should also remain consistent over long horizons, where subject identity and scene structure must be preserved as the camera moves or revisits previous viewpoints. To support minute-level synthesis, we introduce ***Hybrid Context Integration***, which combines static context reference tokens to preserve subject identity and dynamic context memory to maintain the scene appearance. Together, they enable long-horizon human-environment generation with reduced identity drift and improved viewpoint-revisit consistency.

Our contributions are summarized as follows: **1)** We introduce ONE-SHOT, a parameter-efficient compositional human-environment video synthesis framework that factorizes subject appearance, human dynamics, spatial environment, camera trajectory, and text semantics for flexible composition

and high-fidelity synthesis. **2)** We propose canonical-space motion injection with Dynamic-Grounded-RoPE, which grounds scene-independent human motion to the target video region without complex 3D pre-alignment between human motion and scene geometry. This design enables accurate motion and placement control while preserving text-level semantic controllability. **3)** Hybrid Context Integration combines static subject reference and dynamic scene context, maintaining subject identity and scene consistency across minute-scale generations. **4)** Extensive experiments and ablations show that ONE-SHOT outperforms existing methods in structural control, flexible composition, text-level semantic controllability, and long-horizon consistency.

2 Related Work

2.1 Conditional Video Generation and Editing

Large-scale video generation models [2, 1, 11–13] provide a strong foundation for conditional video synthesis. Building on these models, general-purpose conditional generation and editing methods support a wide range of controls, including image/reference and subject conditioning [14–17], structural or layout guidance [18–20], and text-guided editing [21, 22]. For videos, recent unified frameworks extend these controls to multi-task video creation and editing [4, 5], while another line improves temporally consistent editing through feature propagation, keyframe propagation, or attention manipulation [23–25]. Other works explore subject replacement, layered editing, dynamic content insertion, and motion/trajectory control [26, 17, 27–30]. These methods greatly improve controllability for generic objects and scenes. However, compositional human-environment synthesis requires more specialized control: the model must follow articulated human motion, preserve identity, respect scene and camera geometry, and remain consistent over long horizons. This motivates human-centric and 3D-aware designs rather than treating the person as a generic pixel-level object.

2.2 Human-Centric Reenactment

To achieve more granular control over human forms, human-centric reenactment [6, 31–34, 3, 35–39] leverages structured motion priors to drive realistic character animations. Early efforts like Champ [34] incorporated explicit SMPL-based 3D shape priors to improve robustness against large viewpoint changes, while Animate Anyone 2 [39] set a new benchmark for identity preservation through large-scale video pre-training. More recently, building upon the Wan foundational model [1], Wan-Animate [6] employs spatially-aligned skeleton signals and implicit facial features within a multi-stage training pipeline to achieve superior controllability and expressiveness. Simultaneously, Realisdance-DiT [31] emphasizes motion reliability by introducing targeted modifications to the DiT architecture and employing flexible fine-tuning strategies, which significantly enhances stability in rare poses and complex character-object interactions. While these methods achieve remarkable 2D fidelity, they primarily animate characters in isolated or static settings, failing to perceive or respond to the physical constraints of complex 3D environments. Complementary to these generation methods, world-coordinate human reconstruction approaches recover 3D human motion and global trajectories from monocular videos [40–45]. These reconstructions can serve as structured 3D motion priors, while these methods primarily focus on reconstruction rather than compositional generation in novel environments.

2.3 Human-Environment Video Synthesis

Moving beyond character animation in static settings, human-environmental video synthesis [9, 10] addresses the challenge of grounding movements within explicit 3D scenes to model complex interactions. Uni3C [9] achieves unified 3D coordination by explicitly aligning human motion with scene geometry (e.g., point clouds) to resolve artifacts like interpenetration. RealisMotion [10] further introduces physical priors within a decoupled world-space control framework, significantly mitigating foot-skating and enhancing the dynamic realism of interactions. To further unlock generative freedom, works like MotionCtrl [46] and 3DTrajMaster [47] incorporate controllable camera trajectories and 3D viewpoints, enabling the synthesis of complex interactions from dynamic perspectives within a persistent 3D space. Despite these advancements, current frameworks remain hindered by burdensome 3D pre-processing, a loss of generative diversity due to rigid conditioning, and a failure to maintain spatio-temporal consistency over long-horizon sequences. To address these

challenges, our proposed ONE-SHOT disentangles human dynamics from spatial environments, which simultaneously preserves generative diversity and persistent subject consistency.

3 Preliminaries

Video Foundation Models. We build our model upon a pretrained video foundation model (VFM) trained with flow matching in a latent space, and instantiate it with Wan2.1 [1] in this paper. Given a clean training video $\mathbf{x}_0$, a VAE encoder maps it into a latent tensor $\mathbf{z}_0 \in \mathbb{R}^{T\times H\times W\times C}$, where T is the number of latent frames, $H \times W$ denotes the spatial token grid, and C is the channel dimension. After generation, the predicted latent is decoded back to video frames by a VAE decoder. The VFM learns a time-dependent velocity field along a path between the data latent and a noise latent. Concretely, a noise latent $\mathbf{z}_1 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and a flow timestep $\lambda \in [0, 1]$ sampled from a logit-normal distribution are used to construct an intermediate latent $\mathbf{z}_\lambda = \lambda\, \mathbf{z}_1 + (1 - \lambda)\, \mathbf{z}_0$. Following Rectified Flows [48, 49], the ground-truth velocity along this path is $\mathbf{v}_\lambda = \frac{d\mathbf{z}_\lambda}{d\lambda} = \mathbf{z}_1 - \mathbf{z}_0$. A conditional velocity predictor $\mathbf{v}_\theta$ is trained by minimizing the mean squared error

$$\mathcal{L}_{\mathrm{FM}} = \mathbb{E}_{\mathbf{z}_0,\mathbf{z}_1,\lambda}\left[\left\|\mathbf{v}_\theta(\mathbf{z}_\lambda, \lambda, \mathbf{c}) - \mathbf{v}_\lambda\right\|_2^2\right], \tag{1}$$

where $\mathbf{c}$ denotes conditions, such as text prompts and the additional controls introduced in Sec. 4.

Rotary Position Embedding (RoPE). RoPE [50] injects positional information into attention by applying a position-dependent rotation to the query and key features before computing attention logits. For a token at position $\mathbf{p}$ (e.g., $\mathbf{p} = (t, x, y)$ on the latent grid), let $\mathcal{R}(\mathbf{p}) = \mathcal{R}(t, x, y)$ denote the RoPE rotation operator constructed from sinusoidal functions at multiple frequencies. Given the query and key features $\mathbf{Q}$ and $\mathbf{K}$ of this token, RoPE applies

$$\tilde{\mathbf{Q}} = \mathcal{R}(t, x, y)\, \mathbf{Q}, \qquad \tilde{\mathbf{K}} = \mathcal{R}(t, x, y)\, \mathbf{K}. \tag{2}$$

The rotated $\tilde{\mathbf{Q}}, \tilde{\mathbf{K}}$ are then used in attention, making attention scores sensitive to relative positional offsets. In Sec. 4, we introduce a layout-grounded variant of RoPE for our decoupled conditioning.

4 Methodology

Task Definition. We aim to generate a target video of length T, indexed by t. The input consists of a target scene point cloud $\mathcal{P}$ (which can be reconstructed from a source video or multi-view images and is used only for coarse RGB-D scene cues), a camera trajectory given by extrinsics $\{\mathbf{T}_t\}_{t=1}^T$, a subject appearance reference $\mathbf{y}^{\text{app}}$, a human motion sequence $\mathbf{m}_{1:T}$, and a text prompt $\mathbf{s}_{\text{text}}$. Our goal is to synthesize a realistic human-environment video $\mathbf{x}_{1:T}$ where the reference subject follows the specified motion, remains coherent with the target scene under the camera trajectory, and reflects text-specified semantic details:

$$\mathbf{x}_{1:T} \;=\; \mathcal{G}\big(\mathcal{P}, \{\mathbf{T}_t\}_{t=1}^T, \mathbf{y}^{\text{app}}, \mathbf{m}_{1:T}, \mathbf{s}_{\text{text}}; \mathbf{y}^{\text{mem}}\big)\,, \tag{3}$$

where $\mathbf{y}^{\text{mem}}$ denotes an optional scene-context memory used for long-horizon synthesis, which stores previously generated context frames to maintain scene consistency across clips. In practice, generation is performed in the latent space of the VFM, producing $\mathbf{z}_0 \in \mathbb{R}^{T\times H\times W\times C}$ that is decoded back to $\mathbf{x}_{1:T}$.

Decoupled Conditioning Framework. We encode the above inputs into six complementary conditions:

$$\mathbf{c} \;=\; \{\mathbf{c}_{\text{text}}, \mathbf{c}_{\text{env}}, \mathbf{c}_{\text{mot}}, \mathbf{c}_{\text{id}}, \mathbf{c}_{\text{mem}}, \mathbf{c}_{\text{mask}}\}. \tag{4}$$

Specifically, (a) *text conditions* $\mathbf{c}_{\text{text}}$ encode the text prompt $\mathbf{s}_{\text{text}}$ for semantic control; (b) *environmental conditions* $\mathbf{c}_{\text{env}}$ consist of per-frame RGB-D projections of the 3D scene $\mathcal{P}$ under the target camera trajectory $\{\mathbf{T}_t\}_{t=1}^T$; (c) *motion conditions* $\mathbf{c}_{\text{mot}}$ specify human dynamics $\mathbf{m}_{1:T}$ via canonical-space tokens and a coarse layout bounding box $\{\mathcal{B}_t\}_{t=1}^T$; (d) *appearance conditions* $\mathbf{c}_{\text{id}}$ encode the subject appearance reference $\mathbf{y}^{\text{app}}$; (e) *context memory* $\mathbf{c}_{\text{mem}}$ stores scene-context cues $\mathbf{y}^{\text{mem}}$ for long-horizon consistency; and (f) a spatio-temporal *binary mask* $\mathbf{c}_{\text{mask}}$ guides localized synthesis. As shown in Fig. 2, we inject the dense spatial and visual conditions into the frozen VFM backbone by training a lightweight ControlNet-style [18] conditioning branch via Low-Rank Adaptation (LoRA) [51]. Concretely, $\mathbf{c}_{\text{env}}$, $\mathbf{c}_{\text{id}}$, $\mathbf{c}_{\text{mem}}$, and $\mathbf{c}_{\text{mask}}$ are encoded as dense conditioning inputs. In contrast, $\mathbf{c}_{\text{mot}}$ is

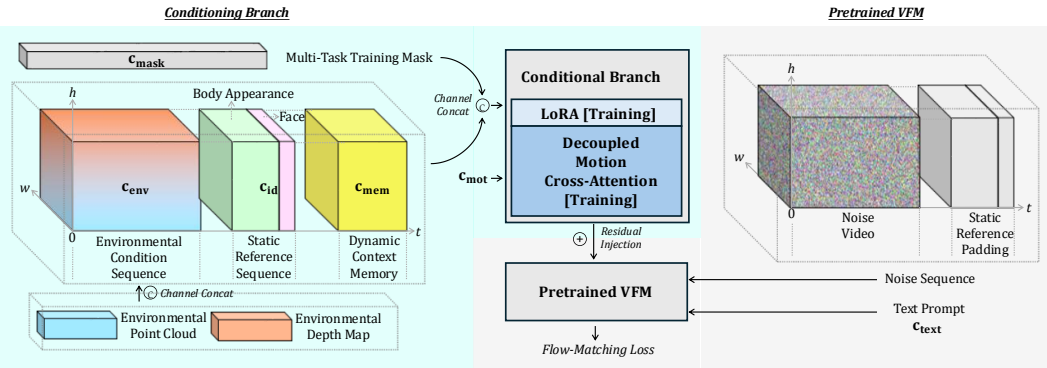


Figure 2: **Model Architecture.** Our model is built upon a *Pretrained VFM* (gray, right) augmented with an additional *Conditioning Branch* (cyan, left) for control injection. The environmental condition $\mathbf{c}_{\text{env}}$ is encoded from 2D-projected point clouds and depth maps, while identity appearance $\mathbf{c}_{\text{id}}$ and context memory $\mathbf{c}_{\text{mem}}$ jointly maintain visual coherence. Human dynamics $\mathbf{c}_{\text{mot}}$ are disentangled from environmental inputs and injected through the proposed Decoupled Motion Cross-Attention.

injected through a dedicated cross-attention pathway, thereby keeping human dynamics decoupled from dense environment conditioning. All trainable conditioning modules are jointly optimized with the multi-task training strategy described in Sec. 4.3.

4.1 Conditioning Branch

The left part of Fig. 2 depicts a ControlNet-style side branch [18] that conditions the frozen VFM through hierarchical residual additions. We organize dense control cues into a unified conditioning context sequence and encode them into layer-wise residual features that are added to selected VFM layers. This branch handles scene, appearance, memory, and mask via concatenation, while human motion is injected within the same branch through the decoupled cross-attention pathway in Sec. 4.2.

Environmental Condition Sequence. We represent the scene and camera trajectory using 2D projections rendered from the point cloud $\mathcal{P}$ under the target extrinsics $\{\mathbf{T}_t\}_{t=1}^{T}$. Let $\pi(\cdot;\mathbf{T}_t)$ denote a rendering/projection operator. We define the environmental condition as an RGB–D sequence

$$\mathbf{c}_{\text{env}} = \left\{\mathbf{RGBD}_t^{\text{env}} \;=\; \pi(\mathcal{P};\mathbf{T}_t)\right\}_{t=1}^{T}, \tag{5}$$

where $\mathbf{RGBD}_t^{\text{env}}$ provides coarse scene geometry and appearance cues from the target camera viewpoint at time t.

Hybrid Context Integration. We construct (i) an identity condition $\mathbf{c}_{\text{id}}$ from a static reference segment $\mathbf{y}_{1:T_{\text{id}}}^{\text{id}}$ (body and face), and (ii) an optional context memory $\mathbf{c}_{\text{mem}}$ from a dynamically updated segment $\mathbf{y}_{1:T_{\text{mem}}}^{\text{mem}}$, enabled only for long-horizon generation:

$$\mathbf{c}_{\text{id}} = E\big(\mathbf{y}_{1:T_{\text{id}}}^{\text{id}}\big)\,, \qquad \mathbf{c}_{\text{mem}} = E\big(\mathbf{y}_{1:T_{\text{mem}}}^{\text{mem}}\big)\,, \tag{6}$$

where $E(\cdot)$ encodes input frames into conditioning tokens. Notably, $\mathbf{y}_{1:T_{\text{id}}}^{\text{id}}$ provides appearance-only cues for decoupled subject conditioning. For long sequences, we follow [52] to retrieve $\mathbf{y}_{1:T_{\text{mem}}}^{\text{mem}}$ from previously generated content under similar viewpoints, stabilizing long-range scene consistency.

Conditioning Context Sequence. Following standard conditioning designs [4], we concatenate environment and appearance segments along the temporal axis and append a training mask sequence $\mathbf{c}_{\text{mask}}$ (introduced in Sec. 4.3), then concatenate them channel-wise to form the initial context:

$$\mathbf{c}_{\text{ctx}} = \text{Concat}_{\text{ch}}\Big(\text{Concat}_t\big[\mathbf{c}_{\text{env}},\ \mathbf{c}_{\text{id}}, \mathbf{c}_{\text{mem}}\big],\ \mathbf{c}_{\text{mask}}\Big). \tag{7}$$

In the next subsection, we introduce a decoupled motion cross-attention module to inject motion conditions into $\mathbf{c}_{\text{ctx}}$, enabling compositional control over both pose dynamics and the global spatial placement of the human subject.

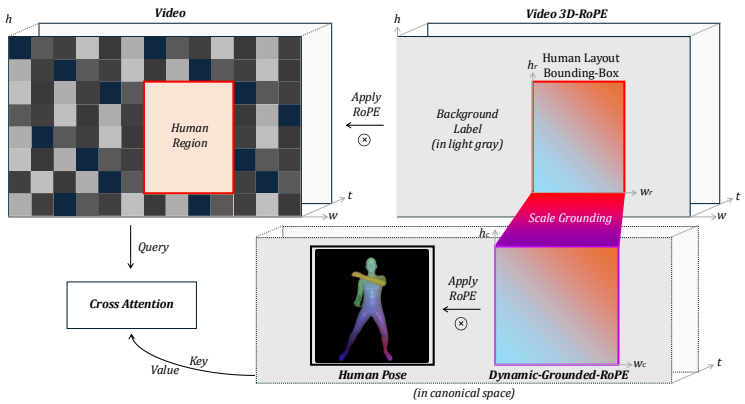


Figure 3: **Decoupled Motion Cross-attention.** The canonical-space human pose is injected into the video through cross-attention, where the proposed Dynamic-Grounded-RoPE bridges the spatial discrepancy between the environment and human spaces.

4.2 Decoupled Motion Cross-Attention

To control both human motion and spatial placement, we decouple human control into (i) local motion in canonical space and (ii) global placement specified by a coarse layout signal on the video token grid (e.g., bounding boxes). Fig. 3 shows the proposed cross-attention pathway for motion injection.

Canonical-Space Human Motion Representation. We parameterize human motion using SMPL-X [53] sequence $\mathbf{m}_{1:T}$. To remove global displacement, we set the global translation to zero and center the body at a canonical origin. We render the canonical SMPL-X mesh into a pose-aligned token map $\mathbf{u}_t \in \mathbb{R}^{h_c \times w_c \times C_u}$, where the rendering camera is chosen such that the body occupies a fixed proportion of the image (e.g., 90%) to normalize scale across videos. Stacking across time yields the motion tokens $\mathbf{c}_{\text{mot}} = \{\mathbf{u}_t\}_{t=1}^{T}$. This representation preserves local articulated motion while removing scene-specific translation and scale, enabling the motion to be composed with different environments and camera trajectories.

Dynamic-Grounded-RoPE. While the motion tokens $\mathbf{c}_{\text{mot}} = \{\mathbf{u}_t\}_{t=1}^{T}$ are formulated in a canonical space, the query features are defined on the video token grid, resulting in a coordinate and scale mismatch between the two domains. As shown in the red box of Fig. 3, we introduce a per-frame placement region $\mathcal{B}_t$ on the video token grid to indicate where the human should appear. This region serves as a coarse placement cue rather than a pixel-accurate human mask (see Sec. C of the Supplementary Material for details). Its parameterization is given by $\boldsymbol{\beta}_t = (x_t^1, y_t^1, x_t^2, y_t^2)$, which defines the rectangular region $\mathcal{B}_t = \{(x, y) \mid x_t^1 \le x < x_t^2,\ y_t^1 \le y < y_t^2\}$. Let $h_r = y_t^2 - y_t^1$ and $w_r = x_t^2 - x_t^1$ denote the height/width of $\mathcal{B}_t$ in token units, and recall (h_c, w_c) as the spatial size of the canonical motion map. We define per-frame scale factors that map video-grid coordinates to the canonical coordinates:

$$s_t^h = \frac{h_c}{h_r}, \qquad s_t^w = \frac{w_c}{w_r}. \tag{8}$$

These scale factors map positions inside $\mathcal{B}_t$ into the coordinate range of the canonical motion map, making target-region queries positionally comparable to canonical motion keys during cross-attention. For a query token at position $\mathbf{p} = (t, x, y)$ on the video grid, we apply a layout-grounded RoPE:

$$\tilde{\mathbf{Q}}(\mathbf{p}) = \begin{cases} \mathcal{R}\big(t,\ s_t^w(x - x_t^1),\ s_t^h(y - y_t^1)\big)\,\mathbf{Q}(\mathbf{p}), & (x, y) \in \mathcal{B}_t, \\ \mathcal{R}\big(t,\ \alpha,\ \alpha\big)\,\mathbf{Q}(\mathbf{p}), & (x, y) \notin \mathcal{B}_t, \end{cases} \tag{9}$$

where the constant α is a large background label (Fig. 3) used to collapse all tokens outside the target region into a shared background positional embedding, distinct from the human-region coordinates. This soft grounding maintains access to motion cues under imprecise boxes (Sec. C of the Supplementary Material).

For motion keys defined on the canonical grid, we apply the standard 3D RoPE without rescaling:

$$\tilde{\mathbf{K}}(\mathbf{p}_c) = \mathcal{R}\big(t,\ x_c,\ y_c\big)\,\mathbf{K}(\mathbf{p}_c), \tag{10}$$

where $\mathbf{p}_c = (t, x_c, y_c)$ indexes positions on the canonical motion map. Together, Eqs. (9)–(10) place target-region queries and canonical motion keys in a comparable positional coordinate system, enabling motion grounding in attention space across the two decoupled domains.

Motion Injection by Cross-Attention. We inject motion via a dedicated cross-attention layer, where queries are taken from the video tokens and keys/values come from the canonical motion tokens. Concretely, at frame t, we extract query features $\mathbf{Q}_t$ from the environmental features conditioned on $\mathbf{c}_{\text{env}}$. We form keys and values from the canonical motion map $\mathbf{u}_t$ as $\mathbf{K}_t, \mathbf{V}_t$. With $\tilde{\mathbf{Q}}_t$ and $\tilde{\mathbf{K}}_t$ obtained from Eqs. (9)–(10), motion injection is computed as

$$\text{Attn}(\tilde{\mathbf{Q}}_t, \tilde{\mathbf{K}}_t, \mathbf{V}_t) = \text{softmax}\left(\frac{\tilde{\mathbf{Q}}_t \tilde{\mathbf{K}}_t^\top}{\sqrt{d}}\right)\mathbf{V}_t. \tag{11}$$

4.3 Model Training

Multi-Task Training. A key advantage of our decoupled design is that environmental and motion supervision can be learned from heterogeneous sources. We therefore adopt a mask-guided mixed-conditioning strategy to train different components with partially available conditions. Specifically, we use a binary mask sequence $\mathbf{c}_{\text{mask}}$ to indicate which region should be synthesized versus preserved, and construct the masked target in latent space as

$$\mathbf{z}_0^{\text{tgt}} = \mathbf{c}_{\text{mask}} \odot \mathbf{z}_0 + (1 - \mathbf{c}_{\text{mask}}) \odot \mathbf{z}_0^{\text{keep}}, \tag{12}$$

where $\mathbf{z}_0$ is the ground-truth video latent, $\mathbf{z}_0^{\text{keep}}$ denotes preserved latents, and $\odot$ is element-wise multiplication with broadcasting.

We use three training modes: (i) *scene-only*, which focuses on reconstructing the environment from $\mathbf{c}_{\text{env}}$ (and optionally the dynamic context memory in $\mathbf{c}_{\text{mem}}$), while disabling motion injection; (ii) *motion-only*, where we train the motion cross-attention to synthesize the human inside the target region $\mathcal{B}_t$ using $\mathbf{c}_{\text{mot}}$, while preserving the outside region with $(1 - \mathbf{c}_{\text{mask}})$; and (iii) *full conditioning*, where $\mathbf{c}_{\text{env}}, \mathbf{c}_{\text{mot}}, \mathbf{c}_{\text{id}}, \mathbf{c}_{\text{mem}}$ are jointly enabled. This schedule encourages specialization of decoupled modules while maintaining compositional generalization at test time. Empirically, we find that competitive controllability can be achieved with substantially less paired task data than fully supervised fine-tuning. More training details are provided in Sec. A of the Supplementary Material.

Training objective. We follow the flow-matching objective (Eq. (1)) and replace the data latent $\mathbf{z}_0$ with the masked target $\mathbf{z}_0^{\text{tgt}}$ defined in Eq. (12).

5 Experiments

5.1 Experimental Setup

Datasets. As mentioned earlier, our method supports training with heterogeneous data sources. We use a mixture of datasets that provide complementary supervision: (i) **EMDB2** [54], which contains sequences captured with dynamic cameras and provides paired global human/camera motion annotations for supervising human motion under moving cameras; (ii) a human-motion subset of **MotionX** [55], which mainly consists of static-camera videos with diverse body motions; (iii) **ARKitScenes** [56], which provides point-cloud captures used to strengthen environmental reconstruction and camera-aware projections; and (iv) **licensed commercial stock-media footage** to increase training-data diversity. In total, our training set contains approximately 50k clips, each 5–20 seconds long. For testing, we use **Traj100** following the same protocol as RealisMotion [10], as well as a curated cross-composition test set for evaluating swaps of identity, motion, and/or scene. More dataset details are provided in Sec. B of the Supplementary Material.

Comparison methods. We compare against representative controllable human-video generation baselines: **RealisDance**, **Uni3C**, **WanAnimate**, and **RealisMotion**. For each baseline, we follow the official inference settings and use the same prompts and evaluation protocol whenever applicable.

Inference protocol. For compositional inference, the coarse first-frame bbox can be specified via an interactive interface and then propagated as described in Sec. C of the Supplementary Material.

Table 1: **Quantitative Results on Traj100 (Self-Reconstruction) and Our Constructed Testset (Cross-Composition).** Lower is better for FID/FVD. Higher is better for others.

Method	Self-reconstruction (Traj100)					Cross-composition (Identity/Motion Swap)					
	FID↓	FVD↓	MS↑	BC↑	SubSim↑	FID↓	FVD↓	MS↑	BC↑	SubSim↑	Face↑
WanAnimate [6]	21.25	224.08	0.893	0.956	0.189	152.07	1435.3	0.892	0.950	**0.368**	0.196
RealisDance [31]	22.17	247.54	0.920	0.946	0.237	155.71	1416.5	0.986	0.962	0.230	**0.257**
Uni3C [9]	55.47	775.38	0.853	0.928	0.185	162.53	1519.5	0.924	0.950	0.321	0.203
RealisMotion [10]	18.89	193.89	0.916	0.940	**0.356**	151.19	1486.8	0.960	0.919	0.300	0.238
ONE-SHOT (**ours**)	**16.88**	**181.17**	**0.940**	**0.959**	0.317	**150.98**	**1382.4**	**0.993**	**0.963**	0.349	0.241

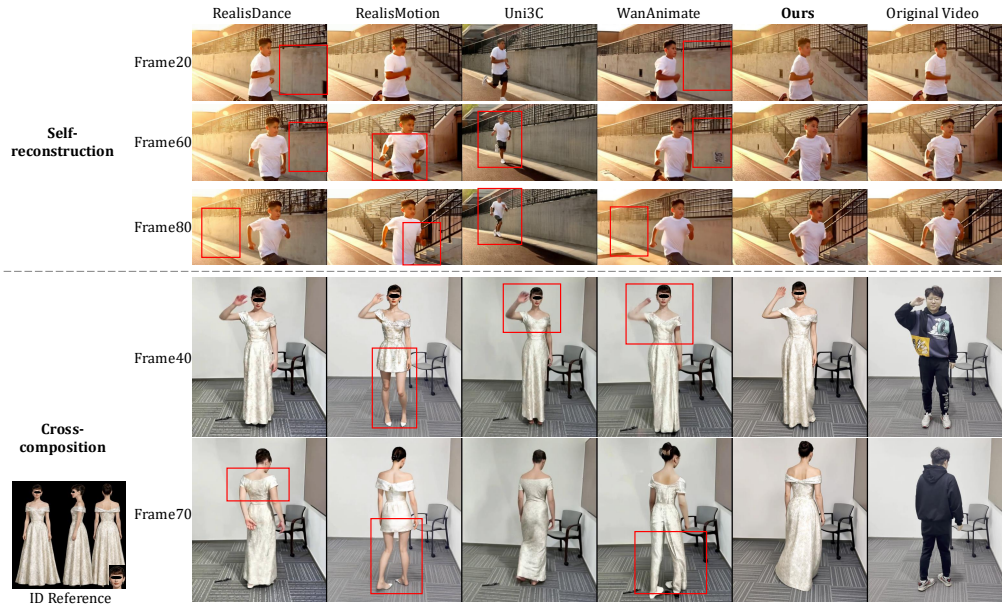


Figure 4: **Self-Reconstruction and Cross-Composition Comparisons.** (a) Our method better preserves both the scene structure and human appearance, and (b) maintains stable subject placement and motion consistency when swapping identity.

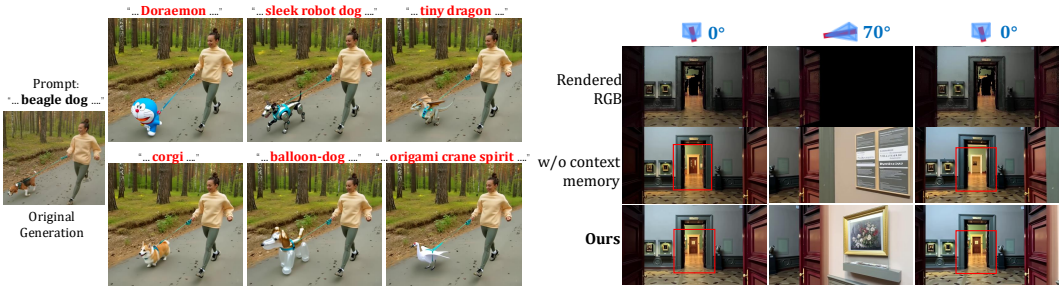
Evaluation metrics. We evaluate both **self-reconstruction** (with paired ground truth) and **cross-composition** (without paired ground truth) using **FID** and **FVD** to measure visual quality and temporal realism, together with task-specific metrics including **Motion Smoothness** (MS), **Background Consistency** (BC), and **Subject Similarity** (SubSim). For cross-composition, we additionally report **Face Similarity** (FaceSim) to assess identity swapping. All metrics are computed using the same number of frames and the same resolution across methods.

5.2 Evaluation Results

Quantitative results. Tab. 1 (left) reports self-reconstruction results on Traj100. Our method achieves the best **FID** and **FVD**, indicating improved visual fidelity and temporal realism. It also attains the highest **Motion Smoothness** and **Background Consistency**, suggesting that our decoupled design stabilizes both human motion synthesis and scene coherence. While RealisMotion obtains a higher SubjectSim, our approach provides a more balanced trade-off across fidelity, motion stability, and background coherence, which is crucial for controllable human-environment generation.

Tab. 1 (right) evaluates cross-composition, where identity and motion are swapped across different scenes. Our method achieves the best performance on **FID/FVD**, **Motion Smoothness** and **Background Consistency**, and remains competitive in **SubjectSim** and **FaceSim**.

Motion and spatial control. Additional evaluation on Traj100 shows competitive motion following, subject placement, and camera control, with the joint-lowest trajectory-failure rate and the lowest foot-ground contact error among the compared methods. Full results and metric definitions are provided in Sec. F (Tab. 6) of the Supplementary Material.



(a) Text-guided editing. (b) Effect of context memory.

Figure 5: (a) **Text-guided editing with compositional controls**. Beyond composing scene, identity, and motion, our method enables instruction-based editing via text prompts, indicating strong compatibility with pretrained VFM and minimal loss of its native text-conditioned editing ability. (b) **Effect of context memory on viewpoint revisit consistency**. Our model maintains structural and temporal consistency even after camera loops, whereas memoryless baselines suffer from appearance drift.

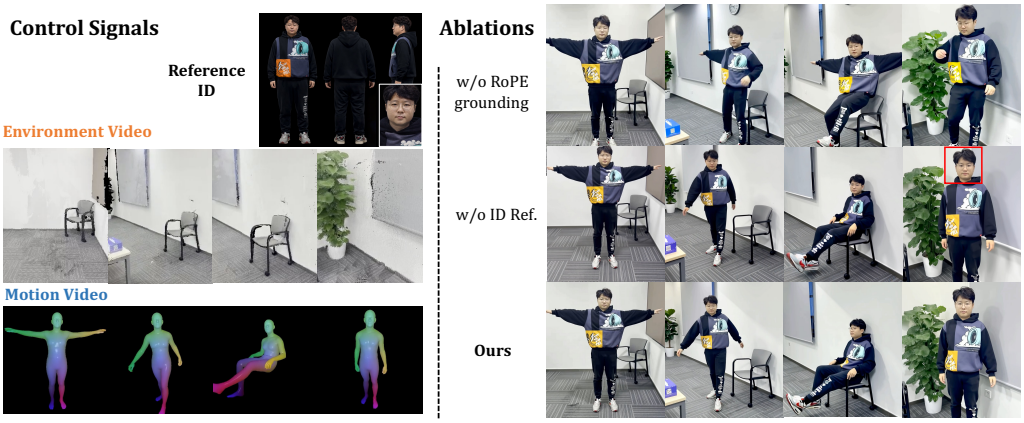


Figure 6: **Ablations of Dynamic-Grounded-RoPE and Face Appearance Integration on Long-Horizon Generation.** Removing RoPE grounding and face reference weakens motion controllability and identity anchoring, respectively, leading to increased pose and identity drift.

Qualitative results. Fig. 4 compares our method with prior approaches in both self-reconstruction and cross-composition. Our method produces more coherent backgrounds and more faithful motion dynamics. In contrast, WanAnimate and RealisDance mainly rely on the first-frame reference and do not explicitly model a persistent 3D-consistent environment, leading to background drifting, structural inconsistencies, or local collapse under camera motion. RealisMotion often yields blurred motion and unrealistic swapped subjects with severe artifacts. Uni3C struggles to disentangle camera motion from human motion, resulting in weak motion following where the subject appears nearly static or moves insufficiently. In cross settings, common failure modes include identity drift, background inconsistency, and motion mis-following, whereas our decoupled motion injection maintains stable placement and consistent motion following.

We can independently swap identity, motion, environment, and camera trajectory as shown in the demo video, while also supporting instruction-based edits via text prompts as shown in Fig. 5a. We also demonstrate the long video generation capability of our method in the demo video, where the scene and human appearance remain consistent over long horizons.

5.3 Ablation Studies

Dynamic-Grounded-RoPE. In Tab. 2, we ablate Dynamic-Grounded-RoPE by (i) removing scale grounding (2nd row) and (ii) replacing it with a naive label RoPE without 2D spatial coordinate–aware scaling (3rd row). Both variants substantially degrade **FID/FVD** and reduce motion smoothness. Without RoPE grounding, the model exhibits incorrect spatial placement and weaker motion following

Table 2: **Ablations on Traj100.**

Ablation	FID↓	FVD↓	MS↑
ONE-SHOT **(ours)**	**16.88**	**181.17**	**0.9404**
w/o RoPE grounding	22.34	229.76	0.9268
label-only RoPE	21.36	221.66	0.9256
w/o face reference	21.37	215.43	0.9378
w/o subject-cropped reference	19.43	200.23	0.9351
w/o context memory	18.17	182.98	0.9330

(see Fig. 6, 1st on the right), while naive label RoPE partially alleviates the issue but remains inferior to the full design. These results suggest that 2D spatial coordinate–aware scaling is important for establishing stable correspondence between canonical motion tokens and the target region on the video grid.

Hybrid Context Integration. As shown in Tab. 2 (4th–5th rows) and Fig. 6 (2nd row on the right), removing face references or using uncropped references hurts reconstruction quality, suggesting that clean appearance cues are necessary for stable identity and texture rendering. Disabling context memory mainly affects long-horizon consistency: results become less stable and inconsistent when the camera revisits previous viewpoints like the case in Fig. 5b. Controlled 60-second viewpoint-revisit evaluations further show that the identity-consistency advantage of face references grows over time (Tab. 4), while context memory improves background consistency across all six revisits (Tab. 5). These findings support the complementary roles of static face references and dynamic context memory in mitigating long-horizon identity and scene-appearance drift (detailed protocols and results are provided in Sec. E of the Supplementary Material).

6 Limitations

ONE-SHOT inherits uncertainty from upstream human and scene reconstruction, and very long generations may still accumulate temporal drift. More fundamentally, compositional generation becomes challenging when the specified human motion is incompatible with the target scene geometry, since the coarse placement cue does not explicitly enforce fine-grained body-scene contact. Detailed failure cases and additional discussion are provided in Sec. G of the Supplementary Material.

7 Conclusion

We presented **ONE-SHOT**, a parameter-efficient framework for compositional human-environment video synthesis. By factorizing human motion, environment geometry, camera trajectories, and human appearance into disentangled signals, ONE-SHOT addresses three key bottlenecks in prior human-centric VFMs: burdensome 3D alignment, over-conditioning, and short-horizon synthesis. Our **canonical-space motion injection** decouples human motion from environmental cues, **Dynamic-Grounded-RoPE** establishes spatial correspondence without dense human-scene pre-alignment, and **Hybrid Context Integration** maintains subject and scene consistency for minute-level generation. Experiments and ablations show that ONE-SHOT achieves a favorable balance between structural control and creative diversity with only lightweight LoRA fine-tuning.

Acknowledgements and Disclosures

This work was conducted during Fengyuan Yang's internship at Baidu, with computing resources provided by Baidu. The authors have no additional competing interests to disclose.

References

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

[2] 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.

[3] Yuxuan Luo, Zhengkun Rong, Lizhen Wang, Longhao Zhang, and Tianshu Hu. Dreamactor-m1: Holistic, expressive and robust human image animation with hybrid guidance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11036–11046, 2025.

[4] 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*, pages 17191–17202, 2025.

[5] Cong Wei, Quande Liu, Zixuan Ye, Qiulin Wang, Xintao Wang, Pengfei Wan, Kun Gai, and Wenhu Chen. Univideo: Unified understanding, generation, and editing for videos. *arXiv preprint arXiv:2510.08377*, 2025.

[6] Gang Cheng, Xin Gao, Li Hu, Siqi Hu, Mingyang Huang, Chaonan Ji, Ju Li, Dechao Meng, Jinwei Qi, Penchong Qiao, et al. Wan-animate: Unified character animation and replacement with holistic replication. *arXiv preprint arXiv:2509.14055*, 2025.

[7] Lijie Liu, Tianxiang Ma, Bingchuan Li, Zhuowei Chen, Jiawei Liu, Gen Li, Siyu Zhou, Qian He, and Xinglong Wu. Phantom: Subject-consistent video generation via cross-modal alignment. *arXiv preprint arXiv:2502.11079*, 2025.

[8] Zhaoyang Li, Dongjun Qian, Kai Su, Qishuai Diao, Xiangyang Xia, Chang Liu, Wenfei Yang, Tianzhu Zhang, and Zehuan Yuan. Bindweave: Subject-consistent video generation via cross-modal integration. *arXiv preprint arXiv:2510.00438*, 2025.

[9] 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*, pages 1–12, 2025.

[10] Jingyun Liang, Jingkai Zhou, Shikai Li, Chenjie Cao, Lei Sun, Yichen Qian, Weihua Chen, and Fan Wang. Realismotion: Decomposed human motion control and video generation in the world space. *arXiv preprint arXiv:2508.08588*, 2025.

[11] Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, et al. Ltx-video: Realtime video latent diffusion. *arXiv preprint arXiv:2501.00103*, 2024.

[12] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024.

[13] Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-sora: Democratizing efficient video production for all, March 2024. URL `https://github.com/hpcaitech/Open-Sora`.

[14] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023.

[15] Teng Hu, Zhentao Yu, Zhengguang Zhou, Sen Liang, Yuan Zhou, Qin Lin, and Qinglin Lu. Hunyuancustom: A multimodal-driven architecture for customized video generation. *CoRR*, abs/2505.04512, 2025.

[16] Zhichao Zuo, Zhao Zhang, Yan Luo, Yang Zhao, Haijun Zhang, Yi Yang, and Meng Wang. Cut-and-paste: Subject-driven video editing with attention control. *Neural Networks*, 181:106818, 2025.

[17] Yuchao Gu, Yipin Zhou, Bichen Wu, Licheng Yu, Jia-Wei Liu, Rui Zhao, Jay Zhangjie Wu, David Junhao Zhang, Mike Zheng Shou, and Kevin Tang. Videoswap: Customized video subject swapping with interactive semantic point correspondence. In *CVPR*, pages 7621–7630. IEEE, 2024.

[18] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023.

[19] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 4296–4304, 2024.

[20] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22511–22521, 2023.

[21] Vladimir Kulikov, Matan Kleiner, Inbar Huberman-Spiegelglas, and Tomer Michaeli. Flowedit: Inversion-free text-based editing using pre-trained flow models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19721–19730, 2025.

[22] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022.

[23] Michal Geyer, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Tokenflow: Consistent diffusion features for consistent video editing. In *ICLR*. OpenReview.net, 2024.

[24] Shuai Yang, Yifan Zhou, Ziwei Liu, and Chen Change Loy. Rerender A video: Zero-shot text-guided video-to-video translation. In *SIGGRAPH Asia*, pages 95:1–95:11. ACM, 2023.

[25] Chenyang Qi, Xiaodong Cun, Yong Zhang, Chenyang Lei, Xintao Wang, Ying Shan, and Qifeng Chen. Fatezero: Fusing attentions for zero-shot text-based video editing. In *ICCV*, pages 15886–15896. IEEE, 2023.

[26] Yao-Chih Lee, Ji-Ze Genevieve Jang, Yi-Ting Chen, Elizabeth Qiu, and Jia-Bin Huang. Shape-aware text-driven layered video editing. In *CVPR*, pages 14317–14326. IEEE, 2023.

[27] Danah Yatim, Rafail Fridman, Omer Bar-Tal, and Tali Dekel. Dynvfx: Augmenting real videos with dynamic content. *CoRR*, abs/2502.03621, 2025.

[28] Jiazhi Guan, Quanwei Yang, Luying Huang, Junhao Liang, Borong Liang, Haocheng Feng, Wei He, Kaisiyuan Wang, Hang Zhou, and Jingdong Wang. Display: Directable human-object interaction video generation via sparse motion guidance and multi-task auxiliary. *arXiv preprint arXiv:2603.09883*, 2026.

[29] Ryan Burgert, Charles Herrmann, Forrester Cole, Michael S Ryoo, Neal Wadhwa, Andrey Voynov, and Nataniel Ruiz. Motionv2v: Editing motion in a video, 2025. URL `https://arxiv.org/abs/2511.20640`.

[30] Cheng Lei, Jiayu Zhang, Yue Ma, Xinyu Wang, Long Chen, Liang Tang, Yiqiang Yan, Fei Su, and Zhicheng Zhao. Ditraj: training-free trajectory control for video diffusion transformer. *arXiv preprint arXiv:2509.21839*, 2025.

[31] Jingkai Zhou, Yifan Wu, Shikai Li, Min Wei, Chao Fan, Weihua Chen, Wei Jiang, and Fan Wang. Realisdance-dit: Simple yet strong baseline towards controllable character animation in the wild. *CoRR*, abs/2504.14977, 2025.

[32] Rang Meng, Xingyu Zhang, Yuming Li, and Chenguang Ma. Echomimicv2: Towards striking, simplified, and semi-body human animation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5489–5498, 2025.

[33] Rang Meng, Yan Wang, Weipeng Wu, Ruobing Zheng, Yuming Li, and Chenguang Ma. Echomimicv3: 1.3 b parameters are all you need for unified multi-modal and multi-task human animation. *arXiv preprint arXiv:2507.03905*, 2025.

[34] Shenhao Zhu, Junming Leo Chen, Zuozhuo Dai, Zilong Dong, Yinghui Xu, Xun Cao, Yao Yao, Hao Zhu, and Siyu Zhu. Champ: Controllable and consistent human image animation with 3d parametric guidance. In *ECCV (55)*, volume 15113 of *Lecture Notes in Computer Science*, pages 145–162. Springer, 2024.

[35] Jiazhi Guan, Quanwei Yang, Kaisiyuan Wang, Hang Zhou, Shengyi He, Zhiliang Xu, Haocheng Feng, Errui Ding, Jingdong Wang, Hongtao Xie, et al. Talk-act: Enhance textural-awareness for 2d speaking avatar reenactment with diffusion model. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–11, 2024.

[36] Yasheng Sun, Zhiliang Xu, Hang Zhou, Jiazhi Guan, Quanwei Yang, Kaisiyuan Wang, Borong Liang, Yingying Li, Haocheng Feng, Jingdong Wang, et al. Cosh-dit: Co-speech gesture video synthesis via hybrid audio-visual diffusion transformers: Y. sun et al. *International Journal of Computer Vision*, 134(3): 85, 2026.

[37] Quanwei Yang, Jiazhi Guan, Kaisiyuan Wang, Lingyun Yu, Wenqing Chu, Hang Zhou, ZhiQiang Feng, Haocheng Feng, Errui Ding, Jingdong Wang, et al. Showmaker: Creating high-fidelity 2d human video via fine-grained diffusion modeling. *Advances in Neural Information Processing Systems*, 37:51039–51062, 2024.

[38] Yuang Zhang, Jiaxi Gu, Li-Wen Wang, Han Wang, Junqi Cheng, Yuefeng Zhu, and Fangyuan Zou. Mimicmotion: High-quality human motion video generation with confidence-aware pose guidance. In *International Conference on Machine Learning*, 2025.

[39] Li Hu, Guangyuan Wang, Zhen Shen, Xin Gao, Dechao Meng, Lian Zhuo, Peng Zhang, Bang Zhang, and Liefeng Bo. Animate anyone 2: High-fidelity character image animation with environment affordance. *arXiv preprint arXiv:2502.06145*, 2025.

[40] Vickie Ye, Georgios Pavlakos, Jitendra Malik, and Angjoo Kanazawa. Decoupling human and camera motion from videos in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21222–21232, 2023.

[41] Yufu Wang, Ziyun Wang, Lingjie Liu, and Kostas Daniilidis. TRAM: Global trajectory and motion of 3D humans from in-the-wild videos. In *European Conference on Computer Vision (ECCV)*, pages 467–487. Springer, 2024.

[42] Yue Chen, Xingyu Chen, Yuxuan Xue, Anpei Chen, Yuliang Xiu, and Gerard Pons-Moll. Human3r: Everyone everywhere all at once, 2025. URL `https://arxiv.org/abs/2510.06219`.

[43] Fengyuan Yang, Kerui Gu, Ha Linh Nguyen, Tze Ho Elden Tse, and Angela Yao. Humans as checkerboards: Calibrating camera motion scale for world-coordinate human mesh recovery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6069–6079, 2025. doi: 10.1109/ICCV51701.2025.00573.

[44] Fengyuan Yang, Tanuj Sur, Tze Ho Elden Tse, and Angela Yao. HumanBA: Human-aware bundle adjustment via global human-camera decoupling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13846–13855, June 2026.

[45] Quanwei Yang, Luying Huang, Kaisiyuan Wang, Jiazhi Guan, Shengyi He, Fengguo Li, Hang Zhou, Lingyun Yu, Yingying Li, Haocheng Feng, and Hongtao Xie. Gesturehydra: Semantic co-speech gesture synthesis via hybrid modality diffusion transformer and cascaded-synchronized retrieval-augmented generation. In *ICCV*, pages 12615–12625. IEEE, 2025.

[46] 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*, pages 1–11, 2024.

[47] Xiao Fu, Xian Liu, Xintao Wang, Sida Peng, Menghan Xia, Xiaoyu Shi, Ziyang Yuan, Pengfei Wan, Di Zhang, and Dahua Lin. 3dtrajmaster: Mastering 3d trajectory for multi-entity motion in video generation. *arXiv preprint arXiv:2412.07759*, 2024.

[48] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

[49] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis, 2024. URL `https://arxiv.org/abs/2403.03206`.

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

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

[52] Jinjing Zhao, Fangyun Wei, Zhening Liu, Hongyang Zhang, Chang Xu, and Yan Lu. Spatia: Video generation with updatable spatial memory, 2025. URL `https://arxiv.org/abs/2512.15716`.

[53] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10975–10985, 2019.

[54] Manuel Kaufmann, Jie Song, Chen Guo, Kaiyue Shen, Tianjian Jiang, Chengcheng Tang, Juan Zarate, and Otmar Hilliges. Emdb: The electromagnetic database of global 3d human pose and shape in the wild, 2023. URL `https://arxiv.org/abs/2308.16894`.

[55] Jing Lin, Ailing Zeng, Shunlin Lu, Yuanhao Cai, Ruimao Zhang, Haoqian Wang, and Lei Zhang. Motion-x: A large-scale 3d expressive whole-body human motion dataset. *Advances in Neural Information Processing Systems*, 36:25268–25280, 2023.

[56] Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Tal Dimry, Yuri Feigin, Peter Fu, Thomas Gebauer, Brandon Joffe, Daniel Kurz, Arik Schwartz, et al. Arkitscenes: A diverse real-world dataset for 3d indoor scene understanding using mobile rgb-d data. *arXiv preprint arXiv:2111.08897*, 2021.

[57] 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, 2025. URL `https://arxiv.org/abs/2511.10647`.

Supplementary Material

This supplementary material provides additional implementation details, data preprocessing descriptions, details on the 2D Bounding Box, and extended discussions to support the main paper.

A Implementation Details

Backbone and trainable modules. Our model is built upon a pretrained video editing framework [4] with a frozen VFM backbone, Wan2.1-14B [1]. We keep the backbone parameters fixed and update only lightweight trainable components: (i) standard LoRA adapters inserted into the conditioning branch, and (ii) the proposed motion cross-attention modules for controllable motion injection (Sec. 4.2 of the main paper). For LoRA, we follow a standard configuration and apply it to the query/key/value projections, output projection, and feed-forward layers (implemented as `to_q`, `to_k`, `to_v`, `to_out.0`, `ffn.net.0.proj`, and `ffn.net.2`). For motion injection, we additionally train the newly introduced motion-conditioning modules, including the extra motion cross-attention layer, the corresponding normalization layers, and the motion token embedding layer. Concretely, after the standard self-attention and text/image cross-attention, we append an additional cross-attention layer that takes motion tokens derived from SMPL-X as conditioning, together with our Dynamic-Grounded-RoPE for spatial grounding.

Optimization and training setup. All main experiments are trained on 8 NVIDIA A100 GPUs for 16k iterations at 640×640 resolution. We use different learning rates for the two trainable parts: 1×10^{-4} for LoRA parameters and 1×10^{-5} for the newly added motion cross-attention modules. To balance robustness and controllability, we adopt a mixed-conditioning schedule: **10%** scene-only samples, **25%** motion-only samples, and the remaining samples use full conditioning. To improve long-horizon generation, we additionally include motion-video context with **50%** probability by using the first 1–9 frames as motion history, which stabilizes early temporal dynamics and reduces drift in long rollouts. We also apply extensive bbox augmentations (e.g., scaling and jittering), which encourage the model to treat the 2D bbox as a "soft guide" for human placement rather than a rigid constraint, thereby improving robustness to bbox misalignment during inference. We set the background label to $\alpha = 100$, a constant much larger than the canonical map size.

B Training Data and Pre-processing

Data sources. Our training data combine public datasets and licensed human-centric footage purchased from the online catalog of a commercial stock-media provider, in an approximate 1:1 ratio. The commercially licensed portion is designed to complement open-source data with diverse real-world compositions that are important for our task, especially outdoor human-motion videos with varied camera behavior and global motion patterns. To improve coverage, we deliberately balance several factors, including videos with and without camera motion, and videos where the subject remains roughly stationary versus exhibits clear global movement. This design helps expose ONE-SHOT to a broader range of compositional settings while remaining easy to reproduce in practice.

Pre-processing philosophy. Our 3D pre-processing is intentionally lightweight and modular. Rather than requiring highly accurate scene reconstruction or explicit human-scene global alignment, we rely on standard off-the-shelf tools to obtain approximate decoupled scene and human signals. Moreover, because ONE-SHOT is trained in a decoupled and mixed-conditioning manner, not every training sample needs to provide every signal. For scene-only videos, we only extract scene-related conditions such as point clouds and camera trajectories; for motion-centric videos, including commercially licensed videos with static cameras, we primarily extract the human-motion signal. This flexibility reduces the dependence on fragile full-pipeline preprocessing.

Decoupled signal construction. When full preprocessing is available, we reconstruct humans in world coordinates using Human3R [42] and estimate dense depth using Depth-Anything v3 [57]. We then recover camera extrinsics and build a background point cloud with the human removed.

This process provides the conditioning inputs required by Sec. 4 of the main paper, including the background point cloud $\mathcal{P}$, camera trajectory $\{\mathbf{T}_t\}_{t=1}^T$, appearance references $\mathbf{y}^{\text{ref}}$ and optional memory $\mathbf{y}^{\text{mem}}$, and the target motion sequence $\mathbf{m}_{1:T}$.

Test sets. For self-reconstruction, we use **Traj100** following the same protocol as RealisMotion [10]. For cross-composition, we curate 50 real seed videos spanning 5 identities, 10 motions, and 10 scenes/camera trajectories, and enumerate cross-combinations of these conditions. We filter candidates for physical and cinematographic feasibility and for valid control inputs across all compared methods, excluding motion-scene or motion-camera conflicts and required-preprocessing failures. All filtering is completed before generation and independently of output quality, yielding 83 common-support feasible combinations. For cross-composition, we compute FID/FVD between each method's 83 generated videos and the same 50 real seed videos as auxiliary distribution-level realism metrics.

C Details on 2D Bounding Box as Coarse Placement

Bbox during Training. During training, to ensure that ONE-SHOT treats the bounding box only as a soft spatial cue indicating where the sparse human should appear, we apply extensive bbox augmentations (e.g., scaling and jittering). This encourages the model to use the 2D prior as a flexible guide rather than a rigid constraint. As a result, ONE-SHOT remains robust to moderate bbox misalignment and can still synthesize natural foot-ground contact. Unlike a hard mask, the shared background coordinate retains motion attention outside the bbox. This preserves access to motion cues when bbox scaling or jitter places body parts outside the specified region.

Bbox during Testing. ONE-SHOT targets compositional generation by freely combining a new scene, human motion, and camera trajectory. Accordingly, we support flexible bbox configuration during inference, including automatically propagating a user-specified first-frame bbox, reusing detected bboxes from source footage when applicable, and even using fixed-size bboxes in tracking or hand-held shots:

1. In general cases (e.g., most examples in our demo video), users specify a coarse square bbox in the first frame via the interface in Fig. 7, denoted as $\mathcal{B}_0$, as a soft signal of the initial human placement. The subsequent placement is then automatically determined by the source SMPL-X motion $\{\mathbf{\Gamma}_t\}_{t=1}^T$ and the target camera motion $\{\mathbf{T}_t\}_{t=1}^T$ and its intrinsic parameters.

   We define the bbox at frame t as $\mathcal{B}_t = \{(u_t, v_t), a_t\}$, where (u_t, v_t) is the bbox center and a_t is the side length. Let the first-frame bbox be $\mathcal{B}_0 = \{(u_0, v_0), a_0\}$. Since the metric human height H can be obtained from the SMPL-X motion $\{\mathbf{\Gamma}_t\}_{t=1}^T$, and in our canonical-space setting the human occupies a fixed proportion (e.g., 90%) of the bbox side length, the first-frame root depth can be easily estimated by perspective projection as

$$Z_0 = \frac{f \cdot H}{0.9 a_0}, \tag{13}$$

   where f is the focal length of the target camera.

   Given the first-frame bbox $\mathcal{B}_0$, we can then back-project the 2D bbox center to obtain the initial 3D human root location $P_0 = (X_0, Y_0, Z_0)$:

$$P_0 = (X_0, Y_0, Z_0) = \left(\frac{(u_0 - c_x) \cdot Z_0}{f}, \frac{(v_0 - c_y) \cdot Z_0}{f}, Z_0 \right), \tag{14}$$

   where c_x, c_y, and f are the intrinsic parameters of the target camera.

   After obtaining the first-frame 3D human root location, we propagate the root trajectory by following the relative motion of the desired SMPL-X sequence:

$$\{P_t\}_{t=1}^T = \{P_0 + \Delta\mathbf{\Gamma}_t\}_{t=1}^T, \tag{15}$$

   where $\{\Delta\mathbf{\Gamma}_t\}_{t=1}^T$ denotes the relative root motion derived from the source SMPL-X motion $\{\mathbf{\Gamma}_t\}_{t=1}^T$.

   For each following frame, the propagated 3D root is projected under the target camera extrinsics $\{\mathbf{T}_t\}_{t=1}^T$ to obtain the bbox center, while the bbox scale is updated according to the projected depth:

$$\{\mathcal{B}_t\}_{t=1}^T = \Pi\big(\{P_t\}_{t=1}^T; \{\mathbf{T}_t\}_{t=1}^T\big), \tag{16}$$

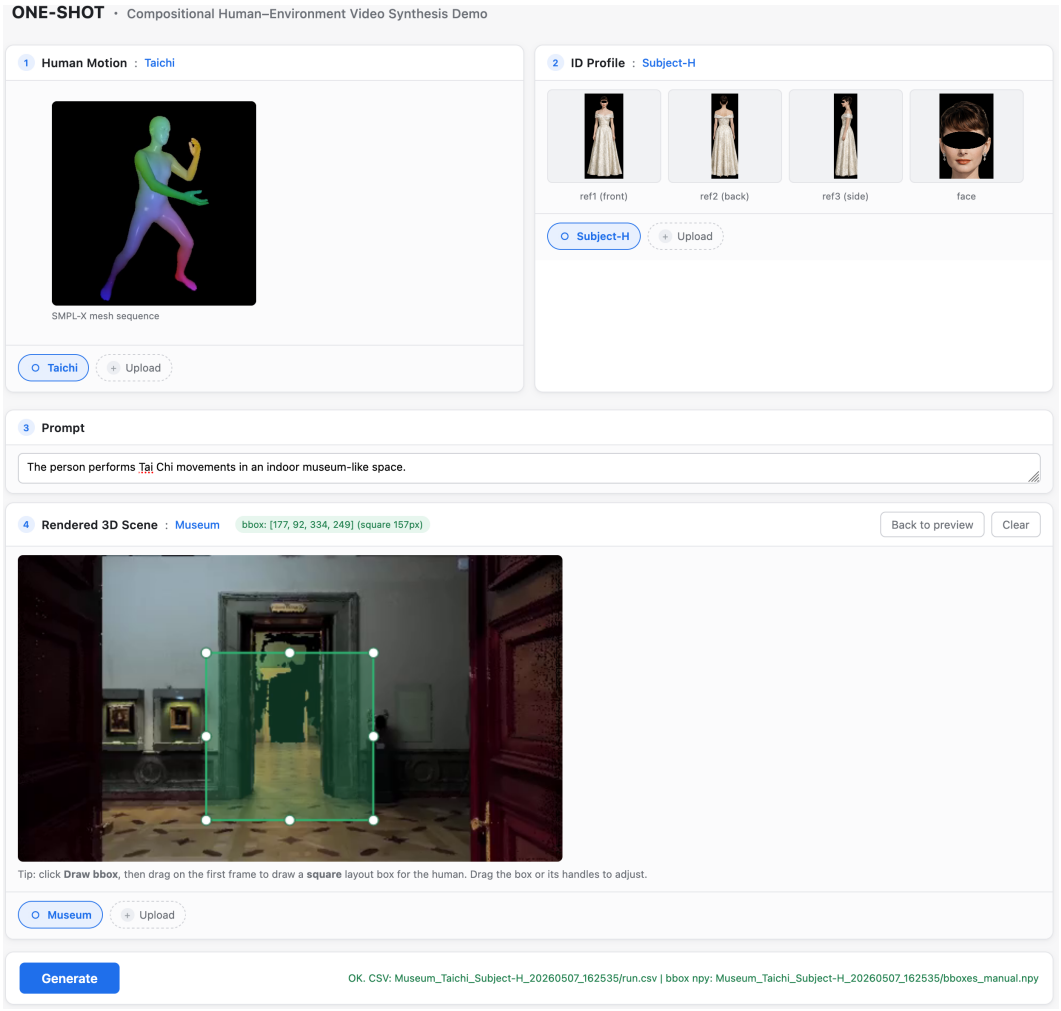


Figure 7: **Interactive inference interface for coarse bbox specification.** Users select the driving motion, identity profile, prompt, and target 3D scene, then only need to draw a coarse square bbox on the first frame of the rendered scene. This matches our test-time formulation, where the bbox serves as a soft initial placement cue rather than a precise human mask.

where $\Pi(\cdot)$ denotes the bbox construction procedure based on projection and depth-dependent scaling. This provides a simple and automatic way to obtain coarse bboxes for arbitrary human motion under arbitrary target camera trajectories, while the only user-specified factor is the coarse first-frame placement.

2. ONE-SHOT is also flexible when users do not want to manually define a bbox. A simple workaround is to reuse detected bboxes when the target scene is derived from footage that originally contained a person. If the target camera trajectory remains the same as that source scene video (e.g., video reenactment), the following bboxes can also be directly reused. Otherwise, they can still be automatically calculated as above under a new target camera trajectory.
3. In other scene-only cases, such as tracking or hand-held shots, the bbox can even remain static across frames because the camera motion itself already mimics a following perspective. This is significantly more user-friendly than previous methods requiring complex and fragile 3D pre-processing.

Interactive Inference Interface. To make coarse bbox specification more intuitive, we provide a lightweight inference interface, as shown in Fig. 7. Users select the driving motion, identity profile, text prompt, and target 3D scene, then draw a square bbox on the first frame of the rendered scene.

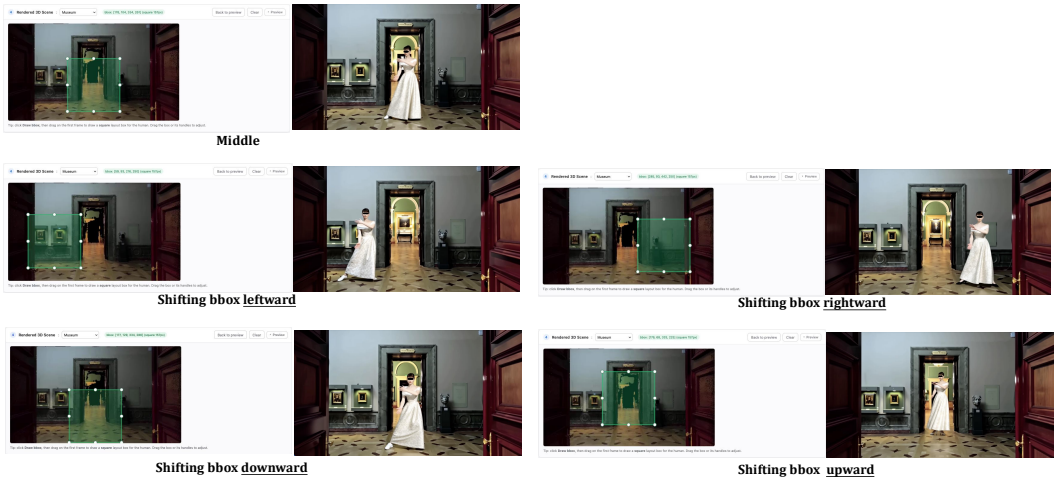


Figure 8: **Robustness to bbox misalignment at inference time.** We test ONE-SHOT with perturbed first-frame bboxes placed higher, lower, left, or right relative to a reasonable placement. Although the bbox only provides a coarse spatial prior, ONE-SHOT still synthesizes humans in the intended region with plausible scene interaction and natural foot-ground contact. This supports our claim that the bbox is treated as a soft guide rather than a rigid constraint.

This interface illustrates our intended test-time usage: the bbox is only a soft placement cue, while ONE-SHOT handles the detailed human synthesis and scene interaction.

Robustness to Bbox Misalignment. To evaluate robustness, we perturb the user-specified bbox by shifting it upward, downward, leftward, or rightward, as shown in Fig. 8. ONE-SHOT still places the subject in the intended region and maintains plausible interaction with the scene, including natural body placement and foot-ground contact. These results further support treating the bbox as a flexible guide rather than a rigid constraint.

D More Discussion

In this section, we provide additional discussion about our method as follows:

1) Fair comparison with baselines. For all baselines, we use their official checkpoints and inference pipelines. Best-effort adaptation is only applied when their interface does not match our compositional task with disentangled control of motion, scene, camera, and identity. The trajectory shift observed in some baselines is not caused by a different target camera trajectory, but by their fragile human-scene global alignment, where scale or magnitude errors can lead to unstable placement or apparent camera mismatch. To further strengthen fairness, we also perform a best-effort WAN-VACE adaptation by fine-tuning it on the same training clips and SMPL-X annotations used by ONE-SHOT so that it can take SMPL-X as compositional control input. Since the standard VACE interface represents controls as frame/mask/reference conditions on the target video grid, we convert the SMPL-X signals into dense target-grid condition maps for fine-tuning. ONE-SHOT remains clearly better (see Tab. 3). WAN-VACE also shows weaker identity preservation and more frequent unstable human scale, placement mismatch, and weaker human-scene alignment.

Table 3: **Comparison with fine-tuned WAN-VACE on self-reconstruction and cross-composition.** WAN-VACE is adapted to our setting by fine-tuning on the same training clips and SMPL-X annotations as ONE-SHOT, with SMPL-X signals converted into dense target-grid condition maps for its video/mask-style conditioning interface. Lower is better for FID/FVD. Higher is better for others.

	Self-reconstruction (Traj100)					**Cross-composition**					
Method	FID↓	FVD↓	MS↑	BC↑	SubSim↑	FID↓	FVD↓	MS↑	BC↑	SubSim↑	Face↑
WAN-VACE (fine-tuned)	20.96	185.09	0.935	0.955	0.311	151.68	1465.3	0.912	0.947	0.182	0.099
ONE-SHOT **(ours)**	**16.88**	**181.17**	**0.940**	**0.959**	**0.317**	**150.98**	**1382.4**	**0.993**	**0.963**	**0.349**	**0.241**

2) How faithfully camera trajectories and environmental layout are preserved. ONE-SHOT uses per-frame RGB-D projections rendered from the target point cloud under the target camera trajectory, so camera motion and scene layout are highly preserved through explicit geometric conditioning.

3) Dependence on 3D preprocessing / pipeline clarity. ONE-SHOT uses standard off-the-shelf preprocessing for scene and human reconstruction, as discussed in Sec. B, rather than first explicitly aligning human motion with scene geometry in 3D as in prior methods. We further apply augmentation/dropout on scene control during training and support pure RGB scene input at inference, making the model less sensitive to scene control quality.

E Minute-Scale Consistency Evaluation

We construct a controlled viewpoint-revisit benchmark using ten 60-second videos in different scenes. Each video contains six 10-second camera loops: starting from an anchor view at $t = 0$, the camera moves away and returns to the same pose at $t = 10, \ldots, 60$s, with the subject in a matched frontal pose at every return. We compare each revisit with the $t = 0$ anchor to measure accumulated identity and scene-appearance drift under matched viewpoints. All values are averaged over the ten videos.

Identity consistency. We compare the full model with a variant without the face reference (Tab. 4). From 10s to 60s, Face-Sim decreases from 0.614 to 0.580 with the full model, compared with 0.597 to 0.483 without the face reference. The full-model advantage grows from +0.017 at 10s to +0.097 at 60s, showing that the static face reference reduces, although does not eliminate, long-horizon identity drift.

Table 4: **Identity consistency as a function of time.** Face-Sim between each revisit and the frontal anchor face at $t = 0$. Higher is better.

Variant	10s	20s	30s	40s	50s	60s
ONE-SHOT full	0.614	0.615	0.603	0.597	0.591	0.580
w/o Face-Ref	0.597	0.589	0.566	0.549	0.510	0.483
Gap	**+0.017**	**+0.026**	**+0.037**	**+0.048**	**+0.081**	**+0.097**

Viewpoint-revisit consistency. We compare the full model with a variant without context memory (Tab. 5). Averaged over the six revisits, memory increases DINO similarity from 0.744 to 0.849 and reduces LPIPS from 0.293 to 0.155 (a 47% reduction). The full model achieves better mean DINO similarity and LPIPS at each revisit time, supporting the role of context memory in mitigating accumulated scene-appearance drift.

Table 5: **Viewpoint-revisit consistency as a function of time.** Each return is compared with the anchor view at $t = 0$. Before computing DINO similarity and LPIPS, we mask out the subject and evaluate only the remaining scene region. Higher DINO and lower LPIPS are better.

Metric / variant	10s	20s	30s	40s	50s	60s	Mean
DINO↑ — full	**0.949**	**0.909**	**0.871**	**0.834**	**0.781**	**0.749**	**0.849**
DINO↑ — w/o memory	0.796	0.859	0.733	0.758	0.690	0.627	0.744
LPIPS↓ — full	**0.057**	**0.111**	**0.153**	**0.175**	**0.204**	**0.230**	**0.155**
LPIPS↓ — w/o memory	0.210	0.266	0.285	0.278	0.369	0.352	0.293

F Motion and Spatial Control Evaluation

Metric definitions. Tab. 6 evaluates motion following, subject placement, global trajectory, camera control, and foot-ground behavior on Traj100. Bounding-box IoU measures the overlap between the generated and reference subject boxes, and 2D PCK@0.1 measures the fraction of keypoints within a normalized distance threshold of 0.1. Traj@50cm is the fraction of clips in which the Sim(3)-aligned global root trajectory exceeds a 50cm error at any frame relative to ground truth. Contact-fidelity is the median GT-referenced foot-to-ground gap error in millimeters, a proxy for ground-contact

Table 6: **Motion and spatial control on Traj100.** All results are evaluated by us under the protocol described in this section. Best results are bold; second-best results are bold and underlined. Contact-fid. is in mm, Foot-skate in mm/frame, and Cam. ATE in m.

Method	Traj@50cm↓	Contact-fid.↓	Foot-skate↓	Cam. ATE↓	bbox IoU↑	PCK@0.1↑
WanAnimate [6]	**<u>0.450</u>**	122.8	58.2	0.181	0.691	0.385
RealisDance [31]	0.510	188.8	**54.3**	0.138	0.606	0.263
Uni3C [9]	0.470	199.5	77.8	0.176	0.273	0.050
RealisMotion [10]	**0.280**	**<u>78.3</u>**	55.9	**0.051**	**0.729**	**0.449**
ONE-SHOT **(ours)**	**0.280**	**69.1**	**<u>55.1</u>**	**<u>0.059</u>**	**<u>0.700</u>**	**<u>0.400</u>**

quality. Foot-skate reports foot-skating error in millimeters per frame, and Camera ATE reports camera absolute trajectory error in meters.

Local 3D motion following. 3D PA-MPJPE measures mean per-joint position error after Procrustes alignment, in millimeters. ONE-SHOT uses canonical-space motion with coarse spatial grounding, allowing the VFM to adapt precise body placement to the target scene. WanAnimate achieves lower PA-MPJPE (24.3 vs. 39.9 mm), while ONE-SHOT obtains lower foot-ground contact error and the joint-lowest trajectory-failure rate under this protocol (Tab. 6).

G Limitation

Our approach relies on the quality of the reconstructed scene point cloud and camera trajectory: extremely inaccurate, sparse, or noisy point clouds (e.g., due to imperfect depth estimation or human removal) can reduce scene consistency and degrade generation quality. Errors from the human estimation module (Human3R in this work) may also propagate to the final results. Conflicting user-specified conditions pose a further challenge: for example, a walking trajectory that passes through a solid wall in the target scene cannot be followed without violating the scene geometry. Under such incompatible conditions, the generated human may exhibit incorrect contact, penetration, floating limbs, or local pose deformation.

We observe qualitative generalization to some uncommon motions, including Tai Chi with a longsword (Fig. 1), but have not systematically evaluated performance as a function of pose rarity or OOD severity. Extreme articulation, rare interactions, and multi-person scenes remain challenging. Temporal drift may also accumulate during very long generations.

H Broader Impact

ONE-SHOT advances controllable human-environment video generation and may support creative content production and long-horizon video synthesis. At the same time, identity-conditioned generation can be misused for misleading media or non-consensual likeness synthesis. To mitigate these risks, access to the capability-bearing checkpoint is gated and requires application approval. Release access will require an institutional affiliation, a brief intended-use description, and agreement to a use policy that prohibits non-consensual likeness synthesis and specifies out-of-scope uses. The official release pipeline will additionally apply synthetic watermarking, with the access restrictions and watermarking mechanism documented in the accompanying model card/data statement.

We also take a conservative approach to data and identity release. Raw training videos will not be released, and any public-facing derived assets will be de-identified through face blurring and removal of identifying metadata. Finally, while our data collection aims to cover varied real-world compositions, we have not systematically characterized demographic composition or performed disaggregated evaluation. We therefore do not make claims about performance consistency across demographic groups.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the main contributions, including canonical-space motion decoupling, Dynamic-Grounded-RoPE, Hybrid Context Integration, and the evaluated compositional human-environment generation setting. The claims are supported by quantitative comparisons, qualitative results, and ablation studies in the experiments (Sec.5).

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The main limitations are discussed in Sec. 6, with further failure-case and generalization analysis in Sec. G of the Supplementary Material.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: The paper does not present theoretical results or formal theorems. The mathematical formulations are used to define the training objective and model components.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The paper describes the model architecture, conditioning signals, training objective, mixed-conditioning schedule, datasets, evaluation protocols, baselines, and metrics (Sec.3, Sec.4, Sec.5.1). Additional preprocessing, training, and evaluation details are provided in the supplementary material (Sec.A, Sec.B, Sec.C). Code is publicly available at `https://github.com/MartaYang/ONE-SHOT-code` to support reproduction.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: Code is publicly available, while access to the checkpoint is gated and raw training videos will not be released. Derived assets will be shared only where licensing and consent permit, subject to the restrictions described in Sec. H.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: In Sec. 5.1, the experimental setup describes the datasets, training clip length, metrics, and ablations. Hyperparameters, preprocessing details, and additional implementation details are provided in the supplementary material (Sec.A, Sec.B, Sec.C).

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

Justification: We do not report error bars or statistical significance tests because training and evaluating large-scale video generation models across multiple random seeds is computationally expensive. Instead, we evaluate across multiple datasets/settings, report standard video-generation metrics, and provide ablation and qualitative comparisons.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The supplementary material Sec. A reports the GPU type and number, training resolution, training iterations, learning rates, and mixed-conditioning schedule used for the main experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Data licensing and evaluation-identity consent are described in checklist items 12 and 13, while privacy protections and release restrictions are discussed in Sec. H. Public data are used under their applicable terms, and the purchased training footage is covered by a written agreement for research use and model training.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.

- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Sec. H discusses creative applications, non-consensual likeness synthesis and misleading-media risks, release safeguards, and the absence of demographic characterization and disaggregated evaluation.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: Access to the capability-bearing checkpoint is gated and requires application approval. Release access will require institutional affiliation, an intended-use description, and agreement to a use policy prohibiting non-consensual likeness synthesis; the official release pipeline will apply synthetic watermarking, with these measures documented in a model card/data statement (Sec. H).

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.

- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Existing datasets and models are cited. MotionX is used under CC BY-NC-SA 4.0, EMDB under ETH Zurich's Terms of Use, and ARKitScenes under Apple's official license terms. The commercial footage is covered by a written research/model-training agreement, with the provider confirming the necessary licensing rights and applicable permissions for identifiable individuals.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The benchmark composition and preprocessing are documented in Sec. B. Two identities are recorded participants with written research-use consent; the other three use AI-generated full-body references derived from public-figure images, which do not establish consent. Public-figure examples have been face-blurred and watermarked; any remaining identifiable material involving the three public figures will be removed from the paper/demo and will not be released. Material from consenting participants will be shared only within their written consent.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: No crowdsourcing or user study was conducted. Reference recordings were obtained with written consent and used solely as inputs for model inference.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: The work does not involve a human-subject experiment. Consented recordings are used solely as reference inputs for inference.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLMs are not used as an important, original, or non-standard component of ONE-SHOT's core method.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.