BODYWORLD: STREAMING HUMAN-CENTRIC WORLDS WITH INDEPENDENT BODY AND CAMERA CONTROL

Anonymous authors
Paper under double-blind review

ABSTRACT

Human-centric video worlds require control over both what a person does and how the action is observed. We introduce BodyWorld, a framework that streams video from a reference image and independently editable body and camera commands. Decoupled geometric conditioning combines articulated body guidance with complementary scene appearance and viewing geometry. To maintain consistency as motion and viewpoint change, a static point-cloud memory anchors the scene, while Pose-Addressable History Memory recalls relevant generated observations to preserve human appearance beyond recent context. A three-stage curriculum learns joint control, introduces block-causal generation through pose-addressed teacher forcing, and distills few-step continuation with self forcing and distribution matching. Experiments across human motion, camera motion, pose recurrence, and long sequences show improved visual fidelity over prior methods and better appearance consistency across visibility changes. With cached history reuse, two-step inference reaches 13.36 generation FPS on a single NVIDIA H200.

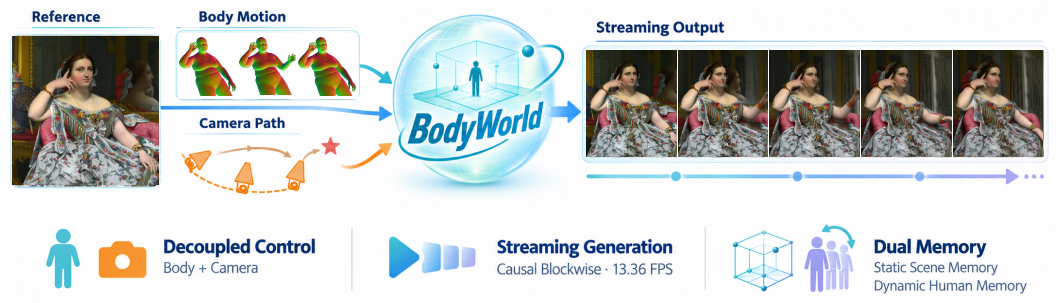

Figure 1: **BodyWorld brings a still image into motion.** Given a reference image and independent body-motion and camera controls, BodyWorld generates video block by block. Here, Ingres's *Madame Moitessier* (1856) becomes a moving portrait: the sitter raises and lowers her left arm as the camera orbits around her.

1 INTRODUCTION

Video generators are becoming interfaces to worlds rather than one-shot renderers of clips. In a video world model, each transition depends on the generated past and an external action, turning synthesis into a controllable process. GameNGen demonstrates neural game simulation, while Genie 3, LingBot-World, WorldPlay, and LingBot-World 2.0 advance navigable, persistent, and event-driven visual environments (Valevski et al., 2025; Parker-Holder & Fruchter, 2025; Robbyant Team et al., 2026; Sun et al., 2025; Gao et al., 2026b). These systems open new possibilities for embodied agents, game engines, virtual reality, and interactive content creation. Yet their controls usually operate at the level of navigation, agent actions, or semantic events; they do not generally expose both frame-accurate articulated human geometry and an independently specified viewpoint.

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

For a human-centric world, interactivity must include the person as an explicitly controlled actor. A user may want a person to turn, bend, or traverse a scene while directing the observer along a separate camera path. These two commands are geometrically coupled in the resulting image, but semantically distinct: one specifies what the person does, while the other specifies how the action is observed. We therefore ask whether *articulated human action can drive each subsequent visual transition alongside independent camera control*.

Two research lines approach this problem from complementary sides. Reference-based human animation offers detailed pose following and appearance preservation through skeletons, dense renderings, or parametric-body guidance (Zhang et al., 2025a; Zhou et al., 2025; Zhu et al., 2024). Concurrent LiveAnimate extends this line to block-causal, distilled, long-form streaming animation (Zhang et al., 2026), while large camera motion remains outside its current scope. Camera- and geometry-conditioned generation instead emphasizes viewpoint control and scene coherence. Uni3C combines 3D-enhanced human and camera control, and Directing the World provides autoregressive SMPL–camera control for world exploration (Cao et al., 2025; Wang et al., 2026b). These advances motivate a chronological interface that brings together precise body geometry, independent viewpoint commands, reference-anchored scene evidence, and few-step continuation from generated history.

We introduce **BodyWorld**, a framework for *human-centric streaming world modeling*. Before generating block k, BodyWorld receives that block's body and camera commands, continues from the committed prefix, emits and commits the result, and only then accepts new commands. Future commands can remain unspecified until their block is generated. Body and camera controls remain independently editable, while their rendered conditions share one coordinate system. This sequence makes human action part of each world transition, offering a basis for interactive action direction and pose-controlled games. Streaming thus allows the user to revise body and camera commands as the generated world unfolds.

We train BodyWorld in three stages. Stage I learns joint human–camera control with bidirectional training. Camera rays specify viewing geometry but not newly exposed appearance, whereas reference point-cloud projections carry observed scene evidence but become sparse away from the source view. We combine them with projected SMPL-family normals (Loper et al., 2015), injecting point and ray signals through separate residual adapters. Stage II initializes the autoregressive model through causal teacher forcing on clean historical latents. Since recent context may lose informative body views when a person turns or revisits a pose, PAHM retrieves a contiguous pose-relevant historical pair alongside five recent latents under a fixed seven-slot budget, keeping every recalled modality at its original timestamp. Teacher forcing leaves a gap between clean training histories and the model's own rollout errors, while reducing denoising steps can further affect quality. Stage III addresses both through generated-history self forcing and distribution matching distillation (DMD), yielding a four-step student (Huang et al., 2025; Yin et al., 2024). Rolling KV reuse and a bounded pose bank of five pair entries further accelerate inference.

We support this study with aligned body, camera, and scene conditions and evaluate pose following, camera control, appearance, recurrence, long-horizon rollout, and efficiency. Our contributions are:

- **A human-centric streaming formulation** that makes articulated body geometry and independent camera motion explicit conditions of each chronological visual transition.
- **A streaming generation framework** combining decoupled geometric conditioning, fixed-budget PAHM, and a three-stage adaptation–causalization–self-forcing/DMD curriculum.
- **Aligned evaluation of the person and world** across joint control, background and appearance preservation, pose recurrence, long-horizon consistency, and efficiency.

2 Related Work

Interactive video world models. GameNGen models game dynamics, while Genie 3 and LingBot-World support navigation through generated environments (Valevski et al., 2025; Parker-Holder & Fruchter, 2025; Robbyant Team et al., 2026). WorldPlay reconstitutes context from past frames to improve long-term geometric consistency; LingBot-World 2.0 adds character actions and text-driven events (Sun et al., 2025; Gao et al., 2026b). MotionStream combines point-trajectory control, including camera-induced tracks, with Self-Forcing and DMD (Shin et al., 2026; Huang

et al., 2025; Yin et al., 2024). Beyond interactive content, DreamDojo simulates robot-action outcomes, DreamZero jointly predicts video and actions for robot control, and GAIA-2 generates controlled multi-camera driving scenes (Gao et al., 2026a; Ye et al., 2026; Russell et al., 2025). BodyWorld instead targets articulated human control with an independently specified camera. It shares causal few-step training with MotionStream, but uses explicit body and scene geometry; its memory recalls human appearance across poses and viewpoints rather than WorldPlay's scene-revisit context.

Controllable human animation. Animate Anyone preserves appearance through ReferenceNet; MagicAnimate combines appearance encoding with temporal diffusion (Hu, 2024; Xu et al., 2024). MimicMotion uses confidence-aware pose guidance, while CHAMP adds SMPL-derived geometry (Zhang et al., 2025a; Zhu et al., 2024). HumanDiT supports variable-length generation with a prefix-latent reference, and UniAnimate enables long-video continuation through first-frame conditioning (Gan et al., 2025; Wang et al., 2025b). RealisDance-DiT minimally modifies a video prior; UniAnimate-DiT uses LoRA and a lightweight pose encoder (Zhou et al., 2025; Wang et al., 2025c). One-to-All handles reference–pose misalignment, Wan-Animate combines body/facial motion transfer with character replacement, and Wan-Animate-2 adds text-controlled viewpoints and a causal Lite variant (Shi et al., 2026; Cheng et al., 2025; Wang et al., 2026a).

For streaming animation, StreamAvatar anchors autoregressive avatars to a reference, while SoulX-LiveAct combines Neighbor Forcing with ConvKV memory for hour-scale generation (Sun et al., 2026; Zhen et al., 2026). Concurrent LiveAnimate uses three-step block-causal generation and pose-matched KV context, with large camera motion remaining a limitation (Zhang et al., 2026). By contrast, we introduce Pose-Addressable History Memory (PAHM), which retrieves a pose-relevant consecutive latent pair from a bounded bank and combines it with five recent latents under a fixed seven-slot context budget, preserving original timestamps and geometric conditions. BodyWorld further couples this memory with independent body–camera commands and reference-anchored scene geometry.

Camera and joint geometric control. CameraCtrl encodes camera poses, while MotionCtrl separates camera and object-motion control (He et al., 2025; Wang et al., 2024b). Wan-Move propagates first-frame features along point trajectories (Chu et al., 2025). For scene consistency, GEN3C renders a 3D cache built from seed and generated frames, and Spatia updates persistent point-cloud memory through visual SLAM (Ren et al., 2025; Zhao et al., 2025). BodyWorld uses a static reference-anchored point cloud for observed appearance and dense Plücker rays for viewing geometry where projections are absent. Uni3C aligns SMPL-X with scene points and also studies complementary Plücker-ray conditioning; 3STC-HIA combines human trajectories and camera movement with terrain-aware retargeting and visibility-guided latent fusion (Cao et al., 2025; Liu et al., 2026). Directing the World further introduces autoregressive SMPL–camera control, but its MMPL backbone first predicts segment-level planning frames, including a future endpoint (Wang et al., 2026b). BodyWorld commits blocks chronologically from generated history and current commands, combining joint geometric control with few-step continuation and command updates after each committed block.

3 BODYWORLD

BodyWorld streams a human-centric world under explicit, frame-level human pose and camera control. Given reference image I_0 and text prompt p, we infer a point cloud $\mathcal{P}_0 = \mathcal{G}(I_0)$ to initialize scene memory $\mathcal{P}$. For block k, body states S_k specify articulation and placement, while camera states C_k independently specify the viewpoint. Their aligned normal maps, scene projections, and camera rays form $U_k(S_k, C_k; \mathcal{P})$. With reference latent z_0 and pose-addressed history H_k containing recent and recalled latents and their conditions, generation follows

$$p_\theta(z^{(1:K)} \mid I_0, p, S, C) = \prod_{k=1}^{K} p_\theta\Big(z^{(k)} \mid z_0, H_k, U_k(S_k, C_k; \mathcal{P}), p\Big). \tag{1}$$

After each block is committed to history, the user can revise body and camera commands. Figure 2 summarizes training: Stage I learns bidirectional human–camera control (Section 3.1); Stage II initializes causal generation through teacher forcing with PAHM (Section 3.2); Stage III combines generated-history self forcing and DMD for few-step streaming (Section 3.3).

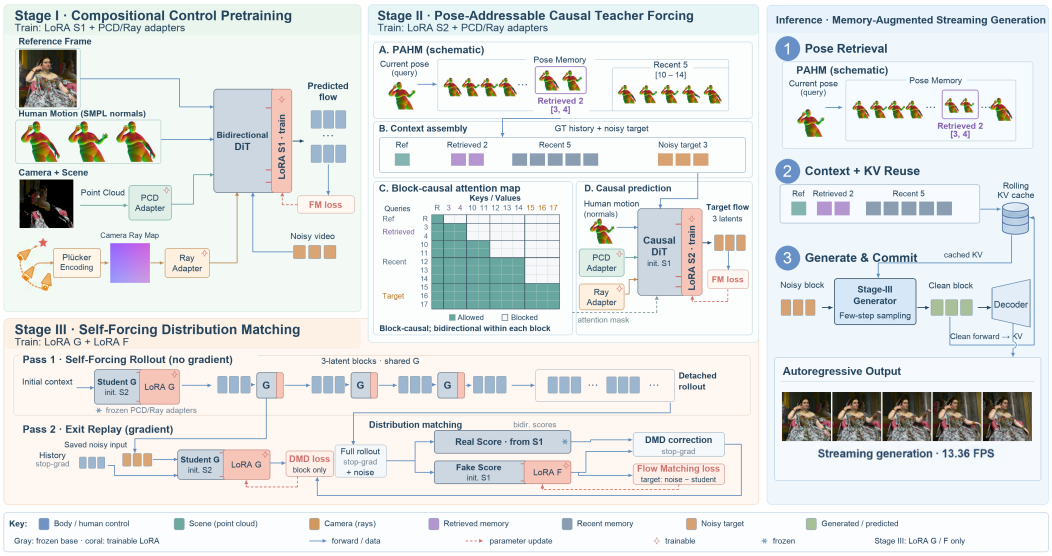

Figure 2: **BodyWorld training curriculum.** Stage I learns complementary human, scene, and camera controls with bidirectional full-clip training. Stage II imposes block causality and selects a pose-relevant historical pair together with recent context. Stage III exposes the generator to its own histories and uses distribution matching to obtain a four-step causal student.

3.1 Stage I: Compositional Human–Camera Control

Stage I combines human articulation and placement control with camera-trajectory control. We render the SMPL-family body state S_t under camera C_t as a normal map N_t (Loper et al., 2015). Camera control uses complementary point-cloud projections P_t and Plücker ray maps R_t: projections supply scene appearance with incomplete spatial coverage, whereas rays encode dense viewing geometry without appearance. Validity masks V_t mark projection support. Body and camera commands are separately editable, but share the target camera when rendered, aligning $U_t = (N_t, P_t, V_t, R_t)$.

Body normals follow the backbone's existing control pathway; point-cloud and ray features enter separate, non-weight-sharing adapters. The validity mask suppresses PCD features without projection support. Zero-initialized linear projections produce layer-specific residuals, which are scaled and added at selected Transformer layers. The shared backbone thus combines independently encoded scene appearance and camera geometry with human guidance.

Starting from Wan2.1-Fun Control (Team Wan et al., 2025), Stage I learns these controls through full-clip bidirectional flow matching (Lipman et al., 2023). Given clean latent z, noise ϵ, and $z_\tau = (1-\tau)z + \tau\epsilon$, the model predicts $\epsilon - z$ conditioned on the reference and rendered controls. Conditioning regularization is detailed in Appendix A.2.

3.2 Stage II: Pose-Addressed Causal Teacher Forcing

To accommodate incoming body and camera commands, Stage II adopts blockwise causal teacher forcing with three-latent blocks. Each training input comprises reference-image tokens, clean ground-truth history selected by PAHM, and a noisy target block. Reference and history tokens use the clean timestep; only the target is noised and supervised by flow matching. Reference tokens attend only to themselves, providing an appearance anchor visible to all temporal blocks. A block-causal mask lets the target read selected history and itself, while preventing history from accessing later blocks. At inference, generated latents replace ground-truth history. Details are in Appendix A.2.

Recalling appearance through pose. Human motion continually changes which body regions are visible. A recent-only window can discard previously generated textures, causing inconsistent ap-

pearance when those regions reappear. PAHM recalls such observations by pose: we crop control normals to the body region, concatenate occupancy masks, and pool and normalize them into descriptors. Averaging descriptors yields the current-block query q_k and a key $d(W)$ for each eligible historical frame pair W. Cosine similarity selects a contiguous pair, which joins the five most recent latents:

$$W_k^\star = \arg\max_{W \in \mathcal{W}_k} \cos(q_k, d(W)), \qquad H_k = \left[z_{W_k^\star}, z_{\text{recent}}^{1:5}\right]. \tag{2}$$

The pair recalls earlier appearance, while recent latents maintain continuity. All selected latents retain their original geometric conditions and timestamps. For accelerated streaming inference, retrieval uses a dynamic bank of at most five pair entries: pose-novel pairs enter the bank, and replacements favor pose diversity. Descriptor and fallback details are in Appendix A.2; bank maintenance is in Appendix B.

3.3 Stage III: Generated-History Few-Step Distillation

At inference, clean teacher-forced histories become imperfect model predictions, and few-step sampling further changes their distribution. We address both with Self Forcing and distribution matching distillation (DMD) (Huang et al., 2025; Yin et al., 2024): the student learns four-step generation while conditioning on its own detached rollouts.

Learning from the student's own trajectory. The generator inherits Stage II's causal attention and PAHM policy. At each update, we uniformly sample a temporal block T and a denoising exit $j \in \{1, \ldots, 4\}$. A complete sequence is generated without a backward graph, re-noising clean predictions between steps and stopping each block at exit j. We then replay block T's exit prediction with gradients, reusing its noisy input and detached generated history. This Monte Carlo estimate of the truncated gradient samples all blocks and denoising exits over training, while retaining generator activations for only one replayed forward.

Matching the video distribution. Let y be the student rollout, with only replayed block y_T retaining gradients, and $y_\tau = (1-\tau)\,\mathrm{sg}(y) + \tau\epsilon$. A frozen Stage-I real-score model supplies the target distribution; a fake-score model learns from detached student samples via flow matching. Both assess the complete noised rollout with bidirectional attention under $c = (z_0, U_{1:K}, p)$. Their clean-latent predictions $D_\cdot = y_\tau - \tau v_\cdot(y_\tau, \tau, c)$ define the DMD correction and surrogate loss:

$$\begin{aligned} \Delta_\tau &= a_\tau [D_{\text{fake}}(y_\tau, \tau, c) - D_{\text{real}}(y_\tau, \tau, c)], \\ \mathcal{L}_G &= \mathbb{E}_{j,T,\tau,\epsilon}\left[\frac{w_T}{2}\,\mathrm{MSE}(y_T, \mathrm{sg}(y_T - \Delta_{\tau,T}))\right]. \end{aligned} \tag{3}$$

Here $v_\cdot$ denotes a flow prediction, a_τ is a detached normalization and clipping factor, and w_T corrects for uniform block sampling. The stop-gradient target applies the fake-minus-real correction only through y_T. The causal generator receives no paired reconstruction loss. Sampling weights and score-training details are in Appendix A.2.

Efficient autoregressive inference. Following cached block-autoregressive inference (Huang et al., 2025; Zhang et al., 2026), we adapt history reuse to PAHM's nonconsecutive frame selection. Each committed block that may be reused undergoes one clean forward to store layer-wise KV indexed by original frame IDs. When PAHM recalls an earlier pose, the pair's cached KV is gathered alongside that of the five recent latents, preserving the keys' original rotary positions. All denoising steps of the current block reuse this selected context and compute only current-token queries, without re-encoding recalled frames. Cache eviction follows the five-entry pose-pair bank and recent/fallback buffers. Across blocks, rolling reuse retains historical states computed in their insertion context, avoiding window recomputation. Section 4.5 reports throughput; Appendix B details the cache variants and implementation optimizations.

4 Experimental Design

We compare bidirectional and autoregressive BodyWorld with prior methods on human motion across domains and camera regimes (Section 4.2), then assess explicit camera control (Section 4.3).

Ablations examine geometric conditioning, self forcing, DMD, and pose-addressed history, including long-sequence continuation and pose recurrence (Section 4.4). Finally, we measure accelerated generation throughput (Section 4.5).

4.1 Data, implementation, and comparison protocol

Evaluation data come from held-out sequences in our adaptation domains, additional datasets excluded from adaptation (OOD), and our newly captured PoseLoop dataset. We organize these sources by evaluation objective into Human Motion, Camera Motion, PoseLoop, and Long Sequence tracks, with a near-static-camera subset to isolate human control. Source composition, overlap, and subset criteria are detailed in Appendix A.3.

BodyWorld uses 480×480 resolution, with 50-step bidirectional and four-step autoregressive sampling. The autoregressive model conditions on generated history and both PCD/ray inputs. Quality comparisons recompute history windows; the efficiency study evaluates rolling-cache inference separately.

Baselines and input parity. We run the official Wan-Move, Uni3C, One-to-All, UniAnimate-DiT, and Wan2.2-Animate pipelines with common references and target timestamps while retaining their native control interfaces (Chu et al., 2025; Cao et al., 2025; Shi et al., 2026; Wang et al., 2025c; Cheng et al., 2025). Wan2.2-Animate additionally receives driving-face RGB (†). Camera Motion compares Uni3C and BodyWorld; Wan-Move's trajectory-mediated camera mode is not included there. Concurrent work Directing the World studies a similar setting, but no public implementation is available for comparison (Wang et al., 2026b).

Metrics and aggregation. We score lossless frames using PSNR, LPIPS, body-crop DINOv2 similarity, detector-agreement PCK, and error in adjacent-frame perceptual changes (Oquab et al., 2024; Zhang et al., 2018). Camera ATE/path and RRE are aligned VGGT proxies (Wang et al., 2025a). Scores average within clips, then within sources, then equally across sources; spatial metrics exclude the reference frame. Split construction, eligibility, definitions, and paired uncertainty are in Appendices A.3 and A.5.

4.2 Human motion under mixed camera regimes

Table 1 compares BodyWorld with five prior methods on in-domain and out-of-domain human motion under mixed camera conditions. BodyWorld Bidir. ranks first on all in-domain metrics and leads the out-of-domain comparison in PSNR, LPIPS, and temporal consistency.

BodyWorld AR retains most bidirectional quality with causal, few-step generation, outperforming external baselines on all in-domain metrics and out-of-domain PSNR/LPIPS. For fairer human-animation comparison with methods lacking camera control, Table 1(c) uses near-static-camera cases. Bidir. leads all metrics there, while AR maintains strong fidelity and temporal consistency. Figure 3 illustrates appearance preservation and local pose alignment.

4.3 Explicit camera motion

Table 2 compares BodyWorld with Uni3C under explicit camera control. Both BodyWorld variants achieve higher visual fidelity, better human appearance and pose alignment, and greater temporal consistency. Camera-trajectory accuracy remains comparable: Uni3C has slightly lower translational error, while BodyWorld AR has slightly lower rotational error.

4.4 Ablation studies

Table 3 isolates conditioning design, generated-history training, and history selection. Figure 4 shows the PAHM comparison; the full qualitative comparisons remain in Appendix C.3, Figures 8–10.

Decoupling scene evidence and viewpoint control. PCD projections anchor observed appearance but become incomplete under camera motion; rays specify the requested view without supplying

Table 1: **Human Motion under mixed and near-static cameras.** Source-macro means; pale gold/blue/green mark first/second/third before rounding. Cam./Human denote control capability and AR past-conditioned continuation. $\checkmark^t$: point trajectories; $\checkmark^s$: segment-level continuation. DINO/PCK are proxies; † uses driving-face RGB.

Method	Cam.	Human	AR	PSNR↑	LPIPS↓	Body DINO↑	PCK↑	Temp.↓
(a) Mixed cameras: in-domain								
Wan-Move	$\checkmark^t$	$\checkmark^t$	–	14.888	0.423	0.683	0.833	0.055
Uni3C	✓	✓	–	17.874	0.313	0.764	0.868	0.025
One-to-All	–	✓	$\checkmark^s$	15.703	0.371	0.740	0.868	0.035
UniAnimate-DiT	–	✓	–	17.265	0.335	0.766	0.877	0.028
Wan2.2-Animate†	–	✓	$\checkmark^s$	16.151	0.393	0.682	0.864	0.051
BodyWorld Bidir. 50-step	✓	✓	–	**19.642**	**0.232**	**0.807**	**0.902**	**0.013**
BodyWorld AR 4-step	✓	✓	✓	19.460	0.250	0.782	0.882	0.014
(b) Mixed cameras: out-of-domain								
Wan-Move	$\checkmark^t$	$\checkmark^t$	–	18.475	0.322	0.709	0.899	0.050
Uni3C	✓	✓	–	20.315	0.262	0.752	0.971	0.017
One-to-All	–	✓	$\checkmark^s$	15.659	0.356	0.741	**0.972**	0.047
UniAnimate-DiT	–	✓	–	21.226	0.244	**0.799**	0.969	0.042
Wan2.2-Animate†	–	✓	$\checkmark^s$	22.411	0.256	0.771	0.964	0.033
BodyWorld Bidir. 50-step	✓	✓	–	**23.490**	**0.203**	0.799	0.955	**0.015**
BodyWorld AR 4-step	✓	✓	✓	23.254	0.220	0.779	0.935	0.021
(c) Near-static cameras								
Wan-Move	$\checkmark^t$	$\checkmark^t$	–	18.080	0.292	0.753	0.893	0.035
Uni3C	✓	✓	–	22.594	0.156	0.820	0.932	0.010
One-to-All	–	✓	$\checkmark^s$	19.861	0.185	0.820	0.933	0.007
UniAnimate-DiT	–	✓	–	23.982	0.124	0.837	0.933	0.007
Wan2.2-Animate†	–	✓	$\checkmark^s$	22.179	0.192	0.787	0.917	0.031
BodyWorld Bidir. 50-step	✓	✓	–	**26.189**	**0.091**	**0.851**	**0.946**	**0.005**
BodyWorld AR 4-step	✓	✓	✓	25.223	0.105	0.825	0.932	0.006

Table 2: **Explicit Camera Motion.** ATE/path and RRE are aligned VGGT proxies; Body DINO/PCK use human cases. Colors follow Table 1.

Method	PSNR↑	LPIPS↓	Body DINO↑	PCK↑	Temp.↓	ATE/path↓	RRE↓
Uni3C	14.985	0.444	0.693	0.823	0.035	**0.077**	2.049
BodyWorld Bidir. 50-step	16.902	**0.354**	**0.730**	**0.865**	0.027	0.083	2.219
BodyWorld AR 4-step	**17.185**	0.378	0.707	0.863	**0.027**	0.083	**2.031**

appearance. Early joint fusion adds projected PCD/ray features to body-control latents before the backbone; our decoupled design injects them through separate multilayer adapters. In Table 3(a), decoupling improves both body PSNR and temporal consistency over early fusion. Compared with PCD-only conditioning, it improves moving-human temporal consistency with a modest trade-off in whole-frame PSNR. Supplementary Figure 8 shows how the decoupled design better follows the requested view change while preserving scene structure during camera motion.

Self forcing and distribution matching. Self forcing trains on the generated histories encountered at inference, while DMD aligns the few-step generator with the teacher's distribution. In Table 3(b), removing self forcing worsens appearance and pose metrics, especially late in long sequences, consistent with accumulated history errors. Removing DMD increases Human Motion PSNR but degrades LPIPS, Body DINO, and PCK, showing that pixel fidelity alone does not capture few-step perceptual and control quality. Supplementary Figure 9 makes these effects visible: the no-DMD variant loses clothing detail, while the no-self-forcing variant develops color and structural artifacts over time.

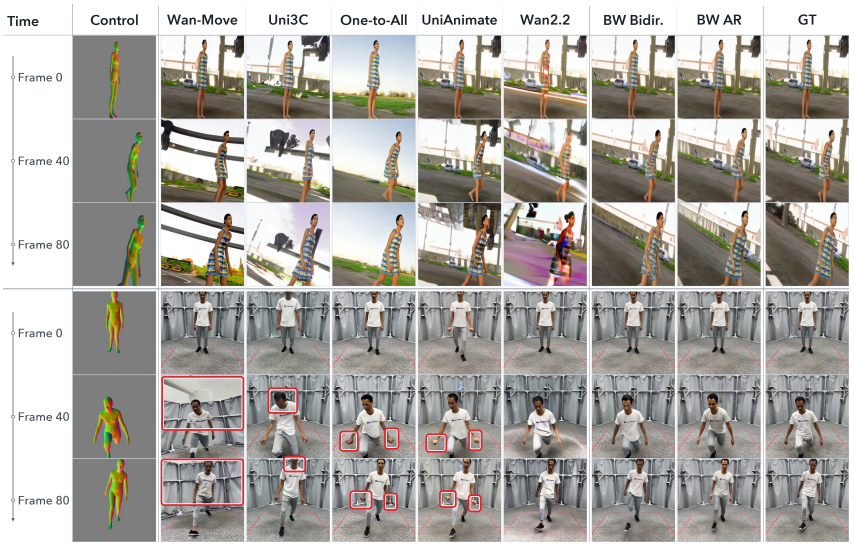

Figure 3: **Qualitative comparison.** Top: BodyWorld better preserves scene structure and human appearance during camera rotation; baselines exhibit background hallucinations and body deformation. Bottom: BodyWorld preserves the background better than Wan-Move and more accurately follows head and hand poses. Red boxes mark baseline errors; native-control comparisons appear in Appendix A.4.

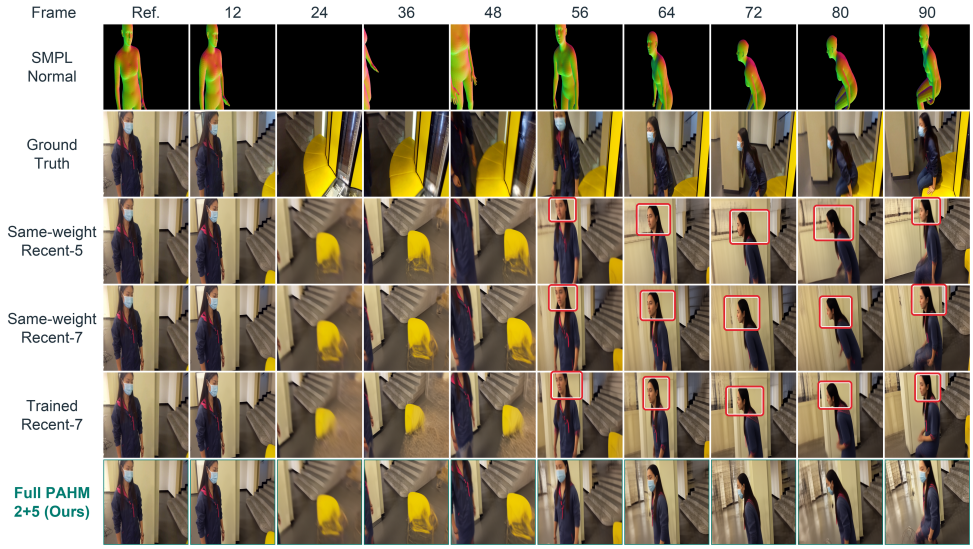

Figure 4: **Pose-addressed memory preserves appearance.** PAHM retains the face mask when the person turns back; recent-only variants lose it at the red-boxed frames. Recent-5 and Recent-7 share the full model's weights; Trained Recent-7 is separately trained. The SMPL row shows the body control.

Pose-addressable appearance memory. A recent window can lose evidence of body regions that temporarily leave view. Table 3(c) isolates retrieval with fixed weights: PAHM improves late-sequence LPIPS and Body DINO over Recent-7 at the same memory budget while maintaining comparable pose accuracy. These gains show the value of selecting relevant observations within a fixed history budget. In the face-mask example in Figure 4, same-weight recent-only variants omit the mask when the face reappears, whereas PAHM retains it. Pose-addressed recall thus preserves appearance details across visibility changes. Full-sequence, length-specific, and paired analyses are provided in Appendix C.2.

Table 3: **Ablation studies.** (a) Matched conditioning comparisons with body control and 50-step sampling; Δ_t measures adjacent-frame RGB-change error. (b) Separately trained variants; Trained Recent-7 uses recent-only history in training and inference. SF denotes self forcing. (c) Same-weight rows reuse the Full (PAHM + SF + DMD) weights from (b), changing only inference-time history without retraining. PAHM 2+5 uses two pose-retrieved and five recent latents; Recent-k uses only the latest k. (b,c) Four-step source-macro results; PoseLoop pools all lengths, and Long-H late uses frames 121–200. Colors follow Table 1, ranked within each panel.

(a) Scene and camera conditioning

	Body PSNR↑			Whole PSNR↑	
Conditioning	ID-H	Moving-H	Moving-H Δ_t↓	ID-H	Scene-C
PCD only	17.196	15.365	0.08397	**19.415**	**17.362**
Ray only	16.892	15.053	0.08901	18.097	12.781
Early joint fusion	17.057	15.125	0.08622	18.478	12.867
Decoupled PCD + ray (ours)	**17.240**	**15.415**	**0.08317**	19.244	17.204

	Human Motion			PoseLoop			Long-H late		
Variant	LPIPS↓	DINO↑	PCK↑	LPIPS↓	DINO↑	PCK↑	LPIPS↓	DINO↑	PCK↑
(b) Separately trained variants									
Full (PAHM + SF + DMD)	**0.240**	**0.781**	**0.900**	0.242	**0.764**	**0.974**	**0.442**	**0.615**	**0.883**
Trained Recent-7 + SF + DMD	0.241	0.779	0.893	**0.241**	0.759	0.974	0.456	0.582	0.878
PAHM + SF, w/o DMD	0.302	0.724	0.827	0.260	0.722	0.951	0.533	0.549	0.796
PAHM + DMD, w/o SF	0.266	0.764	0.865	0.243	0.747	0.968	0.522	0.540	0.810
(c) History selection with fixed full-model weights									
Full PAHM 2+5	**0.240**	**0.781**	**0.900**	**0.242**	**0.764**	**0.974**	**0.442**	**0.615**	0.883
Same-weight Recent-5	0.243	0.775	0.894	0.243	0.748	0.973	0.459	0.561	0.873
Same-weight Recent-7	0.241	0.778	0.897	0.243	0.756	0.973	0.452	0.581	**0.888**

4.5 Inference efficiency

On one H200, BodyWorld AR reaches 9.85 generation FPS in three steps and 13.36 FPS in two steps for 81-frame 480×480 output. With pre-encoded controls, generation timings cover adapters, denoising, scheduling, and cache updates; loading, input encoding, decoding, metrics, and video I/O are excluded. Three steps are the default. Appendix B reports timing, peak memory, and paired quality checks.

5 Conclusion

We presented BodyWorld, a framework for streaming human-centric video generation under explicit body-pose and camera control. Compositional conditioning combines human geometry with complementary scene and viewpoint signals, while PAHM retrieves pose-relevant history to retain human appearance across visibility changes. A three-stage curriculum produces a few-step causal generator, with cached history reuse enabling efficient continuation. Experiments demonstrate improved visual fidelity over prior methods across in-domain, out-of-domain, and camera-control evaluations. The autoregressive model retains most bidirectional quality, and pose-addressed history improves late-sequence appearance consistency at a fixed memory budget.

Limitations and future work. The current framework focuses on a single controlled person rather than multi-person interactions. Our training data emphasize motion-capture-style human activity, with relatively few sequences of sustained exploration through diverse environments, limiting long-horizon human roaming. Extending the framework to interacting people and expanding training data with longer, more varied exploration sequences are directions for future work.

AI USE STATEMENT

The authors led the research and used generative AI tools to assist with method development and experimental code, experiment planning and result interpretation, literature organization, manuscript drafting, editing and translation, and the creation of selected schematic subfigures. AI assistance was used under the authors' direction; the authors take responsibility for the final methods, results, text, and figures.

ETHICS STATEMENT

BodyWorld supports controllable visual synthesis and creative applications, but human appearance and motion synthesis can also be misused for impersonation or deceptive media. Responsible use and release should respect individuals' consent and privacy, comply with dataset licenses, and clearly identify generated media as synthetic. The system should not be used for non-consensual impersonation or deception.

REPRODUCIBILITY STATEMENT

Appendix A.3 documents dataset sources and split construction; Appendices A.2 and A.5 provide training details and evaluation protocols. Upon paper acceptance and institutional approval, we plan to release our newly collected dataset, code, and model checkpoints, subject to applicable data and model licenses.

REFERENCES

Ayce Idil Aytekin, Chuqiao Li, Diogo Luvizon, Rishabh Dabral, Martin Oswald, Marc Habermann, and Christian Theobalt. Physics-based human pose estimation from a single moving RGB camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 3930–3939, 2025. URL https://openaccess.thecvf.com/content/CVPR2025W/RHOBIN/html/Aytekin_Physics-based_Human_Pose_Estimation_from_a_Single_Moving_RGB_Camera_CVPRW_2025_paper.html.

Michael J. Black, Priyanka Patel, Joachim Tesch, and Jinlong Yang. BEDLAM: A synthetic dataset of bodies exhibiting detailed lifelike animated motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. URL https://arxiv.org/abs/2306.16940.

Zhongang Cai, Mingyuan Zhang, Jiawei Ren, Chen Wei, Daxuan Ren, Zhengyu Lin, Haiyu Zhao, Lei Yang, Chen Change Loy, and Ziwei Liu. Playing for 3D human recovery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. URL https://arxiv.org/abs/2110.07588.

Zhongang Cai, Daxuan Ren, Ailing Zeng, Zhengyu Lin, Tao Yu, Wenjia Wang, Xiangyu Fan, Yang Gao, Yifan Yu, Liang Pan, et al. HuMMan: Multi-modal 4D human dataset for versatile sensing and modeling. In *European Conference on Computer Vision*, 2022. URL https://arxiv.org/abs/2204.13686.

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 *ACM SIGGRAPH Asia 2025 Conference Papers*, pp. 1–12, 2025. doi: 10.1145/3757377.3763842. URL https://doi.org/10.1145/3757377.3763842.

Lin-Zhuo Chen, Jian Gao, Shangzhan Zhang, Yihang Chen, Ka Leong Cheng, Yipengjing Sun, Liangxiao Hu, Nan Xue, Xing Zhu, Yujun Shen, Yao Yao, and Yinghao Xu. LingBot-Map: Geometric context transformer for streaming 3D reconstruction. *arXiv preprint arXiv:2604.14141*, 2026. URL https://arxiv.org/abs/2604.14141.

Gang Cheng, Xin Gao, Li Hu, Siqi Hu, Mingyang Huang, Chaonan Ji, Ju Li, Dechao Meng, Jinwei Qi, Penchong Qiao, Zhen Shen, Yafei Song, Ke Sun, Linrui Tian, Feng Wang, Guangyuan Wang, Qi Wang, Zhongjian Wang, Jiayu Xiao, Sheng Xu, Bang Zhang, Peng Zhang, Xindi

Zhang, Zhe Zhang, Jingren Zhou, and Lian Zhuo. Wan-Animate: Unified character animation and replacement with holistic replication. *arXiv preprint arXiv:2509.14055*, 2025. URL https://arxiv.org/abs/2509.14055.

Ruihang Chu, Yefei He, Zhekai Chen, Shiwei Zhang, Xiaogang Xu, Bin Xia, Dingdong Wang, Hongwei Yi, Xihui Liu, Hengshuang Zhao, et al. Wan-Move: Motion-controllable video generation via latent trajectory guidance. In *Advances in Neural Information Processing Systems*, 2025. doi: 10.52202/085713-0014. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/00b67df24009747e8bbed4c2c6f9c825-Abstract-Conference.html.

Jihoon Chung, Cheng-hsin Wuu, Hsuan-ru Yang, Yu-Wing Tai, and Chi-Keung Tang. HAA500: Human-centric atomic action dataset with curated videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. URL https://arxiv.org/abs/2009.05224.

Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. ScanNet: Richly-annotated 3D reconstructions of indoor scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. URL https://arxiv.org/abs/1702.04405.

Yudi Dai, Yitai Lin, Xiping Lin, Chenglu Wen, Lan Xu, Hongwei Yi, Siqi Shen, Yuexin Ma, and Cheng Wang. SLOPER4D: A scene-aware dataset for global 4D human pose estimation in urban environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. URL https://arxiv.org/abs/2303.09095.

Qijun Gan, Yi Ren, Chen Zhang, Zhenhui Ye, Pan Xie, Xiang Yin, Zehuan Yuan, Bingyue Peng, and Jianke Zhu. HumanDiT: Pose-guided diffusion transformer for long-form human motion video generation. *arXiv preprint arXiv:2502.04847*, 2025. URL https://arxiv.org/abs/2502.04847.

Shenyuan Gao, William Liang, Kaiyuan Zheng, Ayaan Malik, Seonghyeon Ye, Sihyun Yu, et al. DreamDojo: A generalist robot world model from large-scale human videos. *arXiv preprint arXiv:2602.06949*, 2026a. URL https://arxiv.org/abs/2602.06949.

Zelin Gao, Qiuyu Wang, Jiapeng Zhu, Jingye Chen, Zichen Liu, Qingyan Bai, Jiahao Wang, Yufeng Yuan, Hanlin Wang, Yichong Lu, Ka Leong Cheng, Haojie Zhang, Jian Gao, Tianrui Feng, Yuzheng Liu, Yao Yao, Yinghao Xu, Xing Zhu, Yujun Shen, and Hao Ouyang. Infinite worlds with versatile interactions. *arXiv preprint arXiv:2607.07534*, 2026b. URL https://arxiv.org/abs/2607.07534.

Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. CameraCtrl: Enabling camera control for video diffusion models. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=Z4evOUYrk7.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017. URL https://arxiv.org/abs/1703.06870.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2106.09685.

Li Hu. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8153–8163, 2024. doi: 10.1109/CVPR52733.2024.00779. URL https://openaccess.thecvf.com/content/CVPR2024/html/Hu_Animate_Anyone_Consistent_and_Controllable_Image-to-Video_Synthesis_for_Character_Animation_CVPR_2024_paper.html.

Chun-Hao P. Huang, Hongwei Yi, Markus Höschle, Matvey Safroshkin, Tsvetelina Alexiadis, Senya Polikovsky, Daniel Scharstein, and Michael J. Black. Capturing and inferring dense full-body human-scene contact. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. URL https://arxiv.org/abs/2206.09553.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. In *Advances in Neural Information Processing Systems*, 2025. doi: 10.52202/085713-5576. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/f4823f831af67a3ef15e41a85434422a-Abstract-Conference.html.

Wei Jiang, Kwang Moo Yi, Golnoosh Samei, Oncel Tuzel, and Anurag Ranjan. NeuMan: Neural human radiance field from a single video. In *European Conference on Computer Vision*, 2022. URL https://arxiv.org/abs/2203.12575.

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. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. URL https://arxiv.org/abs/2308.16894.

Rawal Khirodkar, Jyun-Ting Song, Jinkun Cao, Zhengyi Luo, and Kris Kitani. Harmony4D: A video dataset for in-the-wild close human interactions. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. doi: 10.52202/079017-3407. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/c20b843d0c6b1b40a8e6eb9a44e719c9-Abstract-Datasets_and_Benchmarks_Track.html.

Ruilong Li, Shan Yang, David A. Ross, and Angjoo Kanazawa. AI choreographer: Music conditioned 3D dance generation with AIST++. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. URL https://arxiv.org/abs/2101.08779.

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. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. URL https://arxiv.org/abs/2307.00818.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2210.02747.

Deyin Liu, Jicheng Xu, Lin Yuanbo Wu, Xiaowei Zhao, Xiatian Zhu, Zhe Jin, and Anjan Dutta. 3D scene-adaptive trajectory-controllable human image animation with camera movement. In *European Conference on Computer Vision*, 2026. URL https://arxiv.org/abs/2606.30514.

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Transactions on Graphics*, 34(6):248:1–248:16, 2015. URL https://smpl.is.tue.mpg.de/.

Lukas Mehl, Jenny Schmalfuss, Azin Jahedi, Yaroslava Nalivayko, and Andrés Bruhn. Spring: A high-resolution high-detail dataset and benchmark for scene flow, optical flow and stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. URL https://arxiv.org/abs/2303.01943.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=a68SUt6zFt.

Jack Parker-Holder and Shlomi Fruchter. Genie 3: A new frontier for world models. Google DeepMind, 2025. URL https://deepmind.google/blog/genie-3-a-new-frontier-for-world-models/. Official research announcement, August 5, 2025.

Xuanchi Ren, Tianchang Shen, Jiahui Huang, Huan Ling, Yifan Lu, Merlin Nimier-David, Thomas Müller, Alexander Keller, Sanja Fidler, and Jun Gao. GEN3C: 3D-informed world-consistent video generation with precise camera control. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6121–6132, 2025. doi: 10.1109/CVPR52734.2025.00574. URL https://openaccess.thecvf.com/content/CVPR2025/html/Ren_GEN3C_3D-Informed_World-Consistent_Video_Generation_with_Precise_Camera_Control_CVPR_2025_paper.html.

Robbyant Team, Zelin Gao, Qiuyu Wang, Yanhong Zeng, Jiapeng Zhu, Ka Leong Cheng, Yixuan Li, Hanlin Wang, Yinghao Xu, Shuailei Ma, Yihang Chen, Jie Liu, Yansong Cheng, Yao Yao, Jiayi Zhu, Yihao Meng, Kecheng Zheng, Qingyan Bai, Jingye Chen, Zehong Shen, Yue Yu, Xing Zhu, Yujun Shen, and Hao Ouyang. Advancing open-source world models. *arXiv preprint arXiv:2601.20540*, 2026. URL https://arxiv.org/abs/2601.20540.

Lloyd Russell, Anthony Hu, Lorenzo Bertoni, George Fedoseev, Jamie Shotton, Elahe Arani, and Gianluca Corrado. GAIA-2: A controllable multi-view generative world model for autonomous driving. *arXiv preprint arXiv:2503.20523*, 2025. URL https://arxiv.org/abs/2503.20523.

Shijun Shi, Jing Xu, Zhihang Li, Chunli Peng, Xiaoda Yang, Lijing Lu, Kai Hu, and Jiangning Zhang. One-to-All Animation: Alignment-free character animation and image pose transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4011–4021, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/papers/Shi_One-to-All_Animation_Alignment-Free_Character_Animation_and_Image_Pose_Transfer_CVPR_2026_paper.pdf.

Joonghyuk Shin, Zhengqi Li, Richard Zhang, Jun-Yan Zhu, Jaesik Park, Eli Shechtman, and Xun Huang. MotionStream: Real-time video generation with interactive motion controls. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=v1DKz5Vxr7.

Wenqiang Sun, Haiyu Zhang, Haoyuan Wang, Junta Wu, Zehan Wang, Zhenwei Wang, Yunhong Wang, Jun Zhang, Tengfei Wang, and Chunchao Guo. Worldplay: Towards long-term geometric consistency for real-time interactive world modeling. *arXiv preprint arXiv:2512.14614*, 2025. URL https://arxiv.org/abs/2512.14614. Version 2 revised June 9, 2026.

Zhiyao Sun, Ziqiao Peng, Yifeng Ma, Yi Chen, Zhengguang Zhou, Zixiang Zhou, Guozhen Zhang, Youliang Zhang, Yuan Zhou, Qinglin Lu, and Yong-Jin Liu. StreamAvatar: Streaming diffusion models for real-time interactive human avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10887–10897, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Sun_StreamAvatar_Streaming_Diffusion_Models_for_Real-Time_Interactive_Human_Avatars_CVPR_2026_paper.html.

Team Wan et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. URL https://arxiv.org/abs/2503.20314.

Dani Valevski, Yaniv Leviathan, Moab Arar, and Shlomi Fruchter. Diffusion models are real-time game engines. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2408.14837.

Timo von Marcard, Roberto Henschel, Michael J. Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3D human pose in the wild using IMUs and a moving camera. In *European Conference on Computer Vision*, 2018. URL https://openaccess.thecvf.com/content_ECCV_2018/html/Timo_von_Marcard_Recovering_Accurate_3D_ECCV_2018_paper.html.

Guangyuan Wang, Li Hu, Dechao Meng, Zhongyi Zhang, Peng Zhang, Xindi Zhang, Mingyang Huang, Ruoshi Zhang, Ke Sun, Zhe Zhang, Xingjun Wang, Gang Cheng, Hai Xu, and Bang Zhang. Wan-Animate-2: Pushing the application boundaries of character animation. *arXiv preprint arXiv:2608.06009*, 2026a. URL https://arxiv.org/abs/2608.06009.

Haoyuan Wang, Yabo Chen, Haibin Huang, Chi Zhang, and Xuelong Li. Directing the world: Fast autoregressive video generation with compositional human-camera control. *arXiv preprint arXiv:2606.27964*, 2026b. URL https://arxiv.org/abs/2606.27964.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual geometry grounded transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5294–5306, 2025a. doi: 10.1109/CVPR52734.2025.00499. URL https://openaccess.thecvf.com/content/CVPR2025/html/Wang_VGGT_Visual_Geometry_Grounded_Transformer_CVPR_2025_paper.html.

Wenshan Wang, Delong Zhu, Xiangwei Wang, Yaoyu Hu, Yuheng Qiu, Chen Wang, Yafei Hu, Ashish Kapoor, and Sebastian Scherer. TartanAir: A dataset to push the limits of visual SLAM. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2020. URL https://arxiv.org/abs/2003.14338.

Xiang Wang, Shiwei Zhang, Changxin Gao, Jiayu Wang, Xiaoqiang Zhou, Yingya Zhang, Luxin Yan, and Nong Sang. UniAnimate: Taming unified video diffusion models for consistent human image animation. *Science China Information Sciences*, 68:200103, 2025b. doi: 10.1007/s11432-024-4592-3. URL https://doi.org/10.1007/s11432-024-4592-3.

Xiang Wang, Shiwei Zhang, Longxiang Tang, Yingya Zhang, Changxin Gao, Yuehuan Wang, and Nong Sang. UniAnimate-DiT: Human image animation with large-scale video diffusion transformer. *arXiv preprint arXiv:2504.11289*, 2025c. URL https://arxiv.org/abs/2504.11289.

Zhenzhi Wang, Yixuan Li, Yanhong Zeng, Youqing Fang, Yuwei Guo, Wenran Liu, Jing Tan, Kai Chen, Tianfan Xue, Bo Dai, and Dahua Lin. HumanVid: Demystifying training data for camera-controllable human image animation. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2024a. doi: 10.52202/079017-0635. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/23f3a0f82d79d985b6076bc84d14f66b-Abstract-Datasets_and_Benchmarks_Track.html.

Zhouxia Wang, Ziyang Yuan, Xintao Wang, Yaowei Li, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. MotionCtrl: A unified and flexible motion controller for video generation. In *ACM SIGGRAPH 2024 Conference Papers*, pp. 1–11, 2024b. doi: 10.1145/3641519.3657518. URL https://doi.org/10.1145/3641519.3657518.

Zhongcong Xu, Jianfeng Zhang, Jun Hao Liew, Hanshu Yan, Jia-Wei Liu, Chenxu Zhang, Jiashi Feng, and Mike Zheng Shou. MagicAnimate: Temporally consistent human image animation using diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1481–1490, 2024. doi: 10.1109/CVPR52733.2024.00147. URL https://openaccess.thecvf.com/content/CVPR2024/html/Xu_MagicAnimate_Temporally_Consistent_Human_Image_Animation_using_Diffusion_Model_CVPR_2024_paper.html.

Seonghyeon Ye, Yunhao Ge, Kaiyuan Zheng, Shenyuan Gao, Sihyun Yu, George Kurian, et al. World action models are zero-shot policies. *arXiv preprint arXiv:2602.15922*, 2026. URL https://arxiv.org/abs/2602.15922.

Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. ScanNet++: A high-fidelity dataset of 3D indoor scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. URL https://arxiv.org/abs/2308.11417.

Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Fredo Durand, William T. Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. URL https://arxiv.org/abs/2311.18828.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on*

Computer Vision and Pattern Recognition, 2018. URL https://arxiv.org/abs/1801.03924.

Siwei Zhang, Qianli Ma, Yan Zhang, Zhiyin Qian, Taein Kwon, Marc Pollefeys, Federica Bogo, and Siyu Tang. EgoBody: Human body shape and motion of interacting people from head-mounted devices. In *European Conference on Computer Vision*, 2022. URL https://arxiv.org/abs/2112.07642.

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 *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 74896–74910, 2025a. URL https://proceedings.mlr.press/v267/zhang25v.html.

Yuhong Zhang, Jing Lin, Ailing Zeng, Guanlin Wu, Shunlin Lu, Yurong Fu, Yuanhao Cai, Ruimao Zhang, Haoqian Wang, and Lei Zhang. Motion-X++: A large-scale multimodal 3D whole-body human motion dataset. *arXiv preprint arXiv:2501.05098*, 2025b. URL https://arxiv.org/abs/2501.05098.

Yuxuan Zhang, Haozhong Xiong, Yubo Huang, Jiayi Song, Jinpeng Yu, Haofan Wang, Jiaming Liu, Ruihua Huang, and Liwei Wang. LiveAnimate: Stable long-form streaming human animation in real-time. *arXiv preprint arXiv:2608.11745*, 2026. URL https://arxiv.org/abs/2608.11745.

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

Dingcheng Zhen, Xu Zheng, Ruixin Zhang, Zhiqi Jiang, Yichao Yan, Ming Tao, and Shunshun Yin. SoulX-LiveAct: Towards hour-scale real-time human animation with neighbor forcing and ConvKV memory. *arXiv preprint arXiv:2603.11746*, 2026. URL https://arxiv.org/abs/2603.11746.

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. *arXiv preprint arXiv:2504.14977*, 2025. URL https://arxiv.org/abs/2504.14977.

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 *European Conference on Computer Vision*, pp. 145–162, 2024. doi: 10.1007/978-3-031-73001-6_9. URL https://www.ecva.net/papers/eccv_2024/papers_ECCV/papers/07192.pdf.

A PROTOCOL AND EVALUATION DETAILS

This appendix describes the training implementation, evaluation datasets and metrics, inference optimizations, and additional ablation results.

A.1 CONDITIONING IMPLEMENTATION

We use LoRA rank 64 (Hu et al., 2022), three-latent-frame target blocks, and 480×480 resolution. Stage III initializes the generator from Stage II and uses the Stage-I model as the frozen real-score teacher. Quality evaluations use four-step autoregressive sampling and 50-step bidirectional sampling. Temporal analyses use the source-frame timestamps recorded for each sequence.

The model conditions on body normals, a reference image, scene projections with validity masks, and Plücker camera rays. PCD and ray signals enter separate multilayer residual adapters, both active in the accelerated pipeline. Temporal alignment follows the causal VAE convention: retrieved latents, geometric conditions, and rotary indices share their original global timestamps. The runtime study varies caching and sampling while keeping model weights, resolution, and the seven-latent attention budget fixed.

A.2 TRAINING AND ARCHITECTURE DETAILS

Stage I retains the reference-image anchor and native body-normal pathway of Wan2.1-Fun Control. PCD and ray branches are independently dropped, and projected coverage is degraded during training, discouraging exclusive dependence on either source. The point cloud remains fixed; the video prior synthesizes regions without projected evidence.

Stage II uses clean teacher-forced histories and three-latent blocks. Reference and selected ground-truth historical tokens receive timestep zero. Only the target block is corrupted as $z_\tau^{(k)} = (1 - \tau) z_{\mathrm{GT}}^{(k)} + \tau\epsilon$, and the flow-matching loss is computed only on its prediction against $\epsilon - z_{\mathrm{GT}}^{(k)}$. Reference tokens attend only to the reference. A selected historical block reads the reference, itself, and earlier selected blocks, while the noisy current block reads the reference, selected history, and itself. Tokens within a block may attend to one another; block identities use original temporal indices, preserving causality across blocks even after history selection.

PAHM descriptors and retrieval. Each latent-aligned control normal is cropped to its occupied body region. Masked normals and the occupancy mask form four channels, adaptively average-pooled to 8×8, flattened, and ℓ_2-normalized. The current-block query averages its frame descriptors; each candidate key averages two consecutive historical descriptors. Cosine similarity selects the highest-scoring pair. Candidates exclude latent index zero and end before the five-frame recent window with a two-latent exclusion gap. If seven historical latents are unavailable, descriptors are invalid, or the best similarity is below 0.25, selection falls back to up to seven recent latents. Descriptors are computed from the body controls. Selected latents, normals, PCD projections, rays, rotary indices, and timestamps are gathered together. Stage-II training and the window-recomputed quality benchmark search the available historical prefix; the accelerated inference path restricts candidates to the bank described in Appendix B.

Stage-III sampling and gradient replay. We sample $T \sim \mathcal{U}\{1, \ldots, K\}$ and $j \sim \mathcal{U}\{1, \ldots, 4\}$ independently, sharing exit j across all blocks in a rollout. Each block starts from Gaussian noise. At noise level σ_i, the student predicts $\hat{z}_0 = x_i - \sigma_i v_\theta(x_i, \sigma_i, H_k, U_k)$; for $i < j$, the next input is $(1 - \sigma_{i+1})\hat{z}_0 + \sigma_{i+1}\epsilon_i$ with fresh Gaussian noise. Exit predictions become the generated history for subsequent blocks. After the detached sequence is complete, only block T's exit forward is replayed with gradients, preserving its noisy input, selected history, and conditions. Earlier denoising steps and inter-block dependencies remain detached. For block lengths n_k, Equation 3 uses $w_T = K n_T / \sum_k n_k$, making uniform block sampling an unbiased estimator of the full-sequence, element-averaged *truncated* gradient for a fixed rollout. Training uses seven equal three-latent blocks, so $w_T = 1$.

DMD implementation. The frozen Stage-I real-score model and learned fake-score model operate bidirectionally on the complete noised sequence. Writing D_{real} and D_{fake} for their clean

predictions, we compute $r = \max(\text{mean}\,|\,\text{sg}(y) - D_{\text{real}}|, 10^{-4})$ and $g = (D_{\text{fake}} - D_{\text{real}})/r$, then set $a_\tau = r^{-1}\min(1, 10/\max(\text{RMS}(g), 10^{-8}))$. Statistics are computed per sample over all latent elements and detached. The fake-score flow model is trained on fresh detached rollouts from the same sampled-exit distribution, with target $\epsilon - \text{sg}(y)$ under the same conditioning. The generator uses only the DMD surrogate, with its weight ramped up over the first 100 generator updates; no paired reconstruction loss is added. Reported quality uses all four denoising steps at inference.

A.3 Datasets and split construction

Evaluation draws on held-out sequences from adaptation domains, datasets excluded from adaptation, and newly captured PoseLoop data. Each track groups sequences by the capability being evaluated, with the overlaps specified below. Membership and controls are fixed from the inputs. The main paper's Human Motion, Camera Motion, PoseLoop, and Long Sequence tracks correspond to Human198, Camera90, PoseLoop60, and Long48 in the evaluation records; Long-H denotes the human subset of Long Sequence.

Training data. BodyWorld is adapted on 92,503 eligible 81-frame samples from 14 dataset families (19 source labels). The human-centric sources are 3DPW, BEDLAM, EgoBody, EMDB, GTA-Human, Harmony4D, HumanVid, HuMMan, and SLOPER4D (von Marcard et al., 2018; Black et al., 2023; Zhang et al., 2022; Kaufmann et al., 2023; Cai et al., 2021; Khirodkar et al., 2024; Wang et al., 2024a; Cai et al., 2022; Dai et al., 2023). Motion-X++ contributes six named subcollections, including IDEA400 and HAA500 (Lin et al., 2023; Zhang et al., 2025b; Chung et al., 2021). The scene-only sources are ScanNet++, ScanNet v2, Spring, and TartanAir (Yeshwanth et al., 2023; Dai et al., 2017; Mehl et al., 2023; Wang et al., 2020). The retained sample counts after length and conditioning checks are listed below.

Dataset or source	Samples	Depth / camera source
3DPW	81	LingBot-Map
BEDLAM2 (BEDLAM)	33,176	LingBot-Map
EgoBody	125	LingBot-Map
EMDB	81	LingBot-Map
GTA-Human	2,075	LingBot-Map
Harmony4D	2,214	LingBot-Map
HumanVid	24,232	LingBot-Map
HuMMan	111	LingBot-Map
Motion-X++: Animation 201; HAA500 874; IDEA400 11,301; KungFu 838; Music 2,224; Perform 839	16,277	LingBot-Map
ScanNet++	6,266	native RGB-D/camera
ScanNet v2	6,990	native RGB-D/camera
SLOPER4D	6	LingBot-Map
Spring	18	native depth/camera
TartanAir	851	native depth/camera
Total	92,503	

The selected ScanNet++, ScanNet v2, Spring, and TartanAir clips are scene-only: they contain no controlled person and are included to learn camera motion, scene reprojection, and background changes independently of body control. Together they contribute 14,125 samples; the other 78,378 samples are human-centric. Balanced source sampling expands the iterable training index to 100,079 entries, including repeats introduced only by balancing.

Reference-only scene prior and geometry preprocessing. The four scene-only sources provide depth or RGB-D observations together with camera calibration and poses. Most human-centric sources do not provide dense scene geometry and camera trajectories in a shared convention, so we run LingBot-Map (Chen et al., 2026) on each RGB sequence to estimate per-frame depth, intrinsics, and world-to-camera poses. Although preprocessing retains these per-frame estimates, the scene prior for every training and evaluation sample is constructed only from the reference frame: we back-project its RGB image I_0 with its depth D_0 and intrinsics, remove points in the controlled-person region, and express the result as $\mathcal{P}_0$ in the reference-camera coordinates. Hence point colors

come only from I_0, and the geometric support of $\mathcal{P}_0$ contains only surfaces visible in the reference view. For a requested camera pose C_t, we project this fixed $\mathcal{P}_0$ to obtain the PCD condition and construct the ray map from the requested camera parameters. Depth maps at $t > 0$ are never fused into $\mathcal{P}_0$ or otherwise used by the model, and later RGB frames do not contribute color or geometry to this scene prior. During training, later RGB frames provide video supervision, including the clean histories used by Stage-II teacher forcing. During evaluation, the source sequence supplies the body and camera control trajectories, while its later RGB frames are used only as targets for scoring after generation. Thus the reported generation protocol uses one reference image and its reference-view depth for scene initialization; the remaining frames do not enlarge the visible scene prior.

Human Motion and the near-static subset. Before constructing the training manifest, we reserve complete raw sequences from the adaptation datasets for in-domain evaluation; no window from a reserved sequence is used for training. Human198 contains 91 such sequence-disjoint in-domain cases: 3DPW (10), BEDLAM2 (17), EgoBody (11), EMDB (11), GTA-Human (10), HumanVid (16), HuMMan (14), and SLOPER4D (2). Its 107 out-of-domain cases come from four datasets entirely excluded from adaptation: AIST++ (48), RICH (32), MoviCam (24), and NeuMan (3) (Li et al., 2021; Huang et al., 2022; Aytekin et al., 2025; Jiang et al., 2022). RICH's 32 windows share one raw sequence. Input eligibility requires a controlled person in the reference and valid native conditions for all compared methods; these rules exclude three inputs without a reference person and two without the face conditions required by a baseline. All methods are evaluated on the resulting common set.

The near-static subset requires a rotation of at most 3° and a camera-center displacement of at most 0.03, normalized by median reference depth, relative to the reference pose throughout the sequence. These tolerances accommodate camera-estimation jitter. Selection uses input extrinsics and yields 115 cases (34 ID, 81 OOD), shared by all methods; the other 83 cases contain larger camera motion.

Camera Motion. Camera90 contains 54 human-centric and 36 scene-only cases with explicit camera trajectories. Human cases come from 3DPW, BEDLAM2, EgoBody, EMDB, and HumanVid; the 36 scene cases comprise ScanNet++ (11), ScanNet v2 (10), Spring (4), and TartanAir (11). Camera90 and Human198 share 21 ID inputs: 10 from 3DPW, two from EgoBody, six from EMDB, and three from HumanVid. Results are aggregated within each track. Body DINO and PCK apply to the human subset.

PoseLoop and Long Sequence. PoseLoop60 is our newly captured, training-excluded collection for pose recurrence and long-rollout evaluation. It contains 60 clips: 50 of 81 frames, five of 121 frames, and five of 201 frames. We first average each metric over the scored frames of each clip, then average the 60 clip scores with equal weight. All variants use this aggregation on unrounded scores. Table 7 additionally reports each length group, which contains a distinct set of clips. Temporal degradation within the same sequences is assessed on Long48.

Long48 contains 24 human clips (12 BEDLAM2 and 12 EgoBody) and 24 scene clips (11 ScanNetV2 and 13 TartanAir), each with 201 stored frames. Spatial scoring excludes reference frame 0: full-sequence scores use frames 1–200, whereas the main table's late window uses frames 121–200 of the human subset only. Early and middle windows are 1–80 and 81–120. Table 8 additionally reports full-sequence human and scene results. Temporal metrics include the reference-to-first-target transition.

Capability flags and input parity. Table 1 summarizes native control capabilities and continuation modes. Wan-Move uses surface-point trajectories for human motion and supports camera-induced tracks from projected scene geometry (Chu et al., 2025); Camera90 evaluates the explicit camera interfaces of Uni3C and BodyWorld. Uni3C composes scene and human geometry under camera control (Cao et al., 2025). The $\checkmark^s$ entries for One-to-All and Wan-Animate indicate continuation using preceding-segment frames (Shi et al., 2026; Cheng et al., 2025); UniAnimate-DiT uses overlapping windows (Wang et al., 2025c). BodyWorld performs causal continuation at the latent-block level. Camera flags indicate an explicit viewpoint-control interface.

Input alignment and provenance. Input records retain case and raw-sequence identities, timestamps, control and scene-prior provenance, camera conventions, and content hashes. Body and cam-

era trajectories share a coordinate system. Comparisons use common reference images and target timestamps, with each method receiving its native controls. Wan-Animate's additional driving-face RGB is marked in the main table.

A.4 NATIVE-CONTROL QUALITATIVE COMPARISONS

Figures 5–7 show five synchronized timestamps for each selected case, pairing each method's native control with its output. The controls are point tracks for Wan-Move, combined SMPL/scene renders for Uni3C, pose renders for One-to-All and UniAnimate-DiT, pose and driving-face RGB for Wan2.2-Animate, and body normals for BodyWorld. Reference images and text prompts are shared. For compactness, the panels visualize BodyWorld's body control; its PCD projections and camera rays are also supplied during generation.

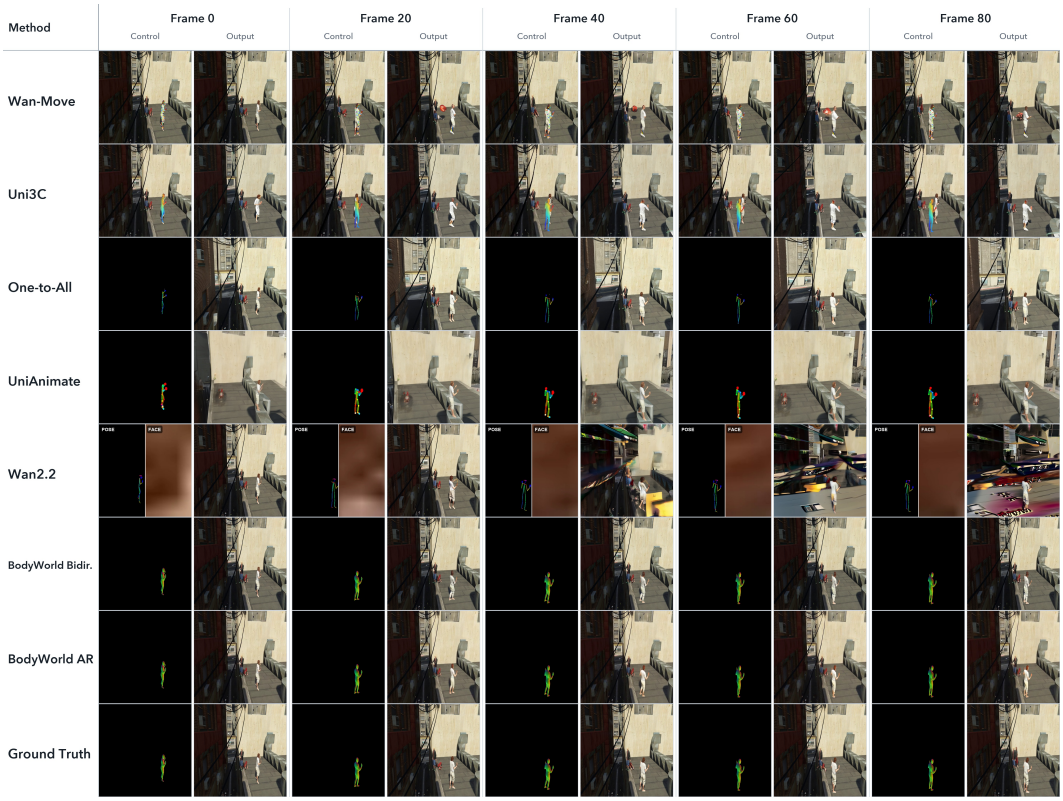

Figure 5: **Native controls and outputs: complex urban alley with a small subject.** Each method occupies one row; Control and Output are paired side by side at Frames 0, 20, 40, 60, and 80.

A.5 METRIC DEFINITIONS

Paired image metrics. We compute image metrics on lossless decoded frames at 480×480. PSNR uses RGB values in $[0, 1]$, and LPIPS uses the AlexNet backbone. Body DINO is the cosine similarity of DINOv2 ViT-L/14 features on matched human crops. Crop boxes come from the target body masks, are padded by 10% of their longest side, and are shared across methods. Spatial metrics exclude reference frame zero. Scores are averaged within clips, within sources, and then equally across sources.

For body PSNR, squared RGB errors are summed within the fixed target body mask and divided by the number of valid scalar pixels. Empty regions are excluded from regional averages. Temporal LPIPS is the mean absolute difference between generated and target adjacent-frame LPIPS distances, including the reference-to-first-target transition. The Stage-I metric Δ_t similarly measures the mean absolute discrepancy between adjacent-frame RGB changes. These metrics assess temporal agreement with the target motion.

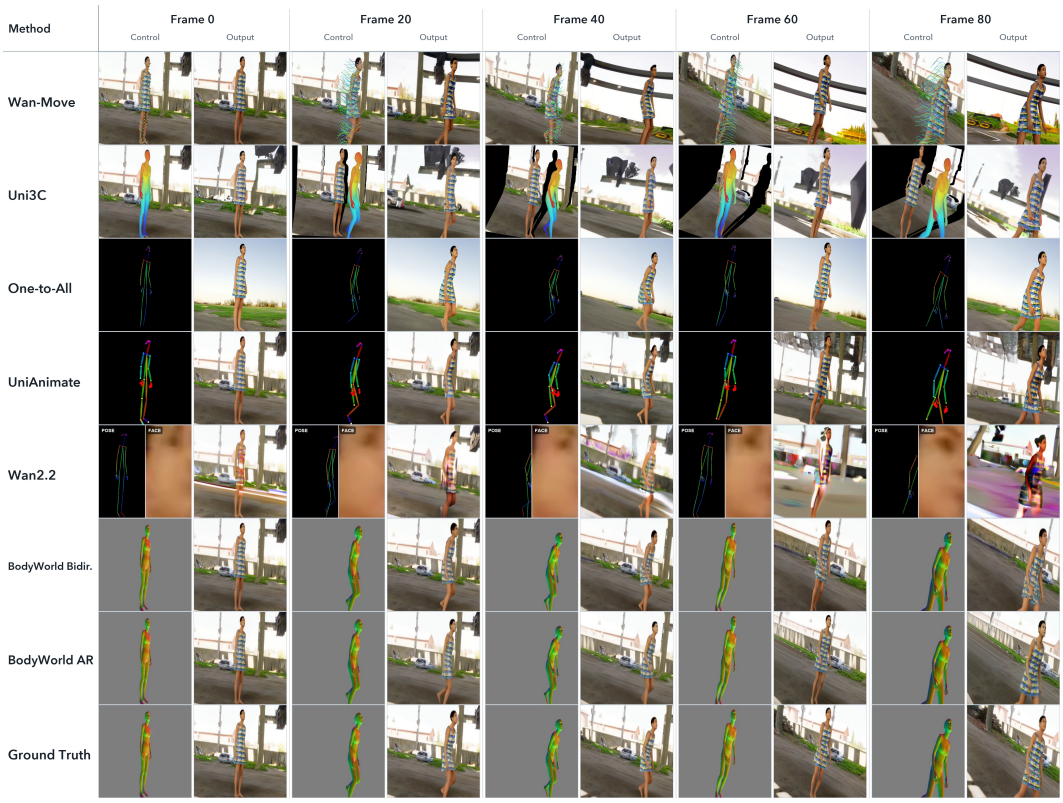

Figure 6: **Native controls and outputs: industrial street with large camera rotation.** Layout and timestamps follow Figure 5.

Pose following. PCK measures agreement between Keypoint R-CNN detections in generated and target RGB (He et al., 2017). We associate detections with the target body box and evaluate the 17 COCO joints. Eligible target joints lie inside the image and have keypoint scores of at least 2. A predicted joint is correct when its score is at least 2 and its error is within 0.1 of the padded target body-box diagonal; missing predictions count as misses. The main benchmark pools hits over eligible joints within each clip before source-macro aggregation. PCK is therefore a detector-based 2D pose proxy.

Camera following. VGGT estimates camera poses from uniformly sampled full-frame generated RGB. One clip-level Sim(3) transform aligns the estimated camera centers to the requested trajectory. ATE/path is the root-mean-square aligned position error divided by the requested path length, computed at the same sampled timestamps. RRE is the mean angular discrepancy, in degrees, between estimated and requested relative rotations at consecutive sampled timestamps. Both metrics are estimator-based proxies for camera following. The near-static subset is evaluated separately for image and human-motion quality.

Appearance and recurrence. PoseLoop return pairs are selected from the input controls and fixed across methods. We resize control frames to 32×32, mean-center and normalize their flattened descriptors, and score a pair by 0.8 times descriptor cosine similarity plus 0.2 times body-mask IoU. The later endpoint lies in the second half of the clip, with a minimum gap of 16 frames and a score threshold of 0.6. We retain up to five high-scoring pairs per clip, suppressing nearby duplicate pairs, yielding 222 pairs across PoseLoop60.

For each pair, a shared crop encloses the union of the two target body masks with 10% padding. DINO RetErr is one minus the cosine similarity between generated endpoint features; LPIPS RetErr measures perceptual distance between the generated endpoints. Pair PSNR uses the union foreground mask. Pose Δ error measures the discrepancy between generated and target keypoint dis-

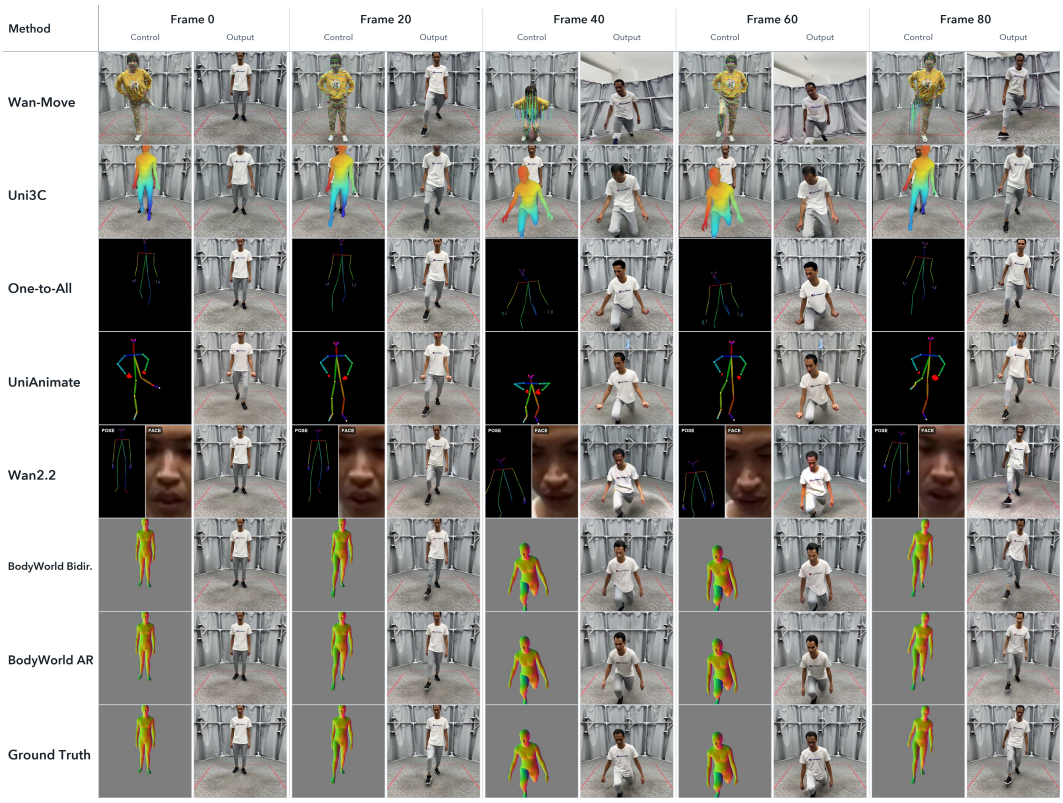

Figure 7: **Native controls and outputs: forward squat with close-up scale change.** Layout and timestamps follow Figure 5.

placements, normalized by the mean body-box diagonal of the two endpoints. These appearance and pose measures jointly characterize consistency when the controlled body returns to a similar configuration.

Temporal windows and paired uncertainty. Long48 temporal comparisons use the same early (frames 1–80), middle (81–120), and late (121–200) windows for all variants. Paired confidence intervals resample raw-sequence groups, preserving the pairing between methods and grouping windows from the same source sequence. Stage-I comparisons use 10,000 bootstrap draws; same-weight PAHM comparisons use 5,000. We report percentile 95% intervals. PoseLoop recurrence tables average eligible pairs, while paired inference first averages within clips and raw-sequence groups.

B Inference runtime details

We measure inference on one NVIDIA H200 at 480×480 and batch size one. Table 4 reports the three- and two-step profiles of the evaluated Stage-III model. Generation timing covers adapters, denoising, scheduling, and cache updates with pre-encoded controls, as defined in Section 4.5.

Table 4: **Single-H200 generation throughput.** Mean over seven post-warm-up clips; peak memory covers all eight. Timings exclude input encoding, VAE decoding, and video I/O.

Rolling-cache profile	Steps/block	Time (s/81 frames)	Gen. FPS	Peak (GiB)
Standard	3	8.223	9.85	54.84
Faster	2	6.061	13.36	54.84

Exact reuse versus rolling inference. For a fixed history window, clean historical tokens attend only to the reference and selected history, so their layer-wise states can be reused exactly across denoising steps. These states are rebuilt when the selected window changes. In eight paired tests on a development model, this yields a mean per-case speedup of $1.727\times$ with bitwise-identical output latents and history selections. Cross-block rolling inference further reuses historical states computed in their insertion contexts and retrieves them from a bounded bank. Table 5 separates these two cache configurations.

Dynamic pose-pair bank. The accelerated path stores at most five anchor entries, each containing two consecutive generated latents, their normalized mean pose descriptor, and original frame indices. After each committed block, eligible pairs outside the recent window and exclusion gap are considered for insertion. We reject invalid descriptors, pairs whose start indices are less than two latent frames from an existing anchor, and pairs whose minimum cosine distance to the bank is below 0.08. When the bank is full, a replacement is accepted only if it increases the minimum pairwise descriptor distance, encouraging pose coverage. Retrieval uses the cosine query and 0.25 threshold defined in Appendix A.2. A short staging buffer supports pair formation and Recent-7 fallback; older unselected latent payloads are pruned. This gives retrieval storage a fixed capacity. Table 4 reports overall peak GPU memory. The quality benchmark uses prefix search, and the accelerated evaluation uses this bounded bank.

Table 5: **Four-step cache comparisons.** The upper pair uses a development model on eight matched clips; the lower group uses the evaluated model on one matched clip. Each group shares model weights, inputs, and denoising-step count.

Model / protocol	Implementation	Time (s/81f)	Gen. FPS	Peak (GiB)
Earlier / eight paired clips	Window recomputation	35.792	2.263	41.30
Earlier / eight paired clips	Exact within-block cache	20.729	3.908	52.15
Evaluated / single clip	Window recomputation	35.662	2.271	40.15
Evaluated / single clip	Initial rolling cache	16.731	4.841	48.85
Evaluated / single clip	Optimized rolling cache	10.823	7.484	54.57

At four denoising steps, optimized rolling inference increases generation rate from 2.271 to 7.484 FPS in the matched single-clip comparison, a $3.29\times$ speedup. Runtime optimizations beyond the initial rolling cache raise throughput from 4.841 to 7.484 FPS. Table 4 separately reports the multi-clip throughput of the final three- and two-step profiles.

Runtime configuration. The runtime combines the bounded bank above with separately cached reference KV. Historical keys are stored after RoPE using stable global frame IDs, checked on retrieval; eviction removes KV no longer required by the bank and recent/fallback buffers. One contiguous KV copy per layer and block provides per-frame views, and gathered history is reused across that block's denoising steps. Three- and two-step schedules use scheduler indices $(1000, 500, 250)$ and $(1000, 250)$, respectively. For seven generated blocks, they require 21 or 14 denoising forwards plus six clean-cache updates: the terminal block is never reused. Clean updates skip the denoiser head and the final Transformer layer's cross-attention and feed-forward tail, whose outputs do not contribute to cached KV.

Conditioning and kernel optimizations. The backbone's text projection and layer-wise text/reference cross-attention KV are computed on first use and reused for the same prompt and reference. The timing benchmark pre-encodes full-clip controls and batches the temporally pointwise PCD/ray adapters, gathering their outputs by original frame IDs. Active LoRA deltas are folded into BF16 base weights, empty adapter wrappers are removed, and the frozen control adapters also use BF16. After reference initialization, only current-block tokens act as queries. Since every supplied key is visible to these queries, the all-true attention mask is omitted for native SDPA; reference initialization retains its causal mask. All timings use inference mode with this combined configuration.

Paired quality check. We compare the accelerated profiles against four-step window recomputation on the eight runtime-development clips, including the timing warm-up, with identical model

weights, inputs, and seeds. Table 6 reports clip-mean differences from lossless decoded frames, with body metrics evaluated using control-normal masks. On these clips, both profiles maintain body PSNR and SSIM, while whole-frame PSNR changes by less than 0.2 dB.

Table 6: **Quality checks for accelerated inference.** Changes relative to four-step window recomputation on eight development clips. Higher PSNR and SSIM are better.

Profile	Body PSNR (dB)	Body SSIM	Whole PSNR (dB)	Whole SSIM
Three-step rolling	$+0.004$	$+0.00473$	-0.160	$+0.00002$
Two-step rolling	$+0.150$	$+0.01432$	-0.049	$+0.00731$

C Additional ablation results

We provide the conditioning protocol, long-sequence and recurrence results, and qualitative comparisons supporting the main ablation table. ID-H, Moving-H, and Scene-C denote in-domain human, moving-camera human, and scene-only camera subsets, containing 96, 44, and 36 cases, respectively.

C.1 Stage-1 conditioning protocol and uncertainty

The Stage-I comparison uses matched training budgets, 81-frame sampling at 480×480, 50 denoising steps, and guidance 1.0. All four variants share the same 168 unique inputs and sampling seed. ID-H96 and the 60-case human-camera set share 24 inputs; results are averaged within each subset. Moving-H44 is selected by input camera trajectories. PSNR and adjacent-change errors are measured before video encoding, with a method-independent body mask for regional scores.

Stage-I conditioning scores average all 81 stored frames, including the reference frame, consistently across the four variants. The main Human Motion and Camera Motion benchmarks exclude the reference frame. This conditioning comparison evaluates PCD-only, ray-only, early joint fusion, and decoupled PCD/ray injection under the same body controls.

The paired comparison with PCD-only clarifies the effect of adding viewpoint information: decoupled conditioning improves moving-human temporal consistency while maintaining similar body PSNR, with a modest trade-off in whole-frame reconstruction. Supplementary Figure 8 shows the corresponding improvement in viewpoint following and scene structure.

Paired uncertainty. Using 10,000 raw-sequence bootstrap resamples and orienting effects so positive favors decoupled conditioning, the ID-H96 and Moving-H44 body-PSNR gains are $+0.0437$ dB (95% CI $[-0.0054, 0.0949]$) and $+0.0506$ dB ($[-0.0346, 0.1417]$). The Moving-H44 adjacent-change improvement is $+0.000802$ ($[0.000093, 0.001620]$). Whole-frame PSNR favors PCD-only: effects are -0.1705 dB ($[-0.2843, -0.0677]$) on ID-H96 and -0.1588 dB ($[-0.3212, -0.0037]$) on Scene-C36.

C.2 Stage-3 long-horizon and history-selection results

All Stage-III results use four-step sampling, generated history, lossless-frame evaluation, and source-macro aggregation on Human198, PoseLoop60, and Long48. Spatial metrics exclude the reference frame; temporal LPIPS includes the reference-to-first-target transition.

Training variants and inference-time interventions. The four training variants share matched training budgets: PAHM + self forcing + DMD, Recent-7 + self forcing + DMD, PAHM + self forcing without DMD, and PAHM + DMD without self forcing. Each generator initializes from Stage II; DMD variants use the frozen Stage-I real-score teacher. Trained Recent-7 uses recent-only history during both training and inference. In the same-weight comparison, all settings use the full PAHM model: Full 2+5 combines two pose-retrieved and five recent latents, Recent-5 drops the retrieved pair, and Recent-7 fills all seven slots with recent latents. Full versus Recent-7 isolates history selection at equal capacity; Recent-5 tests the contribution of the retrieved pair with the five recent slots fixed.

Table 7: **PoseLoop60 across sequence lengths.** The 81-frame split has 50 cases; the 121- and 201-frame splits each have five. Values are source-macro means for each length group.

Length	Variant	PSNR↑	LPIPS↓	Body DINO↑	PCK↑	Temp.↓
81f	PAHM + SF + DMD	19.0011	0.2257	0.7831	0.9741	0.0109
	Recent-7 + SF + DMD	19.0658	0.2243	0.7788	0.9746	0.0108
	PAHM + SF, w/o DMD	19.0607	0.2428	0.7437	0.9483	0.0130
	PAHM + DMD, w/o SF	19.0621	0.2274	0.7716	0.9698	0.0112
121f	PAHM + SF + DMD	18.7304	0.2619	0.7156	0.9824	0.0115
	Recent-7 + SF + DMD	18.7709	0.2621	0.7175	0.9849	0.0114
	PAHM + SF, w/o DMD	18.9066	0.2816	0.6650	0.9681	0.0151
	PAHM + DMD, w/o SF	19.3166	0.2444	0.6959	0.9791	0.0118
201f	PAHM + SF + DMD	16.9257	0.3827	0.6151	0.9649	0.0136
	Recent-7 + SF + DMD	16.6825	0.3868	0.6083	0.9566	0.0126
	PAHM + SF, w/o DMD	17.3464	0.4139	0.5611	0.9604	0.0208
	PAHM + DMD, w/o SF	17.2500	0.4017	0.5500	0.9442	0.0132

Table 8: **Long48 full 201-frame rollouts.** Human and scene subsets each contain 24 cases. Body DINO and PCK apply only to human-centric clips.

Variant	PSNR↑	LPIPS↓	Body DINO↑	PCK↑	Temp.↓
(a) Long48 human, full 201 frames					
PAHM + SF + DMD	17.4233	0.3733	0.6845	0.9012	0.0188
Recent-7 + SF + DMD	17.3537	0.3779	0.6660	0.8990	0.0192
PAHM + SF, w/o DMD	18.0359	0.4297	0.6258	0.8163	0.0309
PAHM + DMD, w/o SF	16.8645	0.4288	0.6382	0.8509	0.0187
(b) Long48 scene-only, full 201 frames; human metrics not applicable					
PAHM + SF + DMD	12.5409	0.5975	N/A	N/A	0.0633
Recent-7 + SF + DMD	12.3674	0.6005	N/A	N/A	0.0607
PAHM + SF, w/o DMD	13.2725	0.7221	N/A	N/A	0.0857
PAHM + DMD, w/o SF	12.2124	0.6583	N/A	N/A	0.0680

On Long48-human, the complete model reduces early-to-late degradation relative to the separately trained Recent-7 variant: PSNR drops by 3.0810 versus 3.3429 dB, LPIPS increases by 0.1537 versus 0.1716, Body DINO drops by 0.1513 versus 0.1805, and PCK drops by 0.0241 versus 0.0306. This pattern is consistent with pose retrieval preserving useful appearance evidence as recent views change. On Long48-scene, the complete model's early-to-late changes are -4.8581 dB PSNR, $+0.2568$ LPIPS, and $+0.0313$ temporal LPIPS. Table 8 reports the full-sequence results for both subsets.

PAHM retrieval activates in all 60 PoseLoop cases and all 24 Long48-human cases, at block activation rates of 0.5102 and 0.7623. With model weights fixed, PAHM improves Human198 LPIPS, Body DINO, and PCK over Recent-7, with paired 95% intervals excluding zero; the PSNR difference is inconclusive. On Long48-human late frames, PAHM improves Body DINO by 0.0341 (95% CI $[0.0129, 0.0578]$) and LPIPS by 0.0099 ($[0.0009, 0.0189]$). The intervals for PSNR, PCK, and temporal LPIPS span zero, placing the clearest benefit in appearance consistency. PoseLoop recurrence supports this interpretation: grouped DINO and LPIPS return errors decrease by 0.00569 ($[0.00109, 0.01063]$) and 0.00161 ($[0.00061, 0.00258]$). These paired estimates average raw-sequence groups; Table 9 averages eligible pairs.

Table 9: **Perceptual recurrence on 222 PoseLoop return pairs.** Values are pair-macro means; paired effects in the text instead average raw-sequence groups. Pairs are selected only from input controls and shared by all methods. DINO/LPIPS return errors and pose-delta error are lower-is-better; pair PSNR is higher-is-better. The measures characterize 2D appearance and detector-based pose recurrence.

History	DINO RetErr↓	LPIPS RetErr↓	Pair PSNR↑	Pose Δ err.↓
Full PAHM 2+5	0.18599	0.29374	14.7216	0.01980
Recent-5	0.19983	0.29650	14.7713	0.01972
Recent-7	0.19090	0.29526	14.7153	0.02036

C.3 QUALITATIVE ABLATION COMPARISONS

Figures 8–10 illustrate the effects of conditioning design, generated-history training, and pose-addressed retrieval, corresponding to panels (a)–(c) of Table 3. Each comparison uses a common reference and synchronized target timestamps.

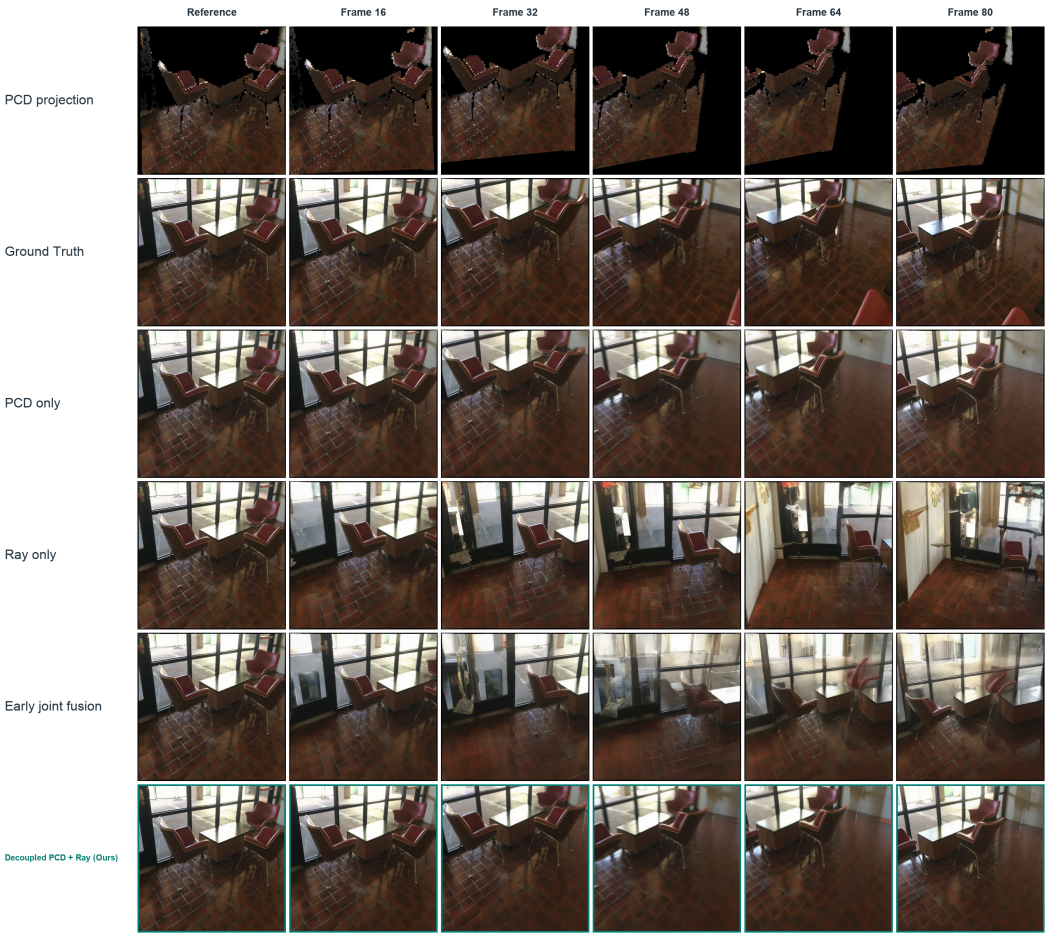

Figure 8: **Qualitative ablation of compositional camera conditioning.** The projected-PCD row visualizes the warped reference-view evidence supplied to the model; the remaining rows show ground truth and generated outputs at the same timestamps. PCD-only conditioning largely follows the scene geometry but misses part of the requested view change, while ray-only conditioning and early joint fusion accumulate pronounced structural errors. The proposed decoupled PCD/ray adapters better preserve the chairs, glass doors, and tiled floor as the camera moves.



Figure 9: **Qualitative ablation of generated-history training.** The SMPL-normal row shows the synchronized body command. Across this 201-frame rollout, removing DMD progressively blurs the person and its clothing texture, whereas removing self forcing introduces color and structured-appearance artifacts. Combining self forcing and DMD better maintains the patterned shirt, controlled body, and otherwise stable scene.

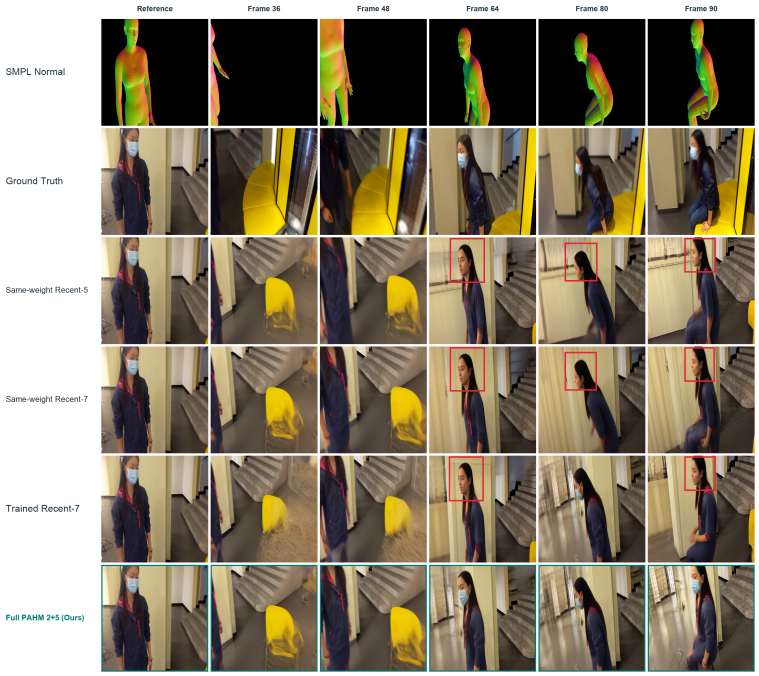

Figure 10: **Qualitative ablation of pose-addressable history.** Same-weight Recent-5 and Recent-7 alter only history selection at inference, isolating PAHM from checkpoint differences; Trained Recent-7 is an additional separately trained baseline. Red boxes mark visible faces where the recent-only variants lose the reference mask. Trained Recent-7 recovers it at Frame 80 but loses it at Frames 64 and 90, whereas full PAHM preserves it at these timestamps.