AESOP: ASYMMETRIC HUMAN–CAMERA GENERATION WITH TRANSLATION-INTENSITY CONTROL

Anonymous authors
Paper under double-blind review

ABSTRACT

Human motion defines an action, while a camera trajectory determines how it is presented. Camera generation for a given human motion and joint human–camera generation are usually treated as separate tasks, although both share an asymmetric dependency: human motion can be generated independently, whereas the camera responds to the realized action. We introduce **AESOP**, a unified framework with an independent human pathway and a shared human-conditioned camera generator. Its asymmetric architecture serves both tasks while preserving the human output during camera generation. Although human context anchors the shot to the action and camera text describes its movement, translation intensity remains underspecified. We therefore construct trajectory pairs that differ in camera translation magnitude while sharing human motion and camera text, then use these pairs to learn an explicit intensity condition. Experiments on the PulpMotion dataset demonstrate strong camera distributional and framing quality in both tasks and effective control over camera translation intensity.

1 INTRODUCTION

Text-conditioned human motion generation can now produce diverse, natural actions that follow language descriptions (Guo et al., 2022; Tevet et al., 2023; Zhang et al., 2024; Chen et al., 2023; Zhang et al., 2023; Guo et al., 2024). Authoring an animation also requires choosing how the action is viewed. Cinematography treats framing and camera movement as constructive choices that organize screen space, direct attention, and shape how viewers experience an event (Bordwell et al., 2024; Brown, 2022). Human motion and camera trajectories therefore play complementary but not interchangeable roles: human motion defines the action, whereas camera trajectory presents that action for the viewer. A human-aware camera is therefore part of motion storytelling.

Prior work approaches this relationship through two broad formulations. Conditional camera methods adapt a trajectory or shot plan to an available human motion (Courant et al., 2024; Kizil et al., 2026), leaving human synthesis outside the task. Coupled systems synthesize human motion and camera trajectories within one generative process (Courant et al., 2026; Cheng et al., 2026), treating their generation as interdependent. These formulations leave an opportunity to unify both tasks around the distinct roles of human action and camera response.

We introduce **AESOP** (**A**symmetric **E**ncoding and **S**ynthesis of **O**bserver **P**aths). Two observations guide our design. First, human action can be determined independently of the camera, while both tasks can share a camera generator conditioned on supplied or generated human motion. Second, human context and camera text specify an action and a shot description but leave translation intensity underspecified. Together, these observations motivate a shared asymmetric architecture and explicit geometric supervision for camera control.

For the architecture, we propose an independent human representation and generator together with a shared human-conditioned camera generator through a common latent interface. Camera training freezes the human pathway; joint inference completes human generation before sampling the camera trajectory. This one-way dependency serves both tasks while preserving the supplied or generated human action during camera generation.

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.

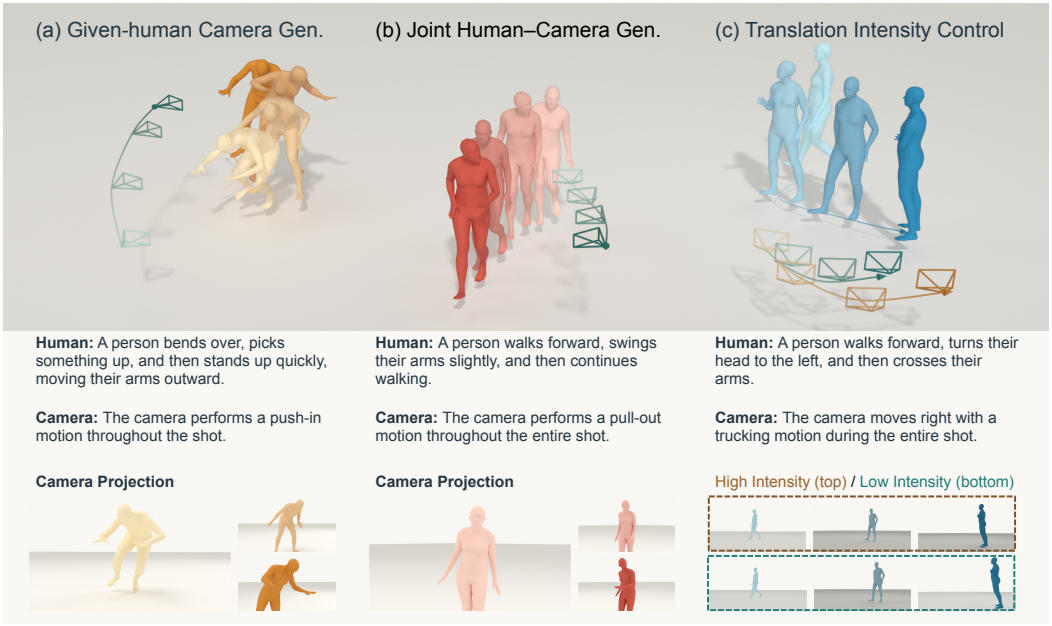


Figure 1: AESOP supports (a) camera generation for a given human motion, (b) joint human–camera generation, and (c) continuous camera translation-intensity control (the amount of camera travel) in both tasks without changing the supplied or generated human motion.

To learn explicit translation intensity, we construct *intensity-paired augmentation* (IPA): weaker and stronger camera trajectories which share human motion and camera text. IPA contracts or expands smooth displacement about the camera center at the first frame, and geometry and visibility checks screen the targets. Auxiliary *direction-paired augmentation* (DPA) supplies opposite-direction Truck and Dolly examples. After original-pair camera training, DPA and then IPA are introduced through continuation while retaining earlier data.

On the PulpMotion dataset, AESOP achieves strong camera distributional and framing quality in both tasks. It reduces given-human r-FPD by 49.9% relative to DIRECTOR-C and joint camera FDC by 87.8% relative to PulpMotion DiT. Its learned intensity condition controls camera travel.

Our contributions are threefold. (1) We introduce a unified asymmetric framework with an independent human pathway and a shared human-conditioned camera generator for given-human and joint generation. (2) We learn explicit camera translation-intensity control from geometry-grounded intensity pairs, with directional pairs providing auxiliary supervision. (3) We evaluate both tasks, demonstrating substantially improved camera distributional and framing quality, strong human distribution quality, and effective camera travel intensity control.

2 Related Work

2.1 Human motion generation

Text-to-motion research has expanded from synthesizing plausible action clips to supporting more detailed and flexible authoring. Paired motion–language data such as HumanML3D (Guo et al., 2022) and MotionMillion (Fan et al., 2025) made semantic fidelity and motion quality central evaluation targets. Diffusion-based models (Tevet et al., 2023; Zhang et al., 2024; Chen et al., 2023) and discrete generative models (Zhang et al., 2023; Guo et al., 2024) offer complementary approaches, with masked autoregressive diffusion extending continuous motion modeling (Meng et al., 2025). As clip synthesis improved, the focus broadened to instructions and controls beyond a single action label: finer supervision addresses spatial and temporal detail (Wu et al., 2025), streaming generation accommodates incoming instructions (Zhao et al., 2025; Xiao et al., 2025), and larger-scale training

expands action coverage and controllability (Wen et al., 2025; Rempe et al., 2026). Video-generation priors also support motion synthesis (Lin et al., 2026).

2.2 Human-aware camera generation

Camera synthesis accounts for the subject it depicts. For an available performance, the problem is to generate a compatible viewpoint and trajectory: DIRECTOR (Courant et al., 2024) introduces character-aware text-to-camera generation, while DanceCamera3D (Wang et al., 2024b) studies camera movement conditioned on dance and music. This line of work establishes human motion as useful camera context. When the performance itself is also generated, fixing human motion as input is insufficient. PulpMotion (Courant et al., 2026) incorporates screen-space framing into human–camera generation, and Towards Storytelling Animations (Cheng et al., 2026) studies joint synthesis of their temporal behavior. AESOP uses a shared complete-sequence human interface to support both supplied and generated performances.

2.3 Explicit and geometry-grounded camera control

Geometric camera interfaces support composition constraints (Lino & Christie, 2015); learned keyframing systems combine camera styles with keyframe and velocity control (Jiang et al., 2021). GenDoP generates camera trajectories from text, optionally conditioned on initial-frame RGB-D observations (Zhang et al., 2025). Actor-relative framing in Auteur (Kizil et al., 2026) and visual preference optimization in VERTIGO (Li et al., 2026) improve how directing intent is expressed or followed. Even with human context, however, one instruction can admit many trajectories, leaving the strength and direction of movement ambiguous. Explicit trajectory interfaces address this ambiguity by supplying the geometry itself: MotionCtrl (Wang et al., 2024a) and CameraCtrl (He et al., 2025) condition video generation on camera paths. This shifts trajectory authoring to the user. AESOP instead supplies training examples that vary direction or translation intensity while holding human motion fixed. Paired directional instructions and a continuous intensity scalar expose these geometric choices without requiring a target path.

3 Method

3.1 Preliminaries

Problem definition. Let $H = (h_0, \dots, h_{T-1})$ denote human motion and $C = (c_0, \dots, c_{T-1})$ a synchronized camera trajectory, with text instructions T_H and T_C. Following PulpMotion (Courant et al., 2026), human motion data encode root motion and heading, local joint positions, and joint rotations; camera trajectory data encode human-relative position, camera rotation, field of view, and translation velocity. **Given-human camera generation** maps (H, T_C) to C. **Joint human–camera generation** maps (T_H, T_C) to (H, C). Both tasks accept an optional translation-intensity condition a, with $a = 1$ specifying default behavior and smaller or larger values requesting weaker or stronger camera-center translation.

Flow matching for human and camera generation. For flow time $\sigma \in [0, 1]$, clean latent z and Gaussian noise ϵ, define $z_\sigma = (1 - \sigma)z + \sigma\epsilon$ and target velocity $\epsilon - z$ (Lipman et al., 2023). The human and camera flows minimize

$$\begin{aligned} \mathcal{L}_H^{\mathrm{FM}} &= \mathbb{E}\Big[\|v_\phi(z_{H,\sigma}, \sigma, T_H) - (\epsilon_H - z_H)\|_2^2\Big], \\ \mathcal{L}_C^{\mathrm{FM}} &= \mathbb{E}\Big[\|v_\theta(z_{C,\sigma}, \sigma, z_H, T_C, a) - (\epsilon_C - z_C)\|_2^2\Big]. \end{aligned} \tag{1}$$

Here z_H and z_C are normalized human and camera latents, and the camera flow receives the complete, fixed human latent sequence. Sampling integrates the predicted velocity from noise at $\sigma = 1$ to a clean latent at $\sigma = 0$.

3.2 Unified asymmetric human–camera generation

AESOP learns its asymmetric representation in Stage 1 and the human and camera generators in Stage 2. Geometry-grounded pairs provide translation-intensity and auxiliary direction supervision.

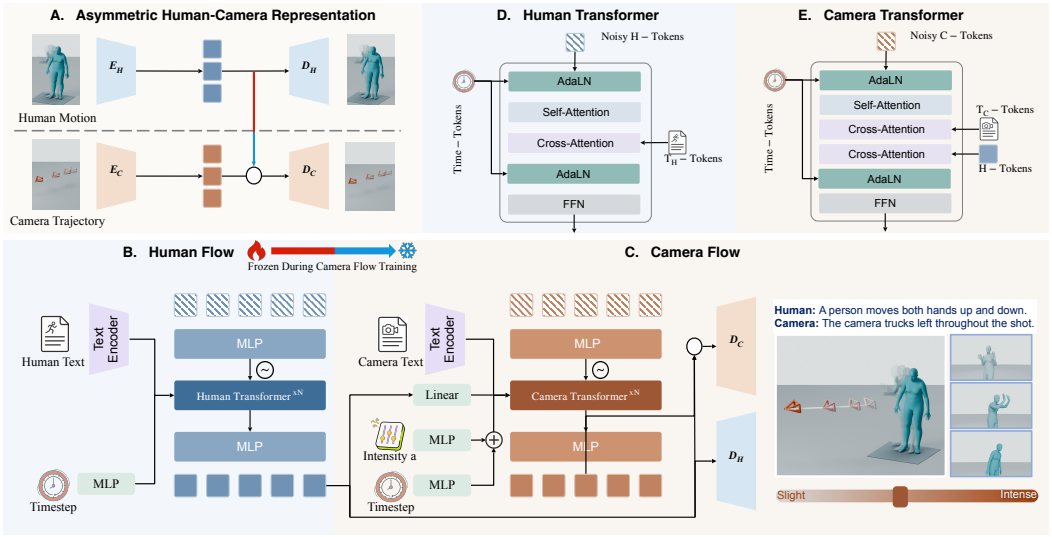


Figure 2: AESOP overview. (A) Separate encoders produce 128-channel human and 64-channel camera latents; the camera decoder reads their concatenation. (B,C) Human generation supplies the complete latent sequence to both human decoding and camera conditioning. Given-human generation instead encodes the supplied motion. (D,E) Human and camera Transformer blocks.

As illustrated in Fig. 2(B,C), AESOP factorizes generation as

$$p(H, C \mid T_H, T_C, a) = p_\phi(H \mid T_H)\, p_\theta(C \mid H, T_C, a). \tag{2}$$

Both tasks use the same conditional camera model $p_\theta(C \mid H, T_C, a)$. Given-human generation encodes the supplied motion, whereas joint generation first samples the complete human latent from $p_\phi(H \mid T_H)$ and reuses it for camera conditioning and human decoding.

Asymmetric human–camera representation. Stage 1 (Fig. 2(A)) first trains a human encoder–decoder and then freezes it while training the camera encoder–decoder:

$$\begin{aligned} z_H &= E_H(H), \quad \hat{H} = D_H(z_H), \\ z_C &= E_C(C), \quad \hat{C} = D_C([\mathrm{sg}(z_H); z_C])\,. \end{aligned} \tag{3}$$

Here $[\cdot;\cdot]$ denotes channel concatenation and sg stops gradients. The camera encoder maps camera trajectory data to a latent sequence, and the camera decoder combines this sequence with the frozen human latents. Human reconstruction depends only on z_H. Supplementary Section B.1 illustrates the difference from PulpMotion's symmetric representation. Stage 1 uses reconstruction and temporal-difference losses, with additional human root and yaw terms; the full objectives appear in Supplementary Section B.2.

Independent human generation. In Stage 2, we train the human flow (Fig. 2(B,D)) with the objective in Eq. 1, then freeze its parameters before camera training. At inference, the human flow samples a complete latent sequence from human text and Gaussian noise, and the frozen human decoder reconstructs the motion. For fixed human text and noise, camera guidance, intensity and camera noise leave the generated human motion unchanged.

Human-aware camera generation. During training, the camera flow (Fig. 2(C,E)) minimizes $\mathcal{L}_C^{\mathrm{FM}}$ over θ, conditioned on the complete 128-channel encoded ground-truth human sequence while predicting velocity in the 64-channel camera latent space. The human flow and Stage 1 representation remain frozen. Training has three phases within Stage 2: original-pair training, DPA continuation, and IPA continuation. Both continuation phases mix original and augmented pairs at predefined sampling ratios, with DPA pairs retained during IPA training. Table S3 gives the schedule.

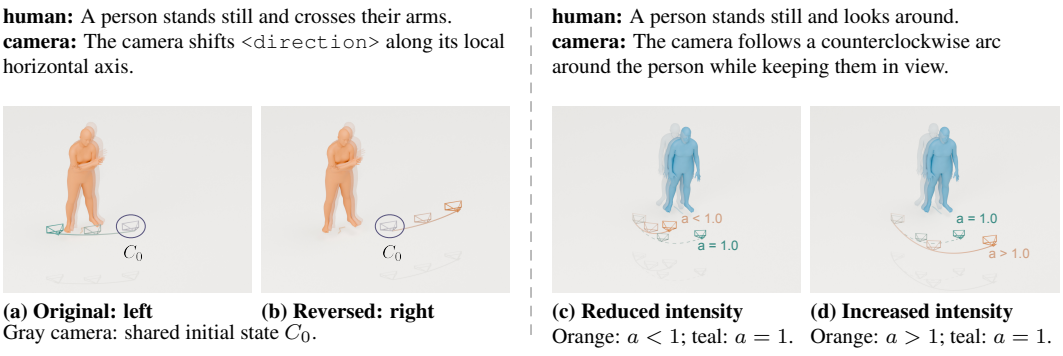
At inference, given-human generation uses the encoded supplied motion, while joint generation uses the completed human-flow latent directly (Fig. 2(B,C)). The same camera flow and decoder serve both tasks. Camera training drops text with probability 0.1 while retaining human context and intensity, enabling classifier-free guidance (Ho & Salimans, 2021) between the text-conditioned velocity v_θ^c and text-dropped velocity v_θ^u:

$$v_\theta^g = v_\theta^u + g\big(v_\theta^c - v_\theta^u\big). \tag{4}$$

The text-conditioned and text-dropped velocity predictions use the same human latent, intensity a, noisy camera latent, and flow time. Here g is the camera-text guidance weight.

3.3 Geometry-grounded camera control

Ordinary paired data rarely show how the same action should be filmed when only camera direction or translation intensity changes. DPA constructs opposite-direction targets, while IPA constructs targets at different translation intensities in physical camera space (Fig. 3).



(a) Original: left **(b) Reversed: right**
Gray camera: shared initial state C_0.

(c) Reduced intensity **(d) Increased intensity**
Orange: $a < 1$; teal: $a = 1$. Orange: $a > 1$; teal: $a = 1$.

Figure 3: Training target construction of DPA and IPA. (a,b) Opposite directions share human motion and the initial camera state. (c,d) Intensity variants retain human motion and camera text.

Data construction. We mine high-confidence atomic Truck and Dolly events for opposite-direction pairs, and translation-active trajectories for weaker/stronger intensity variants. Both constructions fix human motion. Geometry, visibility, framing, dynamics and representation checks screen the targets; Supplementary Section C details the construction and acceptance criteria for both DPA and IPA.

Frames are indexed by $t = 0, \ldots, T-1$. A direction event occupies $[t_s, t_e] \subseteq [0, T-1]$ which may cover a subinterval of the sequence. For direction pairing, let $q^+(t)$ be the original signed event coordinate over $[t_s, t_e]$: lateral camera-center position for trucking, or camera–human radial distance for dolly translation. The opposite target $q^-(t)$ reflects its displacement about the event start. For intensity control, we smooth the original world-space camera-center path p_t to obtain its low-frequency component s_t; $r_t = p_t - s_t$ contains the remaining local fluctuations. The two transformations are

$$\begin{aligned} q^-(t) &= 2q^+(t_s) - q^+(t), \qquad t \in [t_s, t_e], \\ p_t^{(a)} &= p_0 + a(s_t - s_0) + \min(a, 1)(r_t - r_0). \end{aligned} \tag{5}$$

Direction targets (Fig. 3a–b) use paired minimal text templates that differ only in direction and preserve event timing, the displacement magnitude of q, initial camera state, rotation, and field of view; an endpoint offset maintains continuity after the event. Constructed intensity targets (Fig. 3c,d) preserve camera text, rotation, field of view, and sequence timing. They recover p_t exactly at $a = 1$ and amplify only the smooth component when $a > 1$, avoiding amplification of residual jitter.

We also select low-translation sources and pair each unchanged camera target with both $a = 1$ and a non-default intensity. These unchanged-target pairs enter the final training phase alongside active variants, teaching the model to retain low-translation shots under intensity changes. We embed intensity with a two-layer MLP, $e_a = W_2\,\mathrm{SiLU}(W_1[a-1, (a-1)^2]^\top)$, and add e_a to the camera flow's timestep condition (Fig. 2(C)). The learned projections W_1, W_2 are bias-free and W_2 is zero-initialized, giving $e_1 = 0$ throughout training and an initially inactive intensity pathway.

Paired supervision. Both members of each control pair contribute to the flow-matching objective in Eq. 1, sharing human context, Gaussian noise, flow time and camera-text dropout decisions. DPA changes the directional prompt with the target; IPA retains the prompt and changes the intensity label. During target construction, candidate intensities are sampled in $0 < a < 1$ and $1 < a < 2$, with values nearer $a = 1$ proposed more often. Each retained translation-active source supplies one accepted weaker target and one accepted stronger target.

4 EXPERIMENTS

We evaluate camera quality in given-human and joint human–camera generation, human generation quality, and geometry-grounded camera control.

4.1 EXPERIMENTAL SETUP

Data and implementation. We use the public PulpMotion split (Courant et al., 2026), with 162,760 training sequences and 4,053 test sequences, evaluating both tasks at valid sequence lengths. Geometry-grounded training uses 601 direction sources, 8,000 translation-active sources and 2,000 low-translation sources from the training split. Unless stated otherwise, AESOP uses 50-step Euler sampling per stream, human guidance 1, camera guidance $g = 1.5$ and translation intensity $a = 1$. Both tasks share the camera model. Supplementary Sections B.1 and B.2 give architecture and optimization details.

Metrics. Table 1 reports camera distribution distance (FDC), camera-text alignment (CLaTr), movement-caption F1 (Courant et al., 2024), and PRDC recall in the frozen CLaTr embedding space (Naeem et al., 2020). Framing is measured by r-FPD, the distribution distance of projected human joints in normalized image coordinates, and Out (%), the valid-frame rate at which none of nine key joints is in front of the camera and inside the image (Courant et al., 2026). Camera-center ADE measures trajectory error (m) for reconstruction and given-human generation; rotation error (°) compares orientations with the reference camera, including in joint generation. Human metrics use a common frozen evaluator (Petrovich et al., 2023): distribution distance $\mathrm{FD}_{\mathrm{TMR}}$, text alignment TMR and PRDC. CLaTr and TMR report mean nonnegative text cosine similarity scaled by 100. Table S2 adds reconstruction diagnostics.

Repeated-run results use $\mathrm{mean}^{\pm\mathrm{SD}}$, where SD denotes sample standard deviation. Generation and reconstruction comparisons aggregate three independent training seeds; direction-pair results aggregate three paired camera-noise sets from one training seed. Control sweeps fix one checkpoint and share sampling noise across settings.

4.2 COMPARISON WITH PRIOR METHODS

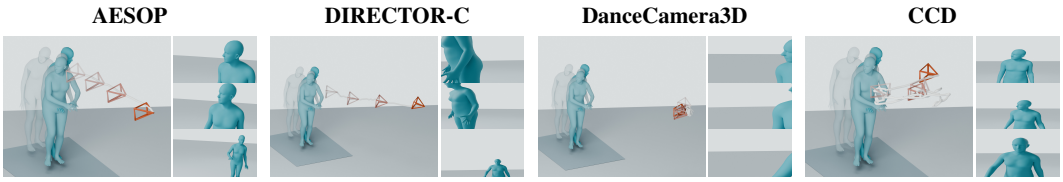
Baselines. Given-human comparisons include DIRECTOR-C (Courant et al., 2024), DanceCamera3D (Wang et al., 2024b), CCD (Jiang et al., 2024), and our CCD-H adaptation. DIRECTOR-C conditions on camera text and the human root-translation trajectory; DanceCamera3D replaces music with camera text. CCD and CCD-H share a human-relative camera representation. CCD uses a text-only network; CCD-H additionally reads human joints and replaces the text prefix with text cross-attention. PulpMotion (Courant et al., 2026) uses DiT (Peebles & Xie, 2023) and MAR (Li et al., 2024) backbones for joint human–camera generation. Table 1 uses original-data endpoints for all baselines. Supplementary Section D.1 details adaptations and training exposure.

Quantitative comparison. AESOP leads both tasks in FDC, CLaTr and caption F1 (Table 1). Given-human FDC drops from DIRECTOR-C's 7.29 to 5.97; joint FDC drops from PulpMotion DiT's 83.80 to 10.18, while CLaTr rises from 44.92 to 69.81. Without augmentation, AESOP attains FDC 6.99/11.39 for given-human/joint generation, below DIRECTOR-C and PulpMotion DiT, respectively. DPA further improves camera quality; IPA adds intensity control while largely preserving generation quality. For framing, AESOP reduces r-FPD from DIRECTOR-C's 1.993 to 0.998 in given-human generation and from DiT's 7.772 to 0.581 in joint generation. Human $\mathrm{FD}_{\mathrm{TMR}}$ is 100.98 versus DiT/MAR's 366.21/319.98, with lower TMR alignment. CCD-H achieves the lowest given-human Out.

Table 1: Quantitative comparison. AE rows report reconstruction; the remaining model rows report generation. AESOP (no aug.) uses original pairs, (+DPA) adds direction pairs, and full AESOP further adds intensity pairs. Human scores are shared across AESOP camera-training phases. Bold marks the best displayed value within each reconstruction/task group, excluding GT; blue marks AESOP.

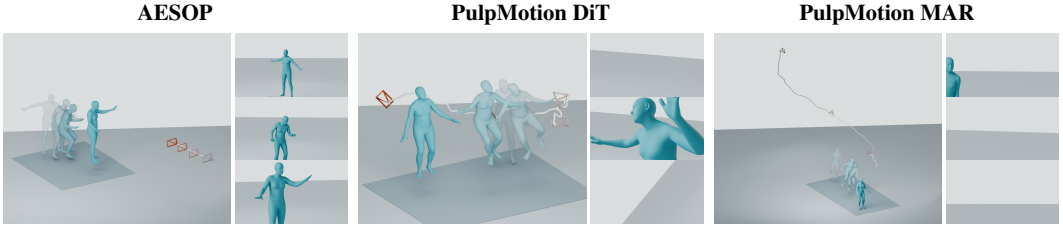
	camera distribution & alignment				camera framing & geometry				human distribution & alignment	
Method	FDC↓	CLaTr↑	F1↑	Recall$_C$↑	r-FPD↓	Out↓	ADE↓	Rot.↓	FD$_{\text{TMR}}$↓	TMR↑
GT reference	≈ 0	70.237	0.945	1.00	0.003	0.71	0.000	0.002	≈ 0	18.398
PulpMotion AE (sym.)	$16.59^{\pm 3.52}$	$59.91^{\pm 1.38}$	$0.739^{\pm 0.032}$	$0.973^{\pm 0.005}$	$0.133^{\pm 0.02}$	$3.93^{\pm 0.55}$	$0.110^{\pm 0.01}$	$1.37^{\pm 0.10}$	$38.79^{\pm 9.99}$	$\mathbf{20.15^{\pm 0.28}}$
AESOP AE (asym.)	$\mathbf{0.11^{\pm 0.10}}$	$\mathbf{70.10^{\pm 0.13}}$	$\mathbf{0.941^{\pm 0.001}}$	$\mathbf{0.999^{\pm 0.0004}}$	$\mathbf{0.089^{\pm 0.01}}$	$\mathbf{2.93^{\pm 0.28}}$	$\mathbf{0.029^{\pm 0.0006}}$	$\mathbf{0.41^{\pm 0.01}}$	$\mathbf{10.05^{\pm 0.89}}$	$17.56^{\pm 0.12}$
Given-human camera generation										
DanceCamera3D	$180.08^{\pm 5.24}$	$20.92^{\pm 0.54}$	$0.252^{\pm 0.002}$	$0.600^{\pm 0.01}$	$2.692^{\pm 0.48}$	$21.72^{\pm 2.33}$	$2.995^{\pm 0.07}$	$63.52^{\pm 1.19}$	–	–
CCD	$359.70^{\pm 6.02}$	$9.53^{\pm 0.19}$	$0.118^{\pm 0.004}$	$0.632^{\pm 0.02}$	$1.937^{\pm 0.19}$	$0.63^{\pm 0.10}$	$2.891^{\pm 0.04}$	$75.58^{\pm 2.09}$	–	–
CCD-H	$375.97^{\pm 39.43}$	$11.32^{\pm 1.58}$	$0.134^{\pm 0.020}$	$0.706^{\pm 0.06}$	$2.942^{\pm 2.35}$	$\mathbf{0.51^{\pm 0.04}}$	$2.712^{\pm 0.14}$	$68.49^{\pm 2.60}$	–	–
DIRECTOR-C	$7.29^{\pm 0.42}$	$64.33^{\pm 0.28}$	$0.848^{\pm 0.001}$	$\mathbf{0.897^{\pm 0.003}}$	$1.993^{\pm 0.04}$	$10.26^{\pm 0.22}$	$2.687^{\pm 0.07}$	$58.16^{\pm 0.36}$	–	–
AESOP (no aug.)	$6.99^{\pm 0.20}$	$69.60^{\pm 0.33}$	$0.905^{\pm 0.006}$	$0.866^{\pm 0.01}$	$1.160^{\pm 0.07}$	$12.56^{\pm 0.43}$	$1.577^{\pm 0.01}$	$33.00^{\pm 0.46}$	–	–
AESOP (+ DPA)	$6.03^{\pm 0.17}$	$70.65^{\pm 0.28}$	$0.915^{\pm 0.004}$	$0.860^{\pm 0.004}$	$1.001^{\pm 0.03}$	$11.51^{\pm 0.21}$	$1.515^{\pm 0.01}$	$32.10^{\pm 0.36}$	–	–
AESOP	$\mathbf{5.97^{\pm 0.16}}$	$\mathbf{70.72^{\pm 0.32}}$	$\mathbf{0.916^{\pm 0.005}}$	$0.863^{\pm 0.01}$	$\mathbf{0.998^{\pm 0.04}}$	$11.47^{\pm 0.22}$	$\mathbf{1.512^{\pm 0.01}}$	$\mathbf{32.06^{\pm 0.41}}$	–	–
Joint human–camera generation										
PulpMotion DiT	$83.80^{\pm 3.30}$	$44.92^{\pm 2.36}$	$0.592^{\pm 0.027}$	$0.711^{\pm 0.02}$	$7.772^{\pm 0.73}$	$35.39^{\pm 1.28}$	–	$\mathbf{66.76^{\pm 1.82}}$	$366.21^{\pm 21.52}$	$\mathbf{24.20^{\pm 0.97}}$
PulpMotion MAR	$131.55^{\pm 9.14}$	$38.64^{\pm 2.42}$	$0.514^{\pm 0.033}$	$0.769^{\pm 0.03}$	$8.169^{\pm 0.65}$	$39.08^{\pm 2.14}$	–	$70.53^{\pm 1.45}$	$319.98^{\pm 7.38}$	$21.57^{\pm 0.87}$
AESOP (no aug.)	$11.39^{\pm 0.67}$	$68.94^{\pm 0.34}$	$0.880^{\pm 0.002}$	$\mathbf{0.826^{\pm 0.01}}$	$0.665^{\pm 0.03}$	$9.13^{\pm 0.24}$	–	$71.56^{\pm 0.53}$	$\mathbf{100.98^{\pm 1.91}}$	$20.46^{\pm 0.34}$
AESOP (+ DPA)	$10.38^{\pm 0.47}$	$69.76^{\pm 0.33}$	$0.887^{\pm 0.005}$	$0.824^{\pm 0.01}$	$\mathbf{0.576^{\pm 0.02}}$	$\mathbf{8.45^{\pm 0.08}}$	–	$71.82^{\pm 0.88}$	$\mathbf{100.98^{\pm 1.91}}$	$20.46^{\pm 0.34}$
AESOP	$\mathbf{10.18^{\pm 0.50}}$	$\mathbf{69.81^{\pm 0.34}}$	$\mathbf{0.891^{\pm 0.006}}$	$0.820^{\pm 0.01}$	$0.581^{\pm 0.02}$	$8.47^{\pm 0.04}$	–	$71.87^{\pm 0.85}$	$\mathbf{100.98^{\pm 1.91}}$	$20.46^{\pm 0.34}$

(a) Given-human camera generation



human: A person walks forward, turns his head to the left, and continues walking.
camera: The camera performs a pull-out motion throughout the entire shot.

(b) Joint human–camera generation



human: A person runs forward, extends their arms outward, and then jumps into an elevator.
camera: The camera executes a push-in motion throughout the entire shot.

Figure 4: **Qualitative comparison on test examples.** (a) Given-human camera generation with a pull-out prompt. (b) Joint human–camera generation with a push-in prompt. Each method has a global view and three selected camera projections, ordered from top to bottom in time. Global views show human–camera geometry at individually fitted scales; projections show framing over time.

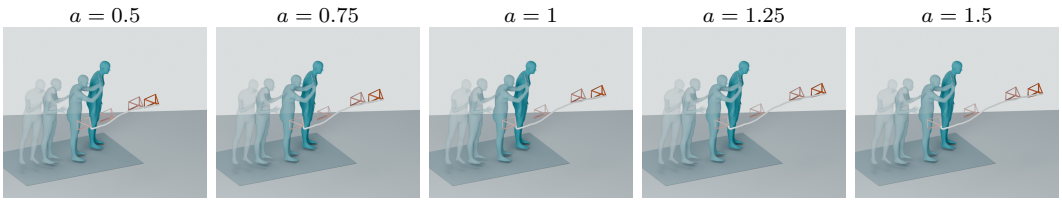
Qualitative comparison. Figure 4 shows AESOP following the pull-out prompt while retaining the subject, whereas DIRECTOR-C and DanceCamera3D crop the subject and CCD misses the requested movement. In the joint example, AESOP progresses from full-body to upper-body framing; DiT alternates between cropped and empty views, while MAR largely misses the human. Supplementary Section H provides more comparisons; Section G shows external examples from HumanML3D and HY-Motion.

Table 2: Human distributional fidelity and coverage in joint generation in the common frozen human embedding space. Bold marks the best value in each column.

Method	Precision ↑	Recall ↑	Density ↑	Coverage ↑
PulpMotion DiT	$0.809^{\pm 0.02}$	$0.215^{\pm 0.02}$	$0.748^{\pm 0.05}$	$0.465^{\pm 0.02}$
PulpMotion MAR	$0.802^{\pm 0.005}$	$0.312^{\pm 0.01}$	$0.705^{\pm 0.02}$	$0.504^{\pm 0.002}$
AESOP	$\mathbf{0.879}^{\pm \mathbf{0.002}}$	$\mathbf{0.835}^{\pm \mathbf{0.01}}$	$\mathbf{0.933}^{\pm \mathbf{0.02}}$	$\mathbf{0.785}^{\pm \mathbf{0.01}}$

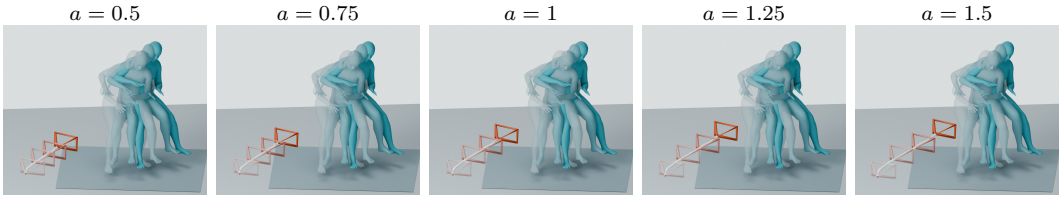
Table 2 evaluates human distributional fidelity and coverage. AESOP improves all four PRDC measures, with recall of 0.835 versus 0.215/0.312 for DiT/MAR. Higher precision accompanies this broader coverage, complementing the lower human distribution distance in Table 1.

(a) Given-human camera generation



human: A person stands still and gestures with both hands, moving them up and down.
camera: The camera moves right with a trucking motion during the entire shot.

(b) Joint human–camera generation



human: A person sits on a stool, moves their left hand towards their face, and then places it back on the instrument.
camera: The camera gradually moves closer to the subject throughout the shot.

Figure 5: **Translation-intensity examples** with shared human context, text and camera noise. Outputs are regenerated at each intensity; numbers report first-to-last camera-center displacement. Fading snapshots indicate temporal order. Global views are fitted separately to each trajectory's spatial extent.

4.3 Representation and Camera Control

Representation quality. The AE rows in Table 1 measure reconstruction by encoding and decoding reference sequences. AESOP reduces camera-center ADE to 0.029 m and rotation error to $0.41°$, compared with 0.110 m and $1.37°$ for PulpMotion. It also yields lower human and camera distribution distances, whereas PulpMotion has higher human-text alignment (20.15 versus 17.56). Supplementary Table S2 reports additional human positional and camera-motion errors.

Translation intensity. Figures 5 and 6 show qualitative examples and an eleven-value scan from $a = 0.5$ to 1.5 (step 0.1). Path length sums camera-center travel; net displacement measures endpoint distance. Mean path length rises from 0.647 to 1.023 m (given-human) and from 0.367 to 0.612 m (joint). We require travel to be nondecreasing in all ten adjacent intervals: 81.50%/81.37% of given-human/joint clips satisfy this for path length, and 78.61%/78.46% for net displacement. For the 2,106 clips with reference net displacement ≥ 0.2 m, path-length rates reach 99.19%/98.86%. Supplementary Section E reports endpoint results.

CLaTr and F1 increase alongside camera travel, while higher r-FPD and Out indicate a framing trade-off. Given-human ADE and rotation error in both tasks vary little. Given-human FDC is lowest near $a = 0.8$, whereas joint FDC decreases across the grid.

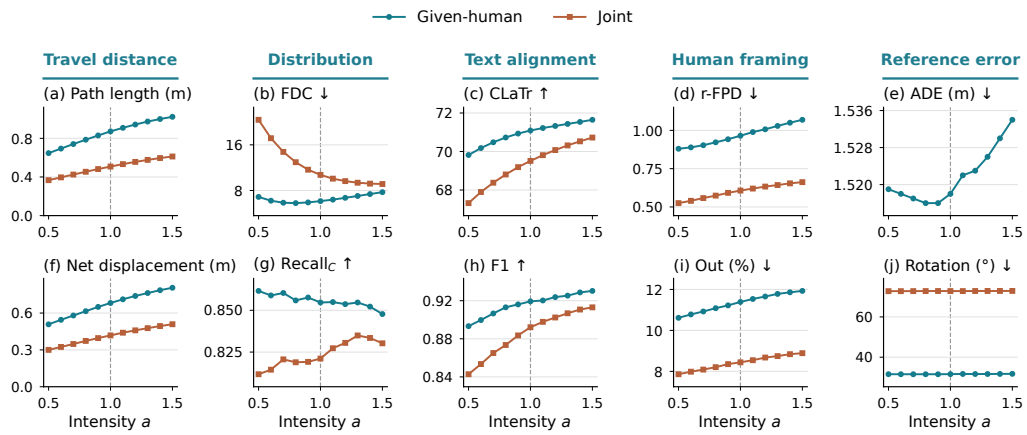


Figure 6: **Translation-intensity response and camera quality.** Columns group travel distance, distribution, text alignment, framing and reference error. Results use one training seed and shared human context, text and camera noise across intensities. ADE is reported for given-human generation. Vertical dashed lines mark $a = 1$.

Geometry-grounded training. We evaluate direction accuracy on original and direction-reversed prompts, comparing translation signs over reference-active axes and time steps. After DPA continuation, AESOP reaches 84.33%/83.11% combined accuracy in given-human/joint generation, exceeding the matched factual-pair control by 1.67/1.38 percentage points (Supplementary Table S6). Accuracy improves in both prompt arms: on reversed prompts, DPA scores 79.00%/79.58%, compared with 77.71%/78.51% for factual continuation. IPA then adds continuous intensity control, with similar default camera quality to the +DPA endpoint in Table 1.

4.4 User study

Table 3: Overall-quality preference for AESOP (%). Scores average clip-level responses, with ties receiving half credit. Intervals are participant-cluster 95% CIs.

Compared with	Pref.	95% CI
Given-human		
CCD	75.3	[64.8, 85.5]
DanceCamera3D	99.2	[97.5, 100.0]
DIRECTOR-C	86.6	[81.5, 91.5]
Joint		
PulpMotion DiT	86.3	[80.1, 91.9]
PulpMotion MAR	92.3	[86.9, 96.9]

We conducted an anonymized paired-comparison study with 29 participants. Each participant completed 20 given-human and 18 joint comparisons, with balanced method pairs and randomized presentation. Participants rate overall camera quality for given-human generation and overall human–camera quality for joint generation. Preference, averaged over clips and equally over opponents, is 87.0% (95% CI: 83.8–90.8) and 89.3% (84.4–93.6), respectively. Table 3 gives the five pairwise comparisons; Supplementary Section F details the protocol and individual criteria.

5 Conclusion

AESOP unifies given-human camera generation and joint human–camera generation through an independent human pathway and a shared human-conditioned camera generator. Geometry-grounded intensity pairs teach explicit translation-intensity control, complemented by auxiliary direction supervision. Experiments on PulpMotion show strong camera distributional and framing quality in both tasks, while the learned intensity condition adjusts camera travel.

Limitations and future directions. The sequence-level intensity condition limits control over individual events, while complete human context and offline sampling limit streaming generation. Future directions include event-wise intensity control and causal, incremental camera generation.

REFERENCES

David Bordwell, Kristin Thompson, and Jeff Smith. *Film Art: An Introduction*. McGraw Hill, 13th edition, 2024.

Blain Brown. *Cinematography: Theory and Practice: For Cinematographers and Directors*. Routledge, 4th edition, 2022.

Xin Chen, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, and Gang Yu. Executing your commands via motion diffusion in latent space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18000–18010, 2023.

Boyuan Cheng, Yingjie Xi, Rui He, Jinhe Na, Ying Cao, Pengjie Wang, Jian J. Zhang, and Xiaosong Yang. Towards storytelling animations: Joint synthesis of human and camera motions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 38376–38386, 2026.

Robin Courant, Nicolas Dufour, Xi Wang, Marc Christie, and Vicky Kalogeiton. E.T. the exceptional trajectories: Text-to-camera-trajectory generation with character awareness. In *European Conference on Computer Vision*, pp. 464–480, 2024.

Robin Courant, Xi Wang, David Loiseaux, Marc Christie, and Vicky Kalogeiton. Pulp motion: Framing-aware multimodal camera and human motion generation. In *The Fourteenth International Conference on Learning Representations*, pp. 98487–98516, 2026.

Ke Fan, Shunlin Lu, Minyue Dai, Runyi Yu, Lixing Xiao, Zhiyang Dou, Junting Dong, Lizhuang Ma, and Jingbo Wang. Go to zero: Towards zero-shot motion generation with million-scale data. In *IEEE/CVF International Conference on Computer Vision*, pp. 13336–13348, 2025.

Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3D human motions from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5142–5151, 2022.

Chuan Guo, Yuxuan Mu, Muhammad Gohar Javed, Sen Wang, and Li Cheng. MoMask: Generative masked modeling of 3D human motions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1900–1910, 2024.

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

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *Conference on Neural Information Processing Systems Workshop on Deep Generative Models and Downstream Applications*, 2021.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020.

Hongda Jiang, Marc Christie, Xi Wang, Libin Liu, Bin Wang, and Baoquan Chen. Camera keyframing with style and control. *ACM Transactions on Graphics*, 40(6):1–13, 2021.

Hongda Jiang, Xi Wang, Marc Christie, Libin Liu, and Baoquan Chen. Cinematographic camera diffusion model. *Computer Graphics Forum*, 43(2):e15055, 2024.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 26565–26577, 2022.

Muhammed Burak Kizil, Enes Sanli, Niloy J. Mitra, Xuelin Chen, Erkut Erdem, Aykut Erdem, and Duygu Ceylan. Auteur: Language-driven cinematographic framing for human-centric video generation. *arXiv preprint arXiv:2606.01900*, 2026.

Mengtian Li, Yuwei Lu, Feifei Li, Chenqi Gan, Zhifeng Xie, and Xi Wang. VERTIGO: Visual preference optimization for cinematic camera trajectory generation. In *European Conference on Computer Vision*, pp. 78–96, 2026.

Tianhong Li, Yonglong Tian, He Li, Mingyang Deng, and Kaiming He. Autoregressive image generation without vector quantization. In *Advances in Neural Information Processing Systems*, volume 37, pp. 56424–56445, 2024.

Jing Lin, Ruisi Wang, Junzhe Lu, Ziqi Huang, Guorui Song, Ailing Zeng, Xian Liu, Chen Wei, Wanqi Yin, Qingping Sun, Zhongang Cai, Lei Yang, and Ziwei Liu. The quest for generalizable motion generation: Data, model, and evaluation. In *International Conference on Learning Representations*, 2026.

Christophe Lino and Marc Christie. Intuitive and efficient camera control with the toric space. *ACM Transactions on Graphics*, 34(4):1–12, 2015.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023.

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.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *The Seventh International Conference on Learning Representations*, 2019.

Zichong Meng, Yiming Xie, Xiaogang Peng, Zeyu Han, and Huaizu Jiang. Rethinking diffusion for text-driven human motion generation: Redundant representations, evaluation, and masked autoregression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27859–27871, 2025.

Muhammad Ferjad Naeem, Seong Joon Oh, Youngjung Uh, Yunjey Choi, and Jaejun Yoo. Reliable fidelity and diversity metrics for generative models. In *International Conference on Machine Learning*, volume 119, pp. 7176–7185. Proceedings of Machine Learning Research, 2020.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *IEEE/CVF International Conference on Computer Vision*, pp. 4172–4185, 2023.

Mathis Petrovich, Michael J. Black, and Gül Varol. TMR: Text-to-motion retrieval using contrastive 3D human motion synthesis. In *IEEE/CVF International Conference on Computer Vision*, pp. 9454–9463, 2023.

Mathis Petrovich, Or Litany, Umar Iqbal, Michael J. Black, Gül Varol, Xue Bin Peng, and Davis Rempe. Multi-track timeline control for text-driven 3D human motion generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshop on Human Motion Generation*, pp. 1911–1921, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, volume 139, pp. 8748–8763. Proceedings of Machine Learning Research, 2021.

Davis Rempe, Mathis Petrovich, Ye Yuan, Haotian Zhang, Xue Bin Peng, Yifeng Jiang, Tingwu Wang, Umar Iqbal, David Minor, Michael de Ruyter, Jiefeng Li, Chen Tessler, Edy Lim, Eugene Jeong, Sam Wu, Ehsan Hassani, Michael Huang, Jin-Bey Yu, Chaeyeon Chung, Lina Song, Olivier Dionne, Jan Kautz, Simon Yuen, and Sanja Fidler. Kimodo: Scaling controllable human motion generation. *arXiv preprint arXiv:2603.15546*, 2026.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *The Ninth International Conference on Learning Representations*, 2021.

Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H Bermano. Human motion diffusion model. In *The Eleventh International Conference on Learning Representations*, 2023.

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

Zixuan Wang, Jia Jia, Shikun Sun, Haozhe Wu, Rong Han, Zhenyu Li, Di Tang, Jiaqing Zhou, and Jiebo Luo. DanceCamera3D: 3D camera movement synthesis with music and dance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7892–7901, 2024b.

Yuxin Wen, Qing Shuai, Di Kang, Jing Li, Cheng Wen, Yue Qian, Ningxin Jiao, Changhai Chen, Weijie Chen, Yiran Wang, Jinkun Guo, Dongyue An, Han Liu, Yanyu Tong, Chao Zhang, Qing Guo, Juan Chen, Qiao Zhang, Youyi Zhang, Zihao Yao, Cheng Zhang, Hong Duan, Xiaoping Wu, Qi Chen, Fei Cheng, Liang Dong, Peng He, Hao Zhang, Jiaxin Lin, Chao Zhang, Zhongyi Fan, Yifan Li, Zhichao Hu, Yuhong Liu, Linus, Jie Jiang, Xiaolong Li, and Linchao Bao. HY-Motion 1.0: Scaling flow matching models for text-to-motion generation. *arXiv preprint arXiv:2512.23464*, 2025.

Bizhu Wu, Jinheng Xie, Meidan Ding, Zhe Kong, Jianfeng Ren, Ruibin Bai, Rong Qu, and Linlin Shen. FineMotion: A dataset and benchmark with both spatial and temporal annotation for fine-grained motion generation and editing. In *IEEE/CVF International Conference on Computer Vision*, pp. 13837–13846, 2025.

Lixing Xiao, Shunlin Lu, Huaijin Pi, Ke Fan, Liang Pan, Yueer Zhou, Ziyong Feng, Xiaowei Zhou, Sida Peng, and Jingbo Wang. MotionStreamer: Streaming motion generation via diffusion-based autoregressive model in causal latent space. In *IEEE/CVF International Conference on Computer Vision*, pp. 10086–10096, 2025.

Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Yong Zhang, Hongwei Zhao, Hongtao Lu, Xi Shen, and Ying Shan. Generating human motion from textual descriptions with discrete representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14730–14740, 2023.

Mengchen Zhang, Tong Wu, Jing Tan, Ziwei Liu, Gordon Wetzstein, and Dahua Lin. GenDoP: Auto-regressive camera trajectory generation as a director of photography. In *IEEE/CVF International Conference on Computer Vision*, pp. 18229–18239, 2025.

Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. MotionDiffuse: Text-driven human motion generation with diffusion model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(6):4115–4128, 2024.

Kaifeng Zhao, Gen Li, and Siyu Tang. DartControl: A diffusion-based autoregressive motion model for real-time text-driven motion control. In *International Conference on Learning Representations*, pp. 23569–23592, 2025.

Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5738–5746, 2019.

SUPPLEMENTARY CONTENTS

A Disclosure of AI Use **14**

B Representation and Implementation **14**

B.1 Symmetric and asymmetric representations 14

B.2 Implementation, training objectives and inference 15

C DPA and IPA Data Construction **16**

C.1 Direction-paired augmentation 16

C.2 Intensity-paired augmentation 17

D Evaluation and Baseline Protocols **19**

D.1 Baseline protocols 19

D.2 Direction-paired training 19

E Translation-Intensity Control **20**

F User Study **20**

G Qualitative results on external motions and prompts **22**

H More Qualitative Results **24**

A Disclosure of AI Use

The authors conceived the research ideas, designed the methods, implemented the core code, and analyzed the experimental results. Under the authors' direction, generative AI tools assisted with supporting code implementation, experimental execution, threshold selection for DPA/IPA data construction, and language refinement. The authors made the final research decisions and take full responsibility for the methods, implementation, experimental results, analyses, and final manuscript.

B Representation and Implementation

B.1 Symmetric and asymmetric representations

PulpMotion's symmetric representation. PulpMotion's shared encoder (Courant et al., 2026) first maps human motion and camera trajectory to a joint latent, which is then split into human and camera components:

$$z^P_{HC} = E_{HC}(H, C), \quad (z^P_H, z^P_C) = \text{split}(z^P_{HC}), \quad z^P_F = W z^P_{HC}. \tag{S1}$$

Independent human and camera decoders reconstruct their respective inputs from z^P_H and z^P_C. A third decoder reconstructs framing from the auxiliary latent z^P_F, obtained by a learned linear map W. The shared encoder and all three reconstruction branches are trained together. Both latents can depend on both inputs: camera reconstruction has access to human context, while human reconstruction also depends on camera input through the shared encoder. Consequently, human representation learning is coupled to camera training, and human encoding requires camera input.

(a) PulpMotion: symmetric representation **(b) AESOP: asymmetric representation**

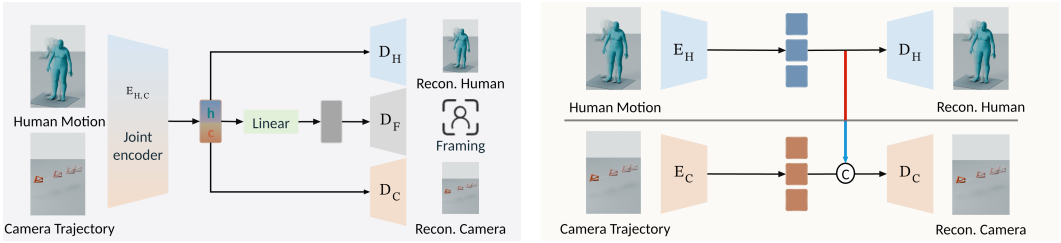


Figure S1: **human–camera representation dependencies.** (a) PulpMotion's symmetric representation encodes both inputs with a shared encoder, reconstructs each from its latent, and learns an auxiliary framing latent from their concatenation (Courant et al., 2026). (b) AESOP separately encodes the two inputs. The human decoder reads only human latents; the camera decoder concatenates camera latents with frozen human latents (circled C). Human representation training is completed before camera training.

Table S1: **Training and reconstruction dependencies.** Both representations provide human context for camera reconstruction. AESOP additionally separates human learning from camera training and makes human reconstruction independent of camera input.

	Training	Reconstruction	
Method	Independent human learning	camera-independent human	human-aware camera
PulpMotion (symmetric)	× Both streams trained together	×	✓
AESOP (asymmetric)	✓ Train human first, then freeze	✓	✓

AESOP's asymmetric representation. AESOP retains PulpMotion's human and camera feature format but uses separate encoders (Fig. S1). Its human decoder reads only the human latent; the camera decoder reads the concatenated human and camera latents.

Table S2 evaluates deterministic encoding and decoding. AESOP has lower human positional, translation-velocity and jerk reconstruction errors, while PulpMotion has lower rotation-velocity

Table S2: **Autoencoder reconstruction on the test split.** Sym./asym. denote PulpMotion's symmetric and AESOP's asymmetric autoencoders. Motion errors are clip-balanced and measured in mm/s, °/s and m/s^3, respectively. MPJPE, ADE and FDE are in meters; aligned MPJPE removes root translation. Bold marks the best value between the autoencoders.

	human reconstruction				camera reconstruction			
Autoencoder	Global MPJPE↓	Aligned MPJPE↓	Root ADE↓	Root FDE↓	FDE↓	Trans. vel.↓	Rot. vel.↓	Trans. jerk↓
PulpMotion (sym.)	$0.1620^{\pm 0.02}$	$0.0617^{\pm 0.01}$	$0.1376^{\pm 0.01}$	$0.3283^{\pm 0.04}$	$0.1677^{\pm 0.02}$	$124.28^{\pm 3.02}$	$\mathbf{10.55}^{\pm \mathbf{0.10}}$	$154.67^{\pm 1.54}$
AESOP (asym.)	$\mathbf{0.1304}^{\pm \mathbf{0.01}}$	$\mathbf{0.0479}^{\pm \mathbf{0.003}}$	$\mathbf{0.1063}^{\pm \mathbf{0.01}}$	$\mathbf{0.2611}^{\pm \mathbf{0.02}}$	$\mathbf{0.0356}^{\pm \mathbf{0.002}}$	$\mathbf{10.11}^{\pm \mathbf{0.81}}$	$10.98^{\pm 0.21}$	$\mathbf{8.55}^{\pm \mathbf{0.40}}$

error. Root-aligned human error is smaller than global error for both representations, indicating the contribution of root-trajectory deviations. Table 1 reports distributional and camera-pose reconstruction quality.

B.2 Implementation, training objectives and inference

Per-frame features. We use PulpMotion's per-frame feature format (Courant et al., 2026), whose human features follow the SMPL-based representation of Petrovich et al. (2024); SMPL is defined by Loper et al. (2015). Each human frame has 199 features describing root height, planar root velocity, yaw velocity, local joint rotations and local joint positions. Each camera frame has 14 features: two field-of-view angles, three camera-minus-human relative-position channels, a six-dimensional rotation representation (Zhou et al., 2019) and three world-space translation increments.

Latent interface. For a sequence of T frames, the human and camera encoders produce $\lceil T/4 \rceil$ latent tokens with 128 and 64 channels, respectively. The human decoder reads only the 128-channel human latent; the camera decoder reads the 192-channel concatenation of human and camera latents.

Noncausal width-256 temporal encoders use stride four; transposed-convolution decoders crop to the requested length. A valid-frame mask marks actual sequence frames and excludes batch padding from losses; latent masks retain $\lceil T/4 \rceil$ tokens. Both flow Transformers have 12 layers, width 512, eight attention heads, FFN multiplier four and dropout 0.1. Camera blocks apply self-attention, camera-text attention, full-human attention and FFN in order, using frozen 512-dimensional CLIP text features (Radford et al., 2021).

For stream $k \in \{H, C\}$, let $\widetilde{z}_k$ denote the encoder output and z_k the normalized latent used by the flow. Normalization uses training-only channel statistics μ_k, s_k and standardized-stream mean/covariance factors m_k, L_k. Cholesky whitening uses ridge 10^{-4}; its inverse precedes decoding:

$$z_k = L_k^{-1}\big((\widetilde{z}_k - \mu_k)/s_k - m_k\big), \qquad \widetilde{z}_k = \mu_k + s_k \odot (m_k + L_k z_k). \tag{S2}$$

Invalid positions are zeroed after either operation. Physical decoding integrates de-normalized translation increments and anchors the first camera center using the decoded relative vector and human root (Courant et al., 2026), using predicted features for initialization. Human-feature, relative-center and translation-increment normalization are distinct from latent whitening.

Stage 1 objectives. Let M mark valid frames and $M_{\Delta,t} = M_t M_{t-1}$ for $t = 1, \ldots, T-1$. Angle brackets average selected scalar elements. For normalized features H, C, cumulative de-normalized yaw ψ, and decoded root trajectory r, the complete reconstruction losses are

$$\begin{aligned} \mathcal{L}_H^{\text{rec}} &= \langle \ell_{\text{SL1}}(\hat{H}, H)\rangle_M + \langle (\Delta\hat{H} - \Delta H)^2\rangle_{M_\Delta} \\ &\quad + 10^{-3}\langle 1 - \cos(\hat{\psi} - \psi)\rangle_M + 3\times 10^{-3}\langle \ell_{\text{SL1}}(\hat{r}, r)\rangle_M, \\ \mathcal{L}_C^{\text{rec}} &= \langle \ell_{\text{SL1}}(\hat{C}, C)\rangle_M + \langle (\Delta\hat{C} - \Delta C)^2\rangle_{M_\Delta}. \end{aligned} \tag{S3}$$

Stage 2 objectives and optimization. We sample $u \sim \mathcal{U}(0,1)$, set $\sigma = 5u/(1+4u)$ and draw $\epsilon \sim \mathcal{N}(0, I)$. The velocity target v^* follows Eq. 1. For n_b valid latent tokens and stream dimension D, define

$$e_b = \frac{\sum_{d,t} M_{b,t}(v_{b,d,t} - v^*_{b,d,t})^2}{D n_b}, \quad \mathcal{L}_{\text{original}} = \frac{1}{B}\sum_b e_b, \quad \mathcal{L}_{\text{continuation}} = \frac{\sum_b n_b e_b}{\sum_b n_b}. \tag{S4}$$

Continuation weights clips by length; original training weights them equally. Each continuation batch has 120 rows, with two rows per control pair. During the 35K DPA phase, direction-pair slots rise from zero to five over the first 2K updates, remain at five through 20K, fall to two through 30K and then to one. During the 15K IPA phase, each batch retains two DPA pair slots; active IPA slots rise to four and null IPA slots to one during the first 1K updates.

Table S3: Training schedule. The final phase additionally trains the intensity MLP at 10^{-4}.

Stage	Phase	Updates	Batch	Learning rate	Frozen components
Reconstruction	human autoencoder	210K	128	5×10^{-5}	None
	camera autoencoder	210K	128	5×10^{-5}	human autoencoder
Generation	human flow	105K	128	2×10^{-4}	Autoencoders
	Original camera	105K	128	10^{-4}	Autoencoders, human flow
	camera + DPA	35K	120	2×10^{-5}	Autoencoders, human flow
	camera + DPA + IPA	15K	120	2×10^{-5}	Autoencoders, human flow

Stage 1 uses AdamW (Loshchilov & Hutter, 2019) with $\beta = (0.9, 0.999)$, zero weight decay, a 1K-step linear warmup and cosine decay to 10^{-6} per phase. Stage 2 uses AdamW with $\beta = (0.9, 0.95)$, weight decay 0.01, gradient clipping 1.0 and EMA decay 0.9999. The human flow warms up for 2K steps and decays its rate by 0.1 at step 80K; original camera training uses a constant rate. Continuations use BF16 and 1K-step warmup.

Inference and conditioning. Sampling follows Section 4.1 with solver shift five and camera-text guidance from Eq. 4. The intensity MLP described in Section 3 has 64 hidden units and a 512-dimensional output added to the camera timestep condition.

C DPA AND IPA DATA CONSTRUCTION

We construct paired camera targets from the 162,760-clip PulpMotion training pool. DPA changes translation direction and its text condition; IPA changes translation intensity while keeping camera text fixed. Both preserve human motion, camera rotation, field of view and timing; the human latent and decoded output are unchanged across each constructed pair. Figures S2 and S3 illustrate the three construction stages and the retained source counts. Table S4 lists the acceptance rules.

C.1 DIRECTION-PAIRED AUGMENTATION

1. Pre-screening. From the training pool, DPA selects 22,783 whole-shot captions containing one Truck or Dolly event. Source-motion checks retain 9,143 clips with stable, sufficient translation. Truck uses a fixed camera-local right axis, obtained by averaging the first $\min(8, \lfloor T/5 \rfloor)$ right vectors, projecting onto the ground plane and normalizing. Dolly uses camera–human root distance.

2. Augmentation. On the accepted interval $[t_s, t_e]$, we reflect the signed coordinate as $q'_t = 2q_{t_s} - q_t$. Truck preserves orthogonal components; Dolly preserves the human-to-camera unit direction. The trajectory is unchanged before the event and carries its endpoint offset afterward. Relative-center and velocity features are recomputed. Paired minimal text templates differ only in direction.

3. Post-checks. Physical checks assess geometry, framing, visible direction effect and non-target leakage, retaining 781 constructed sources (8.54% of the 9,143 eligible sources). Reconstruction checks assess fidelity, direction, response and visibility, yielding 601 sources: 375 Dolly and 226 Truck. These supply 601 original/opposite training pairs with 1,202 targets. Figure S2 illustrates why the decoded direction is checked as well as the raw target.

The DPA visible-effect score is $V = \max(\Delta x/0.08, \Delta\alpha/6^\circ)$ for Truck and $V = \Delta h/0.12$ for Dolly. Here Δx is peak projected subject-center separation in image-width units, $\Delta\alpha$ is peak human-relative azimuth separation, and $\Delta h = \max_t |h'_t/h_t - 1|$ is peak projected box-height ratio change, comparing original/opposite targets within the event. Decoded retention compares this score after versus before reconstruction.

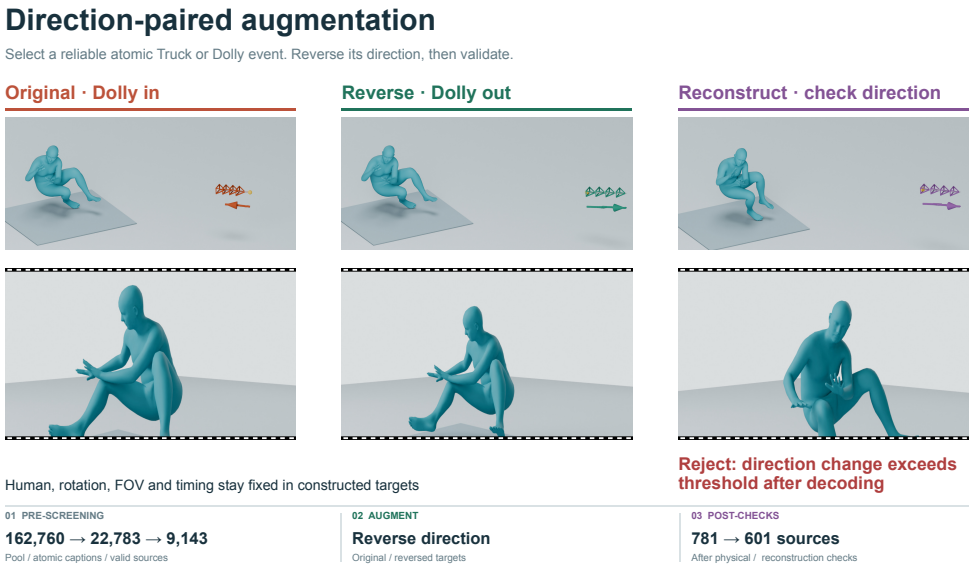


Figure S2: **DPA construction and validation.** Left: the original Dolly-in trajectory. Middle: its opposite Dolly-out target, with human motion, camera rotation, field of view and timing fixed. Right: a reconstruction rejected because its direction change exceeds the acceptance threshold after decoding. The gold point marks the initial camera center; global views show the starting human pose and full trajectory, and projections show the synchronized final frame. The three stages summarize source filtering, direction reversal and target checks; 601 sources pass the final reconstruction check.

C.2 Intensity-paired augmentation

1. Pre-screening. The active branch selects trajectories with sufficient translation to provide weaker/stronger targets. The null branch selects low-translation trajectories and pairs an unchanged target with different intensities, teaching the model to retain low-translation shots under intensity changes. Motion, visibility and distance checks, together with DPA-source exclusion, yield 42,844 active and 75,863 null candidates. Deterministic hash ranking selects 36,000 active sources for target construction.

2. Augmentation. An 11-frame triangular filter with reflection padding decomposes camera centers into $p_t = s_t + r_t$. We apply Eq. 5 and recompute relative-center and velocity features from the transformed trajectory.

For IPA target construction, candidate labels are sampled from $0 < a < 1$ for weaker targets and $1 < a < 2$ for stronger targets. Each interval is divided into four equal-width bins, ordered from smaller to larger a. Their sampling probabilities are $(0.1, 0.2, 0.3, 0.4)$ below one and $(0.4, 0.3, 0.2, 0.1)$ above one, favoring labels closer to $a = 1$. A fixed-seed sampler selects a bin and then a continuous value within it. For each active source, we try up to eight labels per interval and keep the first weaker and first stronger target that pass the checks below. Null sources retain the unchanged camera target while receiving alternating weaker and stronger labels from the same distributions. The eleven-level evaluation uses the narrower range $a \in [0.5, 1.5]$, in steps of 0.1.

3. Post-checks. We retain active sources with an accepted target on both sides of $a = 1$. Physical and construction-representation reconstruction checks yield 14,066 sources (39.07% of the 36,000 constructed sources). After reserving source-video groups for development, the final training set contains 8,000 active sources, yielding 16,000 camera targets labeled $a \neq 1$, and 2,000 null sources. Development control variants are kept separate from control training and the test split.

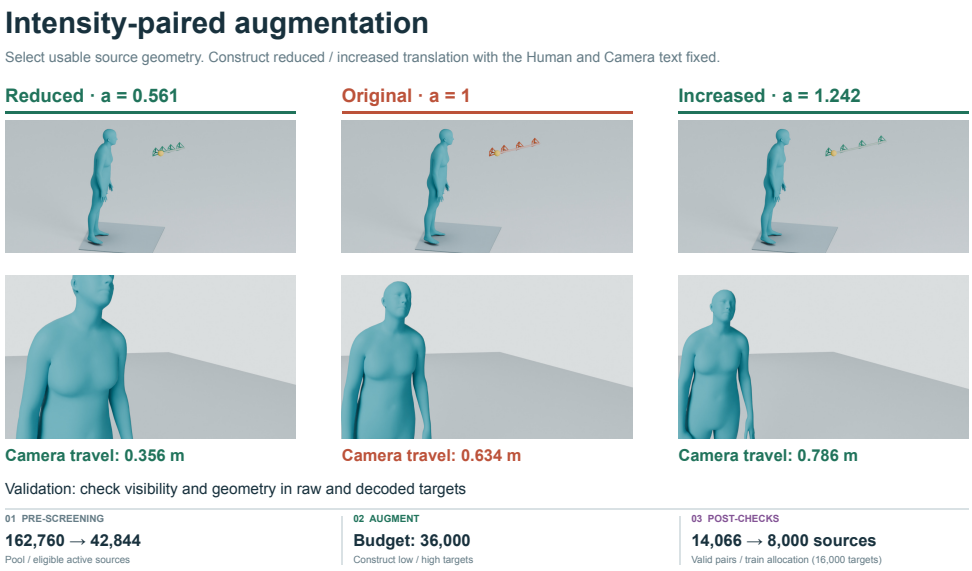


Figure S3: **IPA target construction.** Reduced, original and increased translation share human motion and camera text. Camera travel changes from 0.356 to 0.634 to 0.786 m in this example. The gold point marks the initial camera center; global views and synchronized final-frame projections show the resulting geometry and framing. The three stages summarize the active branch, which supplies 16,000 training targets from 8,000 sources. The accompanying null branch supplies 2,000 low-translation sources with unchanged targets.

Table S4: Construction acceptance criteria for DPA and IPA, grouped by source selection, target geometry, framing and reconstruction.

Operation	Acceptance rules
DPA: direction pairs	
Select source events	Duration ≥ 45 frames; same-sign velocity fraction ≥ 0.80; signed net displacement/path ≥ 0.60. Truck displacement ≥ 0.20 m; Dolly displacement $\geq \max(0.20\text{ m}, 0.10\rho_{t_s})$.
Check target geometry	Opposite-target human distance > 0.25 m; opposite/original response $\in [0.80, 1.25]$. Rotation/FOV unchanged; Truck vertical shift $\leq 10^{-5}$ m and radial relative change ≤ 0.10; Dolly radial-direction change $\leq 0.05°$.
Check framing and effect	No zero-visible frame; visibility ≥ 0.80; opposite outscreen increase ≤ 0.05. Truck: projected center shift ≥ 0.08 image width or viewpoint change $\geq 6°$. Dolly: peak box-height ratio change ≥ 0.12.
Check reconstructed targets	Feature MSE ≤ 0.01181; both direction signs correct; decoded/raw response ratios $\in [0.70, 1.30]$. No framing failure; $V_{\text{dec}} \geq 1$ and $V_{\text{dec}}/V_{\text{raw}} \geq 0.70$.
IPA: intensity pairs	
Select sources	Duration ≥ 1 s. Active: path ≥ 0.4118 m, span ≥ 0.15 m, max-step/path ≤ 0.20. Null: path ≤ 0.10 m. Source visibility ≥ 0.80 and human distance ≥ 0.25 m; exclude DPA sources.
Check magnitude and direction	Active only: absolute requested/actual path-ratio error ≤ 0.08; path/speed-ratio difference ≤ 0.03; direction cosine ≥ 0.90 for $a \geq 0.25$; $\lvert a-1\rvert\times$ original path ≥ 0.15 m.
Check dynamics and framing	Active only: speed/acceleration/jerk p95 ≤ 0.1416 m/frame, 0.03499 m/frame^2, 0.01315 m/frame^3. Human distance $\in [0.25, 12.7716]$ m; visibility decrease ≤ 0.05; severe-outscreen increase ≤ 0.03.
Check reconstruction	Active: construction-representation decoded path-ratio error ≤ 0.15.

Construction checks. Table S4 groups the acceptance criteria. Dynamics limits use the 99.5th percentiles of a 1,000-clip training-data pilot, in per-frame units. A joint is visible when its projection is finite, lies inside the image and has camera-space depth $> 10^{-4}$ m. DPA visibility averages this indicator over joints and frames; outscreen is its complement. IPA visibility counts frames with any visible joint; severe-outscreen counts frames with fewer than half the joints visible. The decoded IPA gate uses the construction representation, and null targets remain unchanged.

D Evaluation and Baseline Protocols

D.1 Baseline protocols

We follow the evaluation setup in Section 4.1 and the AESOP training schedule in Table S3. Given-human evaluation supplies reference human motion; joint evaluation uses each method's generated motion. Table S5 lists adaptations and sampling configurations.

Table S5: Baseline inputs, conditioning and sampling. For PulpMotion, the tuple denotes conditional-text guidance, autoregressive-context guidance, and auxiliary projection guidance, respectively. A negative autoregressive-context weight selects standard CFG; zero auxiliary projection weight disables that guidance term. MAR generates autoregressively, with a diffusion sampler inside each iteration. EDM, DDIM and DDPM follow Karras et al. (2022), Song et al. (2021) and Ho et al. (2020), respectively.

Method	Inputs and conditioning	Sampling and guidance
DIRECTOR-C	Camera text and human root trajectory	EDM: 10 Euler/Heun steps; camera CFG 1.4
DanceCamera3D	Camera text and full human motion; text replaces music	DDIM: 50 steps, $\eta = 1$; human/camera CFG 1.75/1
CCD	Camera text; text-only network	DDPM: 1,000 steps; camera CFG 2
CCD-H	Camera text and human joints; text cross-attention	DDPM: 1,000 steps; camera CFG 2
PulpMotion DiT	Human and camera text; symmetric representation	DDPM: 50 steps; $(g_c, g_m, g_z) = (11, -1, 0)$
PulpMotion MAR	Human and camera text; symmetric representation	18 autoregressive iterations, each with 50 DDPM steps; $(g_c, g_m, g_z) = (3.5, 2, 0)$
AESOP	Human motion and camera text (given); human text and camera text (joint)	Euler: 50 steps per stream; human/camera CFG 1/1.5

Generative training presents 13.44M examples to each camera-only baseline and 26.88M to each PulpMotion joint model. AESOP presents 13.44M examples to each human/camera flow, followed by 6M camera continuation examples. These counts include repeated presentations and both members of control pairs; Table S3 gives AESOP's phase durations and batch sizes.

Human-relative camera decoding. CCD and CCD-H decode camera position relative to the supplied human head in a heading-aligned frame, with a head-directed look-at rotation adjusted by predicted screen position (Jiang et al., 2024). AESOP and PulpMotion anchor camera position to the human root at the first frame and accumulate world-space increments (Courant et al., 2026).

D.2 Direction-paired training

Paired sampling and scoring. Each prompt arm contains the same 3,194 clips with active reference translation, giving 6,388 clip–prompt cases per task. Original and reversed prompts are processed by the same text encoder; human latents, camera noise and valid-length masks are shared within each pair. Given-human uses reference motion; joint generation samples the human first.

Direction accuracy measures the sign of each reference-active translation axis at fixed reference time positions. Each clip first averages sign correctness over those positions and axes, then clips receive equal weight. Reversed targets negate the corresponding reference signs. Reference-inactive positions are excluded; generated still predictions at active positions count as errors. Combined accuracy averages the two prompt arms.

Table S6: Direction accuracy (%). DPA and factual continuations share the starting checkpoint, 35K updates, batch allocation and sampling schedule; factual pairs replace directional counterfactuals in the control. Bold marks the best mean in each column. Section D.2 defines paired sampling and scoring.

	Given-human generation			Joint generation		
Checkpoint	Original↑	Reversed↑	Combined↑	Original↑	Reversed↑	Combined↑
Camera original	$86.38^{\pm 0.86}$	$75.66^{\pm 0.68}$	$81.02^{\pm 0.15}$	$83.15^{\pm 0.13}$	$76.42^{\pm 0.43}$	$79.78^{\pm 0.28}$
35K factual-pair control	$87.60^{\pm 0.43}$	$77.71^{\pm 0.13}$	$82.66^{\pm 0.17}$	$84.95^{\pm 1.13}$	$78.51^{\pm 0.17}$	$81.73^{\pm 0.51}$
+ DPA	$\mathbf{89.66}^{\pm \mathbf{0.67}}$	$\mathbf{79.00}^{\pm \mathbf{0.22}}$	$\mathbf{84.33}^{\pm \mathbf{0.28}}$	$\mathbf{86.63}^{\pm \mathbf{0.77}}$	$\mathbf{79.58}^{\pm \mathbf{0.50}}$	$\mathbf{83.11}^{\pm \mathbf{0.14}}$

E Translation-Intensity Control

For the endpoint change from $a = 1$ to 1.5, path length is nondecreasing for 97.31% of given-human test outputs and 97.01% of joint test outputs. Anonymous visualizations show intensity changes on selected HumanML3D examples in the given-human and joint settings.

F User Study

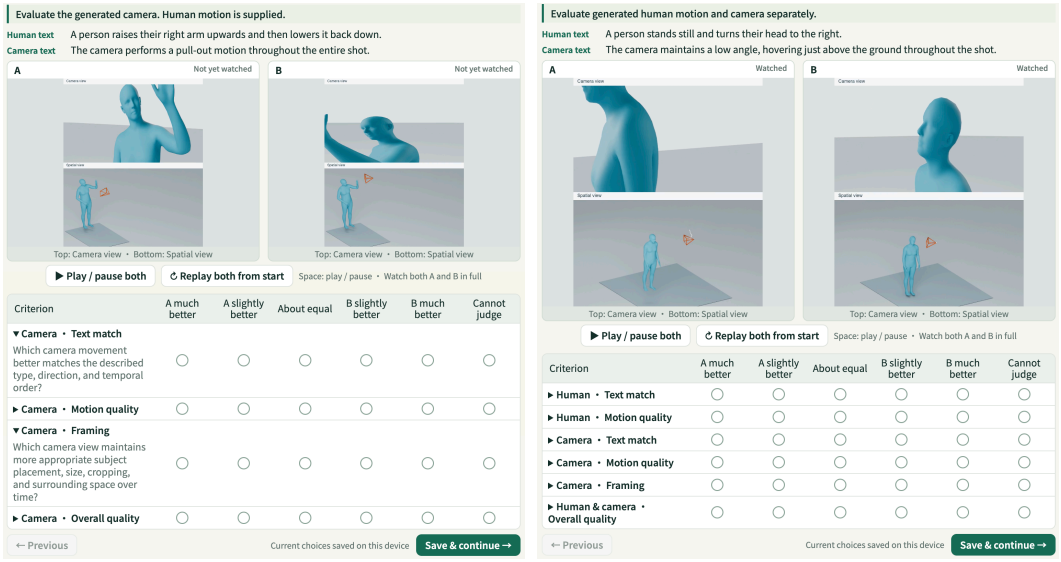
Study design. The anonymous study follows the task allocation in Section 4.4 and includes baseline–baseline comparisons. Each questionnaire receives a randomized, balanced assignment with equal target exposure per method pair, balanced task order, and randomized left–right presentation and trial order. Figure S4 shows synchronized camera and spatial views.

Evaluation criteria. Both tasks assess camera-text alignment, camera motion quality and framing. Given-human trials additionally ask for overall camera quality; joint trials add human-text alignment, human motion quality and overall human–camera quality. Each criterion uses five responses from "A much better" to "B much better," including a tie, plus "Cannot judge." Overall quality is judged directly. Both videos must reach 90% playback coverage before a trial can be saved.

Analysis. For each clip, method pair and criterion, we compute preference as $(W + 0.5T)/(W + T + L)$, where W, T and L count wins, ties and losses. "Cannot judge" responses are counted separately. Within each task and criterion, pairwise scores average equally over clips, and method-level scores average equally over opponents. Overall quality is the primary outcome for each task. We estimate 95% confidence intervals using 10,000 participant-cluster bootstrap resamples with the stimulus set fixed.

Results. The 29 complete questionnaires comprise 580 given-human and 522 joint trials, covering all 120/60 clip–pair comparisons. Of 5,452 criterion responses, 5,238 are judgeable and 214 are "Cannot judge."

Figure S5 reports all task-specific criteria. AESOP has the highest preference on each criterion: joint human-text alignment and motion-quality scores are 80.4% and 80.9%, while camera alignment, motion-quality and framing scores range from 87.5% to 89.9%. Section 4.4 and Table 3 report overall preference.



(a) Given-human camera generation **(b) Joint human–camera generation**

Figure S4: **User-study interface.** Anonymized A/B videos show the generated camera view above an external spatial view. Participants compare text alignment, motion quality, framing and task-specific overall quality. The joint task additionally evaluates human motion.

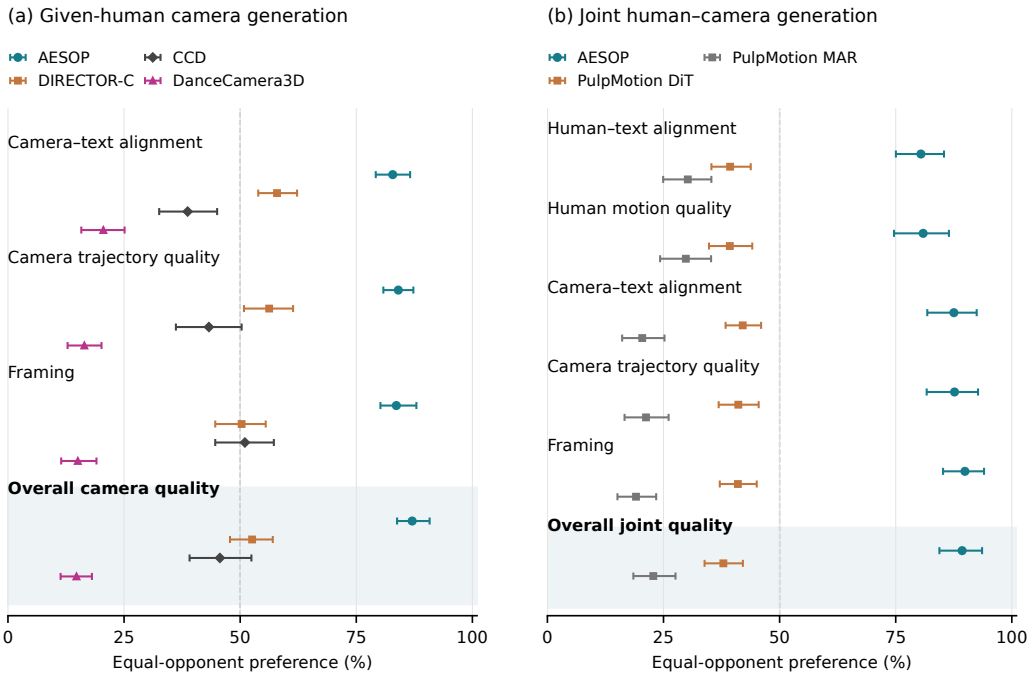


Figure S5: **User-study preferences across all criteria.** Points show clip-balanced, equal-opponent preference scores; bars show 95% participant-cluster bootstrap intervals. The vertical dashed line marks 50% preference. Shading highlights the separately rated overall-quality endpoints.

G QUALITATIVE RESULTS ON EXTERNAL MOTIONS AND PROMPTS

We demonstrate the same PulpMotion-trained AESOP model in three settings: camera generation from HumanML3D motions, joint generation from HumanML3D action descriptions, and camera generation from externally synthesized human motions.

Ground-truth human motions. We adapt HumanML3D (Guo et al., 2022) motions to the human representation and generate cameras conditioned on these motions and camera prompts. Figure S6 shows a static camera observing running in place, a pull-out widening the view of a walking and turning figure, and trucking accompanying jumping and spinning. Videos are available in the anonymous HumanML3D given-human visualization.

Joint generation from action descriptions. We use ground-truth HumanML3D action descriptions to generate both human motion and camera trajectory (Figure S7). A pull-in tightens the view of arm stretching, while a static camera observes an approaching runner. Videos are available in the anonymous HumanML3D joint visualization.

Synthesized human motions. We condition the same camera generator on motions synthesized by HY-Motion. Figure S8 shows static views for bowing and rhythmic stepping, a pull-out for a surfboard pop-up, and a pull-in for stepping back and blocking. Videos are available in the anonymous HY-Motion given-human visualization.

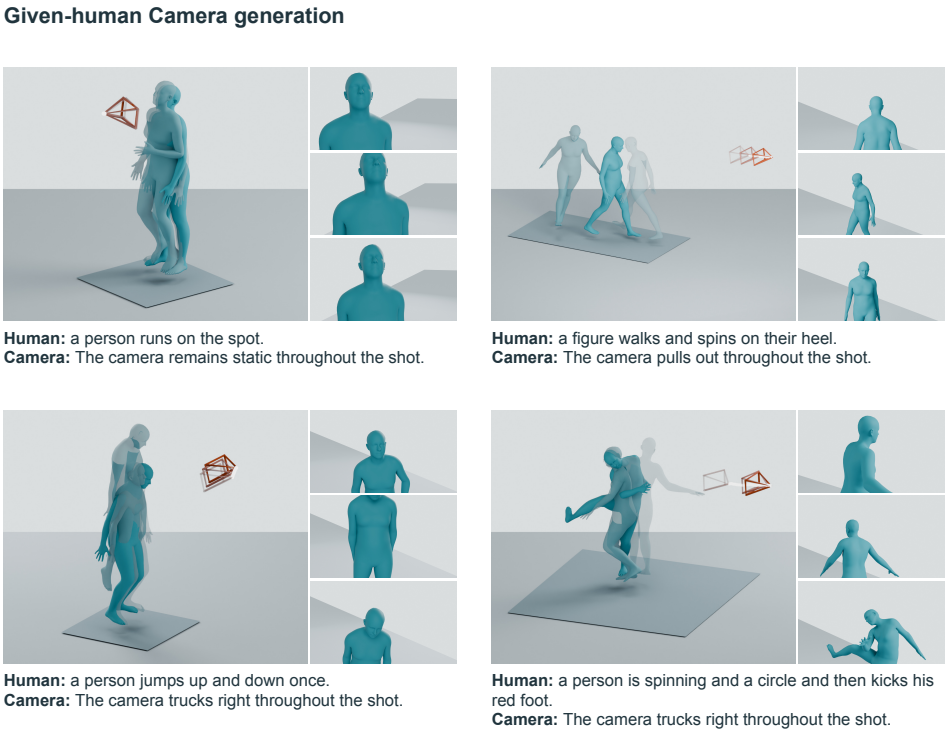
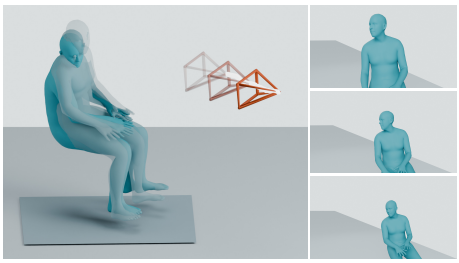
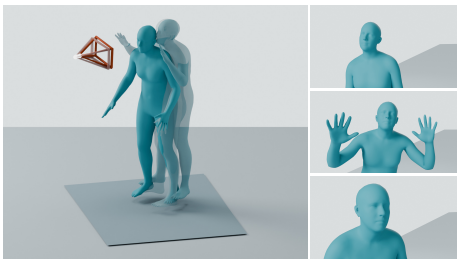


Figure S6: **Camera generation conditioned on HumanML3D motions.** Each example shows a global view and three successive camera views; lighter poses indicate earlier frames. Global views are fitted separately to each scene.

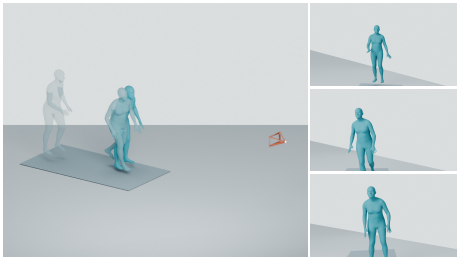
Joint Human–Camera generation



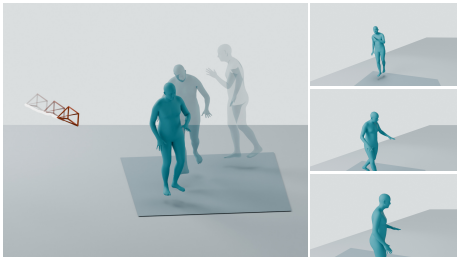
Human: the man sits on a stool and looks around.
Camera: The camera pulls out throughout the shot.



Human: person stretches both arms up and then put arms down.
Camera: The camera pulls in throughout the shot.



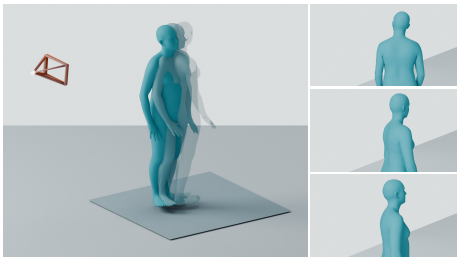
Human: a person runs forward and stops short.
Camera: The camera remains static throughout the shot.



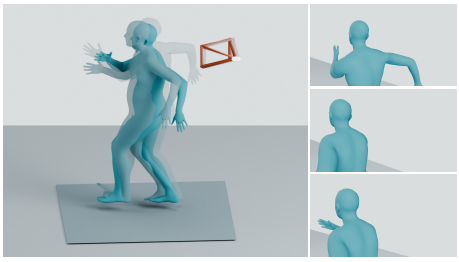
Human: a person walks three steps to his right, then five steps to his left, and finally three steps to his right.
Camera: The camera trucks right throughout the shot.

Figure S7: **Joint generation from HumanML3D action descriptions.** Each example shows generated human motion and camera trajectory in a global view and three successive camera views.

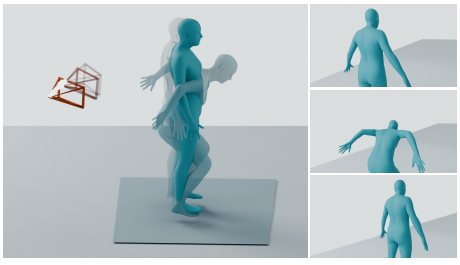
Camera generation from HY-Motion outputs



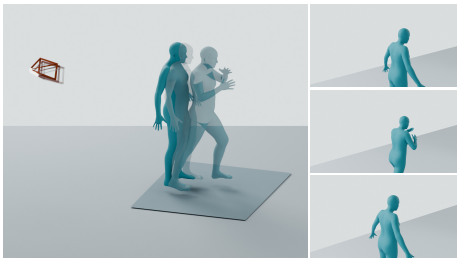
Human: a person bows down facing the front then bows down facing the right.
Camera: The camera remains static throughout the shot.



Human: A person swings their arms alternately to the rhythm while stepping.
Camera: The camera remains static throughout the shot.



Human: A person does a pop-up on a surfboard.
Camera: The camera pulls out throughout the shot.



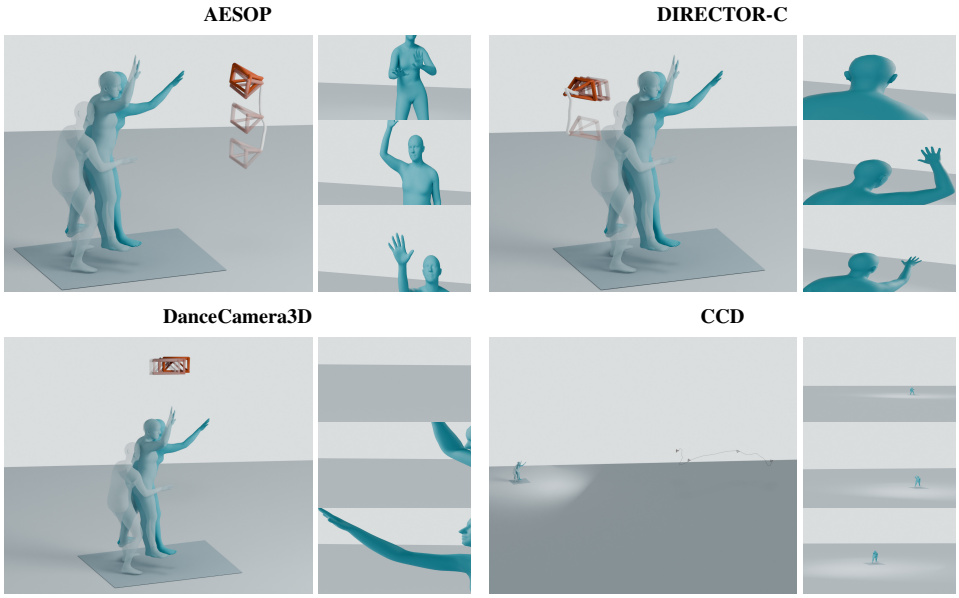
Human: A person steps back with their left foot, then blocks with their right arm.
Camera: The camera pulls in throughout the shot.

Figure S8: **Camera generation conditioned on HY-Motion outputs.** Four synthesized human motions are paired with static, pull-out and pull-in camera prompts. Each example includes a global view and three successive camera views.

H MORE QUALITATIVE RESULTS

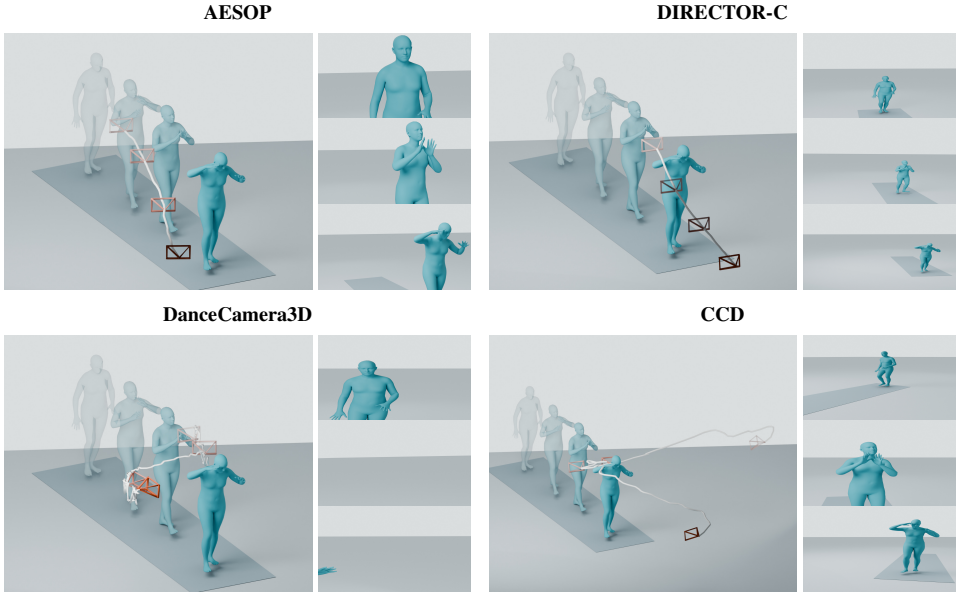
Figures S9 and S10 show two additional examples for each task. Each method shows a global view and three camera projections in temporal order from top to bottom. Anonymous comparisons are available for given-human camera generation and joint human–camera generation.

(a) Example 1



human: A person bends forward, lifts their right leg, and then extends their right arm upwards.
camera: The camera starts with a boom top shot.

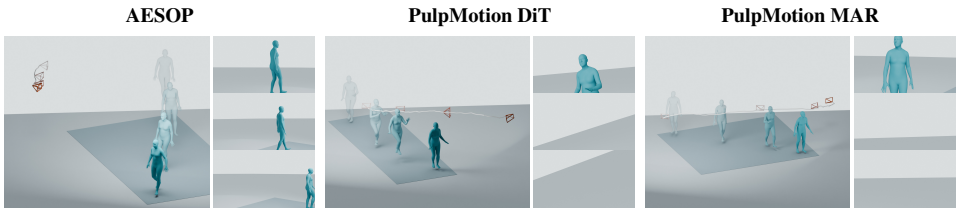
(b) Example 2



human: A person walks forward along the sidewalk, maintaining a steady pace.
camera: The camera performs a pull-out motion throughout the entire shot.

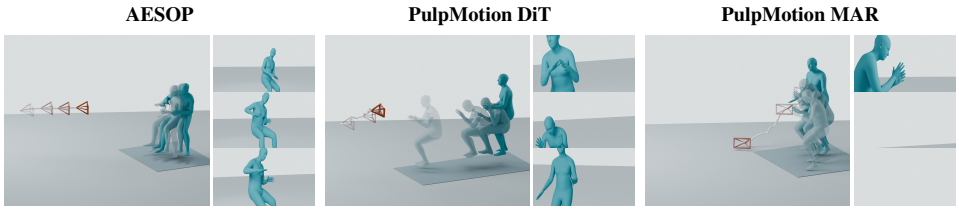
Figure S9: **Additional given-human camera generation examples.** Global views are fitted separately to each trajectory.

(a) Example 1



human: A person walks forward, swinging their arms naturally.
camera: The camera moves right along a dolly or truck during the entire shot.

(b) Example 2



human: A person sits on a stool, moves their left hand towards their face, and then places it back on the instrument.
camera: The camera gradually moves closer to the subject throughout the shot.

Figure S10: **Additional joint human–camera generation examples.** Global views are fitted separately to each trajectory.