CAT2M: CAMERA-AWARE TEXT TO MOTION GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent world models have achieved notable progress in synthesizing realistic videos with controllable camera movements; however, achieving fine-grained human motion control under dynamic viewpoints remains a significant challenge. Meanwhile, existing text-to-motion methods typically assume fixed camera settings, which limits their applicability in world model settings where human motion should be consistent with the text description under changing camera viewpoints. To address this gap, we introduce a new task: *camera-aware text-to-motion generation*, where human motion is generated based on both text descriptions and given camera trajectories. To facilitate this task, we construct the **CAT-Motion** dataset, a large-scale dataset with over 100K samples of paired human motion and camera trajectories, each annotated with rich captions including viewpoint-specific motion descriptions. Based on this dataset, we propose **CAT2M**, a unified framework for motion generation with text and camera trajectory conditions. A shared coordinate frame anchored to the first camera preserves human-camera relative geometry, while temporal camera cross attention conditions generation on the changing viewpoint. We further introduce differentiable screen-space supervision through a cumulative training scheme that retains diffusion and 3D joint objectives, constraining the projected appearance alongside the underlying 3D motion. Extensive experiments demonstrate that our approach outperforms existing baselines, showing strong ability in producing human motions that are semantically accurate and geometrically consistent under camera constraints and text descriptions.

1 INTRODUCTION

With recent advances in video generation (Brooks et al., 2024; He et al., 2024; Wang et al., 2024a; Zheng et al., 2024; Guo et al., 2023; Yu et al., 2024; Cao et al., 2025; Hu, 2024) and world models (Bruce et al., 2024; Team et al., 2026; Zheng et al., 2026; Mao et al., 2025; Huang et al., 2025; Wu et al., 2025), controllable synthesis of dynamic visual scenes has become increasingly practical. Generating human behavior that follows user intent is an important goal in these settings, requiring control over both what a person does and how the motion appears from the intended viewpoint. Descriptions such as "approaching the camera," "leaving the frame," or "staying centered on screen" depend on the relative geometry of the person and the camera: the same 3D motion can produce different apparent directions, framing, and body visibility under different viewpoints. A motion may therefore perform the intended action in 3D yet fail to exhibit the specified behavior within the frame. However, camera control and human motion generation are still predominantly studied separately, leaving a key question insufficiently addressed: how can we generate articulated 3D human motion that realizes a textual description under a prescribed camera trajectory?

Existing motion generation methods seldom address this challenge. Standard text-to-motion models (Zhu et al., 2025; Guo et al., 2022a; 2024; 2022b) generate body motion without a prescribed camera sequence as an explicit condition, whereas the few methods that incorporate camera motion generally treat it as an output modality (Courant et al., 2025) rather than as an external control signal for human motion generation. Consequently, they cannot directly support the generation of human motion that is simultaneously aligned with textual intent and controllable under prescribed camera viewpoints.

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.

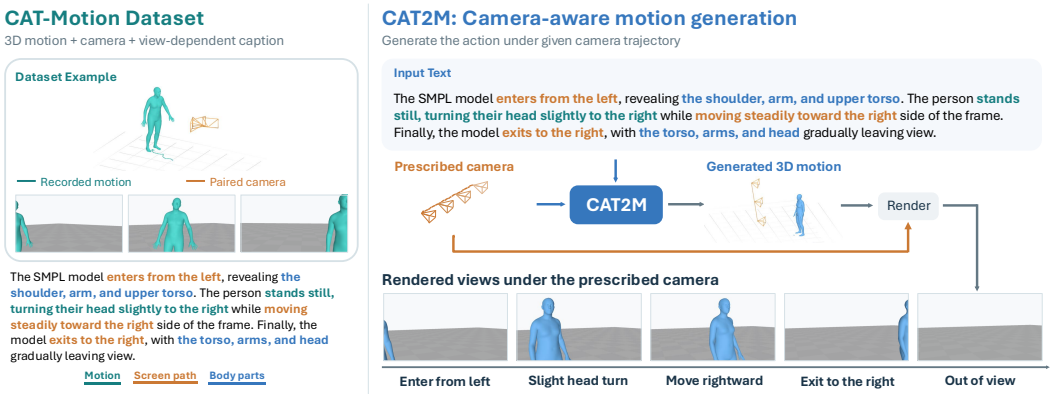


Figure 1: **Overview of our work**. CAT-Motion pairs 3D human motions with camera trajectories and view-dependent captions, jointly capturing human motion, on-screen movement, and body-part visibility. Given a short caption and a prescribed camera trajectory, CAT2M generates 3D motion to realize both the described action and its intended appearance in the frame.

To fill this gap, we introduce a new task termed *camera-aware text-to-motion generation*, which aims to generate 3D human motion conditioned on both textual descriptions and given camera trajectories, thereby enabling explicit and coordinated control of human dynamics and camera viewpoints. However, existing datasets primarily pair text with human motion (Guo et al., 2022a; Plappert et al., 2016; Lin et al., 2023; Zhang et al., 2025b) or emphasize camera trajectories (Zhou et al., 2018; Courant et al., 2024; Zhang et al., 2025a), with limited descriptions linking human actions to their appearance under specific viewpoints. This lack of coupled annotations limits supervision for learning camera-aware motion. Meanwhile, standard text-to-motion methods typically omit camera conditions. Although PulpMotion (Courant et al., 2025) explores joint human-camera modeling, it does not focus on text-driven control of observed human behavior under prescribed camera trajectories. Consequently, plausible 3D motion may still produce on-screen behavior inconsistent with textual intent, leaving camera-aware human motion generation underexplored.

To provide viewpoint-specific supervision, we construct **CAT-Motion**, containing over 100K paired human-motion and camera-trajectory samples. Our structured captioning pipeline describes entering, in-frame motion, and leaving stages, capturing screen position, motion direction, and visible body parts. These stage-wise annotations are combined with action descriptions to produce sequence-level captions that connect human actions to their visual manifestation under the corresponding camera trajectories.

Building on this dataset, we propose **CAT2M** for generating 3D human motion from text and prescribed camera trajectories. We represent human motion and cameras in a shared coordinate frame anchored to the first camera to preserve their relative geometry, and use temporal camera cross attention to condition generation on the changing viewpoint. To constrain how the generated motion appears within the frame, we introduce differentiable screen-space supervision through cumulative training that retains the diffusion and 3D joint objectives. Together, these designs connect the underlying 3D motion to its observed appearance. Experiments on the CAT-Motion dataset demonstrate improved condition-motion alignment and competitive camera-aware alignment, while ablations examine the contributions of camera conditioning, geometric objectives, and motion representation.

Our main contributions are as follows: **(1)** We introduce a novel task of *camera-aware text-to-motion generation*, where human motion is generated conditioned on both textual guidance and given camera trajectories. **(2)** We construct the **CAT-Motion** dataset, a large-scale dataset with over 100K samples of paired human motion and camera trajectories, with each sample labeled with multi-faceted captions describing both the isolated human motion and the human-camera relationship. **(3)** We propose **CAT2M**, a new framework for camera-aware text-to-motion generation, combining a shared camera-referenced motion representation, temporal camera cross attention, and differentiable screen-space supervision in a cumulative training scheme that retains 3D geometric objectives.

2 Related Work

2.1 Text to Motion Generation

Text-to-motion (T2M) generation synthesizes realistic human motion from natural language, providing intuitive control over motion semantics. Early approaches map text to motion using generative adversarial networks (Ahn et al., 2017; Goodfellow et al., 2020) and encoder-decoder architectures (Ahuja & Morency, 2019; Ghosh et al., 2023), often by learning shared text-motion embeddings. Subsequent variational autoencoders (Guo et al., 2022a) improve motion diversity and temporal coherence, while extensions (Athanasiou et al., 2022; Petrovich et al., 2022; Tevet et al., 2022a) incorporate attention (Vaswani et al., 2017) and multi-sequence conditioning to enhance generation quality. Inspired by text-to-image generation, recent work explores discrete tokens and diffusion (Guo et al., 2022b; Kim et al., 2023; Zhang et al., 2023). VQ-VAE (van den Oord et al., 2018) combined with autoregressive transformers generates motion in a language-like manner, while MDM (Tevet et al., 2022b) and related diffusion models (Chen et al., 2023) improve realism, diversity, and text alignment. Further advances investigate efficient latent representations (Petrovich et al., 2021; Rombach et al., 2022), accelerated sampling (Lu et al., 2022; Song et al., 2020), and pretrained models for stronger semantic understanding (Radford et al., 2021; Tevet et al., 2022a). Hybrid autoregressive-diffusion paradigms (Li et al., 2024; Meng et al., 2025) pursue better generation quality and controllability, alongside broader efforts toward unified multimodal understanding and generation (Jiang et al., 2023; Park et al., 2025; Wang et al., 2025; Wu et al., 2024). However, existing datasets (Courant et al., 2025; Guo et al., 2022a; Lin et al., 2023; Plappert et al., 2016; Zhang et al., 2025b) provide text-motion pairs without explicit captions describing human-camera relationships. This limits direct supervision for generating motions whose appearance under a given camera follows viewpoint-dependent descriptions.

2.2 Camera Control and Trajectory Modeling

Camera control and trajectory modeling have been widely studied for filmmaking, virtual environments, and video generation. Early approaches (Bares et al., 2000; Christie & Languénou, 2003; Christie & Olivier, 2009) use handcrafted rules or 3D optimization to enforce cinematic constraints such as composition, visibility, and smoothness. Their dependence on carefully designed objectives limits scalability and generalization. Learning-based methods (Courant et al., 2024; Gschwindt et al., 2019; Huang et al., 2021; Jiang et al., 2021; 2024) instead learn cinematic camera behaviors from example videos or structured representations, while reinforcement learning (Bonatti et al., 2020; Huang et al., 2019; Xie et al., 2023) learns camera policies for specific environments. Character-centric methods (Wang et al., 2024b;c) account for the visibility of human body joints, particularly when generating camera movements for dance performances. More recent work supports diffusion-based trajectory generation (Jiang et al., 2024) and text-conditioned camera control (Zhang et al., 2025a), reducing reliance on curated reference videos. These approaches principally generate camera motion, rather than directly controlling the articulated human motion observed through the camera. In parallel, camera-controlled video generation (Bahmani et al., 2025; Guo et al., 2023; He et al., 2024; 2025; Wang et al., 2024a) incorporates camera parameters to steer synthesized viewpoints. Such methods focus on visual appearance and scene-level consistency, while human motion often remains implicit or weakly modeled. Thus, advances in camera planning and video synthesis do not directly resolve the task of generating articulated 3D human motion that follows text under a prescribed camera trajectory. Our work takes the camera as a given condition and generates the human motion, linking what is performed to how it is observed.

3 CAT-Motion Dataset

3.1 Dataset Description

CAT-Motion pairs 3D human motions with camera trajectories, intrinsics, and captions describing how the motions appear from the corresponding viewpoints. It contains over 100K paired samples, totaling 9.53M frames across camera variants. Each variant includes a source action description and camera-aware captions. Table 5 compares CAT-Motion with existing motion and camera datasets.

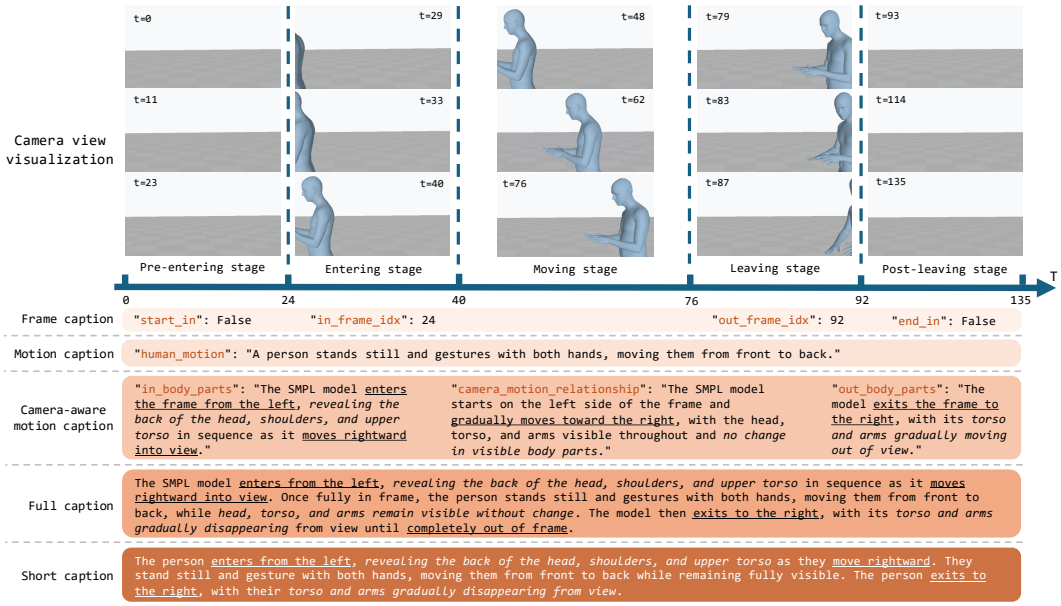


Figure 2: **An example from CAT-Motion.** Stage-wise captions describe entry, in-frame motion, and exit, capturing movement direction and body-part visibility. These descriptions are combined with the source motion caption and visibility metadata to form a full caption, which is then condensed into a short caption used as the text condition for motion generation.

3.2 Dataset Construction

Camera trajectory curation. We build our dataset upon the PulpMotion dataset (Courant et al., 2025), which provides paired human motion and camera trajectories. Existing human–camera datasets rarely include scenarios in which the subject enters or exits the frame, while such cases are common in real-world settings and critical for generation tasks. To address this gap, we redesign the camera trajectories to emphasize such situations, resulting in 91% of the samples involving the subject entering and/or leaving the frame during the sequence. We design five trajectory augmentation strategies to transform the original camera trajectories, with detailed descriptions provided in the supplementary materials.

Camera-aware motion annotation. We design a structured annotation pipeline to generate detailed captions describing the relationship between human motion and camera viewpoints using the Qwen3-VL (Bai et al., 2025) model. Each video sequence is segmented into three segments based on human visibility. We first identify the entry and exit timestamps of the human, and divide the full sequence into three segments: (1) an entry segment consisting of the first N_{entry} frames after the human appears, (2) an exit segment consisting of the last N_{exit} frames before the human disappears, and (3) a middle segment containing all remaining frames between the entry and exit segments. Each segment is captioned independently with different prompt instructions tailored to its temporal role. These stage-wise descriptions are then combined with the source action caption into a chronological full-sequence description. We retain the detailed captions and additionally provide concise versions as CAT2M's text conditions. This hierarchical annotation strategy improves both fine-grained temporal descriptions and overall narrative consistency. An example of our annotated data is shown in Figure 2.

Human pose representation. Camera-aware generation requires preserving human-camera relative geometry, rather than an arbitrary world coordinate system. However, existing human motion representations (Guo et al., 2022a) are typically normalized to a canonical coordinate system, where the human is aligned to a common ground plane and faces a fixed forward direction. Such normalization removes global orientation and translation information, making them unsuitable for our task. To address this limitation, we adopt a 147-dimensional human pose representation that preserves

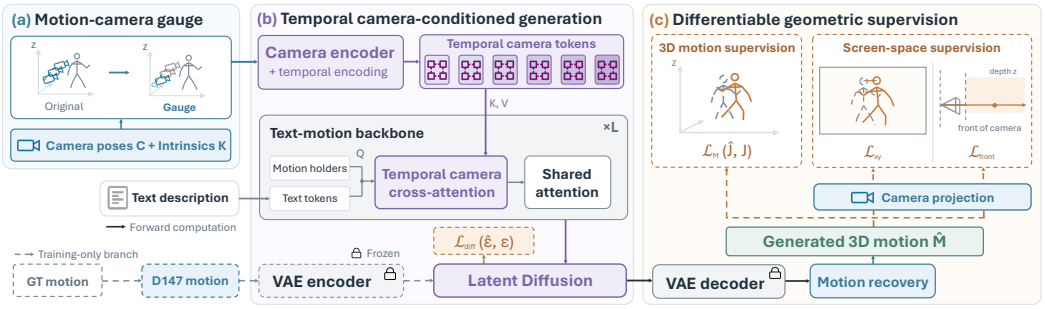


Figure 3: **Overview of our CAT2M framework.** Human motion and camera trajectories are expressed in a shared gravity-preserving first-camera coordinate system. Camera parameters are then encoded into temporal tokens that condition the text-motion backbone through cross attention, guiding latent diffusion to generate motion from text and prescribed cameras. Generated latents are decoded through a frozen VAE and recovered as 3D motion. During training, 3D joint supervision, projected joint alignment, and a front-of-camera depth constraint jointly encourage geometric and screen-space consistency.

global spatial information. We jointly transform human motion and camera poses into a fixed, gravity-preserving coordinate frame defined by the first prescribed camera. This frame remains unchanged throughout the sequence, removing arbitrary global translation and horizontal orientation while preserving the relative geometry. Each motion frame is encoded by a 147-dimensional SMPL (Loper et al., 2015) vector,

$$m_t = [r_t, q_t^0, \ldots, q_t^{23}] \in \mathbb{R}^{147}, \tag{1}$$

where $r_t \in \mathbb{R}^3$ is the absolute root translation, $q_t^0 \in \mathbb{R}^6$ encodes the root orientation in the shared frame, and $q_t^1, \ldots, q_t^{23} \in \mathbb{R}^6$ encode local joint rotations relative to their parents. By retaining both global position and motion dynamics information, the proposed representation enables the model to capture fine-grained human–camera spatial relationships and generate camera-aware human motion.

4 Camera-Aware Text to Motion Generation

4.1 Problem Formulation

The proposed new task aims to generate human motion conditioned on a camera trajectory and a text description of the motion observed within the frame. Formally, given a text description $\mathcal{D}$, a camera-to-world pose sequence $\mathcal{C} = \{c_t\}_{t=0}^{T-1}$, and camera intrinsics $I = [f_x, f_y, c_x, c_y]$, we generate a motion sequence $\mathcal{M} = \{m_t\}_{t=0}^{T-1}$ whose appearance under the prescribed cameras matches the description. We use the D147 representation and shared, gravity-preserving first-camera coordinate frame described in Section 3.2, with $m_t = [r_t, q_t] \in \mathbb{R}^{147}$. Here, $c_t \in \mathbb{R}^{3\times4}$ and r_t are expressed in this shared frame. The generation process can be formulated as sampling from the conditional distribution $p(\mathcal{M} \mid \mathcal{C}, \mathcal{D}, I)$.

4.2 Text-Motion Framework

VAE training. To bridge the gap between continuous motion representations and the autoregressive generation paradigm, we first train a variational autoencoder (VAE) (Kingma & Welling, 2013; Chen et al., 2023) to compress motion sequences into a compact latent space. Specifically, given a D147 representation motion sequence of T frames $\mathcal{M} = \{m_0, m_1, \ldots, m_{T-1}\}$, where $m_t \in \mathbb{R}^{147}$, the VAE encoder $\mathcal{E}_{\text{VAE}}$ maps the motion representations into a latent representation $z \in \mathbb{R}^n$. This latent representation serves as the target space for motion generation during training, while the VAE decoder $\mathcal{D}_{\text{VAE}}$ is used to reconstruct the final motion output from the generated latent.

Conditional latent generation. We adopt a bimodal architecture to jointly model language and motion (Zhu et al., 2025). We use text and motion streams with modality-specific processing and

shared attention. The text stream is initialized from pretrained GPT-2, and the resulting condition states $h = F_\theta(\mathcal{D}, \mathcal{C}, I)$ guide a diffusion head (Li et al., 2024). For a VAE latent z, diffusion operates on $z_0 = sz$, where s is the latent scale associated with the pretrained VAE. With diffusion timestep τ, the noise-prediction objective is

$$\begin{aligned} z_\tau &= \sqrt{\bar{\alpha}_\tau}\, z_0 + \sqrt{1-\bar{\alpha}_\tau}\, \epsilon, \qquad \epsilon \sim \mathcal{N}(0, \mathbf{I}), \\ \mathcal{L}_{\text{diff}} &= \mathbb{E}_{z_0,\epsilon,\tau}\left[\text{MSE}(\epsilon, \epsilon_\theta(z_\tau, \tau, h))\right], \end{aligned} \tag{2}$$

where $\bar{\alpha}_\tau$ is the cumulative noise schedule. At inference, the generated latent is divided by s, decoded by the VAE, and transformed back through $\mathcal{N}^{-1}$.

4.3 Temporal Camera Conditioning

To enable our model to generate camera-aware human motion, it is necessary to incorporate both camera extrinsics and intrinsics as conditioning signals. Camera frames and language tokens have different temporal structures. Rather than interpolating the camera sequence to token positions and adding the features pointwise, we let token states attend to the prescribed camera sequence. A linear camera encoder projects each 16-dimensional vector, formed by the flattened pose and scaled intrinsics, to the token embedding dimension. Temporal encoding based on the valid sequence length provides time information, yielding camera tokens $E = \{e_t\}_{t=0}^{T-1}$. Before shared attention, we apply

$$H_c = H + \text{CA}(H, E), \tag{3}$$

where H denotes token features. Camera cross attention CA uses token states as queries and camera features as keys and values, without additional learnable QKV projections. Padding is masked, while all valid camera frames remain accessible because the full trajectory is prescribed. This provides time-aware, content-dependent conditioning without imposing a one-to-one token-frame alignment. Pixel intrinsics are retained separately for geometric projection.

4.4 Differentiable Screen-Space Supervision

Camera conditioning supplies viewpoint information, but does not itself define a loss on the observed outcome. We therefore complement 3D joint supervision with losses on the projected motion. From the same diffusion training step, we recover a clean latent prediction and its 3D joints:

$$\begin{aligned} \widehat{z}_0 &= \frac{z_\tau - \sqrt{1-\bar{\alpha}_\tau}\, \epsilon_\theta(z_\tau, \tau, h)}{\sqrt{\bar{\alpha}_\tau}}, \\ \widehat{\mathcal{M}} &= \mathcal{N}^{-1}(\mathcal{D}_{\text{VAE}}(\widehat{z}_0/s)), \qquad \widehat{J} = \text{FK}_{\text{SMPL},22}(\widehat{\mathcal{M}}; \beta = 0). \end{aligned} \tag{4}$$

The decoder parameters remain frozen, but gradients through decoding, root reconstruction, and forward kinematics reach the diffusion head and condition network. For a sequence with valid frame set $\mathcal{T}$, we supervise the recovered 3D joints using

$$\mathcal{L}_{\text{MPJPE}} = \frac{1}{N|\mathcal{T}|} \sum_{t \in \mathcal{T}} \sum_{j=1}^{N} \left\| \widehat{J}_{t,j} - J_{t,j} \right\|_2, \tag{5}$$

where $N = 22$ is the number of joints, and $J = \text{FK}_{\text{SMPL},N}(\mathcal{M}; \beta = 0)$ denotes the reference joints in the same frame. We further supervise their projected appearance. Let $\widehat{u}_{t,j}$ and $u_{t,j}$ denote predicted and ground-truth joint projections in normalized device coordinates (NDC) under the same camera c_t and pixel intrinsics I, and let $\widehat{d}_{t,j}$ denote predicted camera depth. We define

$$\begin{aligned} \mathcal{L}_{\text{xy}} &= \langle \rho_\delta(\|\widehat{u}_{t,j} - u_{t,j}\|_2) \rangle_\Omega, \\ \mathcal{L}_{\text{front}} &= \left\langle b_{m,\vartheta}(\widehat{d}_{t,j}) \right\rangle_\Omega, \end{aligned} \tag{6}$$

where ρ_δ is the Huber loss with $\delta = 0.1$, and $b_{m,\vartheta}$ is the softplus depth barrier with margin $m = 0.10\,\text{m}$ and temperature $\vartheta = 0.05\,\text{m}$. The eligible set Ω is determined solely by ground truth: valid frames, finite camera-space joint coordinates, and positive ground-truth depth. We use all 22 joints and do not require them to lie inside the image. Predicted out-of-frame or behind-camera points are not excluded. The operator $\langle \cdot \rangle_\Omega$ first averages within each sequence, then across sequences with eligible points.

Table 1: Comparison to baseline methods.

Method	Condition-Motion Alignment				Camera-Aware Alignment		
	Bi-R@1 ↑	Bi-R@3 ↑	$\mathrm{FD}_{\mathrm{motion}}$ ↓	MMDist ↓	$\mathrm{FD}_{\mathrm{frame}}$ ↓	ProjE ↓	LLM Score ↑
PulpMotion-Cond	3.57%	8.29%	0.289	0.964	0.578	4.004	7.62
MDM-Cam	1.57%	3.78%	0.417	1.088	2.375	17.578	7.40
MoMask-Cam	2.05%	4.97%	0.183	1.038	1.166	4.720	7.54
MLD-Cam	22.97%	39.46%	0.083	<u>0.752</u>	<u>0.095</u>	<u>3.285</u>	<u>7.96</u>
MotionGPT3-Cam	<u>23.36%</u>	<u>39.55%</u>	<u>0.064</u>	0.756	0.190	3.787	7.06
CAT2M (ours)	**29.80%**	**49.63%**	**0.055**	**0.710**	**0.019**	**1.120**	**8.43**

Table 2: Camera conditioning and training loss ablations of CAT2M.

Variant	Conditions/losses				Condition-Motion Alignment				Camera-Aware Alignment		
	Gauge	TCA	$\mathcal{M}$	$\mathcal{S}$	Bi-R@1 ↑	Bi-R@3 ↑	$\mathrm{FD}_{\mathrm{motion}}$ ↓	MMDist ↓	$\mathrm{FD}_{\mathrm{frame}}$ ↓	ProjE ↓	LLM Score ↑
Raw-Add	–	–	✓	✓	12.94%	25.74%	0.063	0.798	0.042	1.494	8.01
Gauge-Add	✓	–	✓	✓	27.90%	47.36%	0.057	0.719	<u>0.022</u>	1.386	8.35
TCA Only	–	✓	✓	✓	21.10%	36.76%	0.060	0.749	0.025	<u>1.192</u>	8.26
w/o $\mathcal{M}, \mathcal{S}$	✓	✓	–	–	25.19%	42.51%	0.077	0.737	0.081	3.130	8.08
w/o $\mathcal{M}$	✓	✓	–	✓	25.85%	43.34%	0.078	0.733	0.075	2.007	8.30
w/o $\mathcal{S}$	✓	✓	✓	–	<u>29.27%</u>	<u>48.79%</u>	**0.054**	<u>0.712</u>	0.023	1.644	<u>8.36</u>
CAT2M (Ours)	✓	✓	✓	✓	**29.80%**	**49.63%**	<u>0.055</u>	**0.710**	**0.019**	**1.120**	**8.43**

Optimization. After VAE pretraining, we freeze its parameters and first train the generator with $\mathcal{L}_{\mathrm{diff}}$ and $\mathcal{L}_{\mathrm{MPJPE}}$. We then add $\mathcal{L}_{\mathrm{xy}}$ and $\mathcal{L}_{\mathrm{front}}$ while retaining both original objectives. We find this two-stage training strategy helpful for the model to achieve better performance.

5 Experiments

5.1 Experiment Setup

Dataset. We conduct all experiments on the **CAT-Motion** dataset, a large-scale dataset containing over 100K samples of paired human motion and camera trajectories, each annotated with captions describing human motion under specific camera viewpoints. The dataset captures diverse camera behaviors and human actions, with explicit annotations of human-camera interactions, including entering, in-frame motion, and leaving stages. For motion representation, we adopt the D147 pose representation as described in Section 3.2. Compared to the commonly used 263-dimensional representation in prior works (Guo et al., 2022a;b; Zhu et al., 2025), our representation is more compact and better suited for modeling camera-aware motion dynamics.

Evaluation metrics. We evaluate our method from three perspectives: condition-motion alignment, camera-aware alignment, and perceptual quality. For condition–motion alignment, we follow previous works (Guo et al., 2022b;a; Jiang et al., 2023) and report (1) Bi-R@k to measure the semantic consistency between the input conditions and the generated motion, (2) $\mathrm{FD}_{\mathrm{motion}}$ to measure the Fréchet distance between generated and reference motion embeddings, and (3) Multimodal Distance (MMDist) to evaluate the distance between paired joint-condition and generated motion embeddings. To further evaluate camera-aware consistency, we follow (Courant et al., 2025) and report (4) $\mathrm{FD}_{\mathrm{frame}}$ that measures the Fréchet distance between the distributions of generated and reference frame-level 2D framing features, and (5) ProjE to measure the mean projection error between generated and reference motions in normalized screen coordinates over 22 body joints. We additionally report (6) LLM Score for perceptual evaluation of the rendered outputs with a maximum score of 10. Since our task involves joint modeling of text, camera, and motion, existing evaluation encoders are not directly applicable; therefore, we additionally train a multimodal encoder on the CAT-Motion dataset to align text, camera trajectory, and motion into a shared embedding space. Details of the metrics and multimodal encoder can be found in Appendix.

Implementation details. We implement our model in PyTorch and conduct all experiments on NVIDIA RTX A6000 GPUs. The motion latent representation is learned using a variational autoencoder following the design of MLD (Chen et al., 2023). The diffusion head is based on the

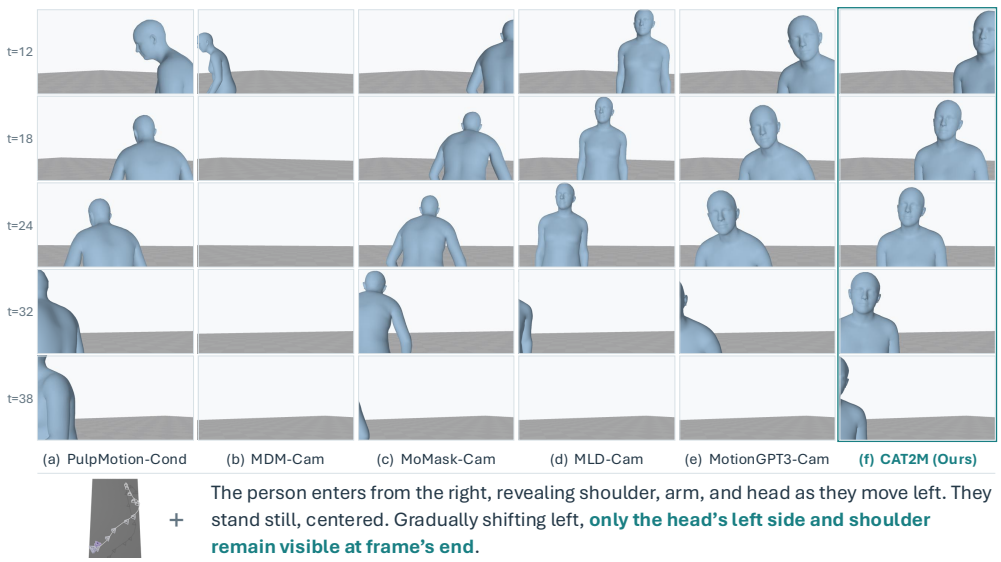


Figure 4: **Comparison between our proposed CAT2M and baselines.** Our model generates motions that match the text description under given camera trajectory.

architecture proposed in (Li et al., 2024), while the text and motion branches follow the design of (Zhu et al., 2025). We use the AdamW optimizer (Loshchilov & Hutter, 2019) throughout all training stages. The VAE is trained for 85 epochs with a learning rate of 1×10^{-4}. For the motion generation model, stage 1 and stage 2 are trained for 140 and 20 epochs, with learning rates of 1×10^{-4} and 1×10^{-5}, respectively.

5.2 Comparison to Baseline Methods

Baselines. We compare with five camera-conditioned adaptations: MDM-Cam (Tevet et al., 2022b), MLD-Cam (Chen et al., 2023), MoMask-Cam (Guo et al., 2024), MotionGPT3-Cam (Zhu et al., 2025), and PulpMotion-Cond (Courant et al., 2025). The first four incorporate camera features into their respective motion-generation backbones. PulpMotion-Cond fixes the camera component of its joint latent and generates only the human-motion component. These adaptations retain the underlying generation mechanisms. Their motion representations, camera encoders, and conditioning interfaces are detailed in Appendix.

Quantitative Results. Table 1 shows that our model CAT2M achieves the best results across all seven metrics. Bi-R@1 and Bi-R@3 exceed MotionGPT3-Cam by 6.44 and 10.08 percentage points, respectively. The $\mathrm{FD}_{\mathrm{motion}}$ and MMDist are also the lowest among the compared methods. For camera-aware alignment, CAT2M reduces $\mathrm{FD}_{\mathrm{frame}}$ by 80.00% and ProjE by 65.91% relative to MLD-Cam, the strongest baseline on both metrics. LLM Score increases from the best baseline score of 7.96 to 8.43. These results demonstrate improvements in joint-condition alignment, motion-distribution matching, and projected appearance.

Qualitative Results. Figure 4 is a comparison of CAT2M with baseline methods on a typical example. The description requires the person to enter from the right, stand near the center, and gradually shift left while remaining partially visible at the end. CAT2M closely follows this progression: the person appears at the right boundary, maintains a standing posture through the central frames, and moves toward the left boundary while retaining part of the head and shoulder in the final frame. In contrast, MDM-Cam places the person near the left boundary at the start, with no person visible in the subsequent sampled frames. PulpMotion-Cond and MoMask-Cam roughly reproduce the right-to-left traversal, but fail to preserve the specified final head-and-shoulder visibility. MLD-Cam and MotionGPT3-Cam also exhibit leftward screen-space movement, but the person is completely out of view by the end. These comparisons highlight that matching the overall movement direction alone is insufficient, CAT2M more faithfully captures the temporal framing and body-part visibility specified by the text.

5.3 Ablation Study

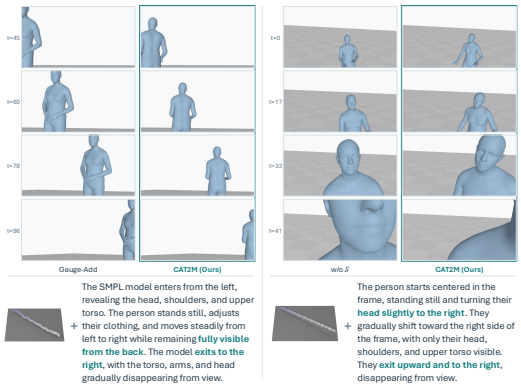


Figure 5: **Qualitative ablations of CAT2M.** We compare the full model with Gauge-Add (left) and w/o $\mathcal{S}$ (right), the strongest ablated variants in camera conditioning and training loss ablation groups, respectively.

To verify the effectiveness of our model design, we conduct ablation experiments on two key aspects: camera conditioning and training losses. For camera conditioning, we examine the first-camera coordinate frame (Gauge) and temporal camera cross attention (TCA) through three variants: Raw-Add uses world coordinates and additive camera injection, Gauge-Add introduces Gauge while retaining additive injection, and TCA Only uses cross attention without Gauge. All three variants retain both geometric objectives. For training losses, we remove 3D joint supervision $\mathcal{L}_{\text{MPJPE}}$ (w/o $\mathcal{M}$), screen-space supervision $\mathcal{L}_{\text{xy}}$ and $\mathcal{L}_{\text{front}}$ (w/o $\mathcal{S}$), or both, while retaining Gauge and TCA. The diffusion objective is used in all variants. Table 2 compares these variants with the full model.

With additive conditioning, Gauge raises Bi-R@1 from 12.94% to 27.90%, highlighting the value of a shared human-camera reference frame. TCA also improves raw-coordinate generation, increasing Bi-R@1 from 12.94% to 21.10%. Combining Gauge and TCA further raises Bi-R@1 to 29.80% and reduces ProjE from 1.386 for Gauge-Add to 1.120. Thus, the shared geometric reference and temporal access to camera conditions make complementary contributions.

The objective ablations show complementary effects of 3D and screen-space supervision. With $\mathcal{S}$ retained, adding $\mathcal{M}$ raises Bi-R@1 from 25.85% to 29.80% and reduces $\text{FD}_{\text{motion}}$ from 0.078 to 0.055. Adding $\mathcal{S}$ to the 3D-supervised variant reduces ProjE from 1.644 to 1.120 (31.87%) and FD_{frame} from 0.023 to 0.019, while $\text{FD}_{\text{motion}}$ remains similar. The combined model also achieves the highest retrieval and LLM scores among the ablations, supporting the use of both objectives for matching the underlying motion and its projected appearance.

Figure 5 compares CAT2M with two strong ablated variants. In the left example, Gauge-Add reproduces the left-to-right traversal but shows the subject from the front, contradicting the requested rear view. CAT2M maintains the back-facing orientation while moving toward the right boundary. In the right example, w/o $\mathcal{S}$ produces an increasingly large frontal close-up rather than the specified upward-right exit. CAT2M instead shifts the head toward the upper-right boundary, with the head outside the frame by the end. Our generated sequence more closely follows the intended exit direction and visibility transition.

6 Conclusion

In this work, we introduce the task of *camera-aware text-to-motion generation*, addressing the challenge of generating human motion that is consistent with text descriptions under given camera trajectories. To support this task, we construct the CAT-Motion dataset, providing over 100K paired human motion and camera trajectory samples with rich stage-wise and sequence level captions capturing human actions and their viewpoint specific appearance. Building on this dataset, we propose CAT2M, which combines a kinematic motion representation in a shared, gravity-preserving first-camera coordinate frame with temporal camera cross attention to model human-camera relative geometry and changing viewpoints. We further complement 3D joint supervision with differentiable screen-space objectives, explicitly connecting the underlying motion to its projected appearance. Experiments demonstrate that CAT2M outperforms camera conditioned baselines in condition-motion alignment and camera-aware alignment, while ablation studies highlight the complementary contributions of the shared representation, temporal camera conditioning, and geometric supervision. Human evaluation further supports the quality of the generated motions. In summary, we propose a new task focusing on the interaction between human motion and camera viewpoints, accompanied by a purpose-built dataset and a model specifically designed to tackle it.

AI USE STATEMENT

In this work, we used generative AI tools for dataset annotation, feedback on research methodology and experimental design, and automated evaluation. Specifically, Qwen3-VL was used to generate captions for the dataset. ChatGPT was used to provide feedback on methodological formulations and experimental design, and to assist with paper writing. GPT-5.4 Mini was used as a multimodal judge to evaluate generated motions. We also used Codex to assist with method implementation and debugging. Formal theorem proving and proof writing are not applicable to this empirical study. Additionally, we used ChatGPT for literature discovery and reference checking, drafting and revising manuscript text, and preparing figures and tables. We assessed AI-generated dataset annotations through a human audit. Automated model-based evaluations were complemented by quantitative metrics and a separate human preference study. The annotation prompts, human audit, and automated evaluation protocol are documented in the appendix. We have reviewed all AI-assisted work. AI-assisted code was reviewed and tested by the authors for correctness. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Hyemin Ahn, Timothy Ha, Yunho Choi, Hwiyeon Yoo, and Songhwai Oh. Text2action: Generative adversarial synthesis from language to action, 2017. URL https://arxiv.org/abs/1710.05298.

Chaitanya Ahuja and Louis-Philippe Morency. Language2pose: Natural language grounded pose forecasting, 2019. URL https://arxiv.org/abs/1907.01108.

Nikos Athanasiou, Mathis Petrovich, Michael J Black, and Gül Varol. Teach: Temporal action composition for 3d humans. In 2022 International Conference on 3D Vision (3DV), pp. 414–423. IEEE, 2022.

Sherwin Bahmani, Ivan Skorokhodov, Guocheng Qian, Aliaksandr Siarohin, Willi Menapace, Andrea Tagliasacchi, David B Lindell, and Sergey Tulyakov. Ac3d: Analyzing and improving 3d camera control in video diffusion transformers. In Proceedings of the Computer Vision and Pattern Recognition Conference, pp. 22875–22889, 2025.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. arXiv preprint arXiv:2511.21631, 2025.

William Bares, Scott McDermott, Christina Boudreaux, and Somying Thainimit. Virtual 3d camera composition from frame constraints. In Proceedings of the eighth ACM international conference on Multimedia, pp. 177–186, 2000.

Rogerio Bonatti, Wenshan Wang, Cherie Ho, Aayush Ahuja, Mirko Gschwindt, Efe Camci, Erdal Kayacan, Sanjiban Choudhury, and Sebastian Scherer. Autonomous aerial cinematography in unstructured environments with learned artistic decision-making. Journal of Field Robotics, 37 (4):606–641, 2020.

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. 2024. URL https://openai.com/research/video-generation-models-as-world-simulators.

Jake Bruce, Michael Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, Yusuf Aytar, Sarah Bechtle, Feryal Behbahani, Stephanie Chan, Nicolas Heess, Lucy Gonzalez, Simon Osindero, Sherjil Ozair, Scott Reed, Jingwei Zhang, Konrad Zolna, Jeff Clune, Nando de Freitas, Satinder Singh, and Tim Rocktäschel. Genie: Generative interactive environments, 2024. URL https://arxiv.org/abs/2402.15391.

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

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.

Marc Christie and Eric Languénou. A constraint-based approach to camera path planning. In International Symposium on Smart Graphics, pp. 172–181. Springer, 2003.

Marc Christie and Patrick Olivier. Camera control in computer graphics: models, techniques and applications. In ACM SIGGRAPH ASIA 2009 Courses, pp. 1–197. 2009.

Robin Courant, Nicolas Dufour, Xi Wang, Marc Christie, and Vicky Kalogeiton. E.t. the exceptional trajectories: Text-to-camera-trajectory generation with character awareness, 2024. URL https://arxiv.org/abs/2407.01516.

Robin Courant, Xi Wang, David Loiseaux, Marc Christie, and Vicky Kalogeiton. Pulp motion: Framing-aware multimodal camera and human motion generation. arXiv preprint arXiv:2510.05097, 2025.

Anindita Ghosh, Noshaba Cheema, Cennet Oguz, Christian Theobalt, and Philipp Slusallek. Synthesis of compositional animations from textual descriptions, 2023. URL https://arxiv.org/abs/2103.14675.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. Communications of the ACM, 63(11):139–144, 2020.

Mirko Gschwindt, Efe Camci, Rogerio Bonatti, Wenshan Wang, Erdal Kayacan, and Sebastian Scherer. Can a robot become a movie director? learning artistic principles for aerial cinematography. In 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 1107–1114. IEEE, 2019.

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. 5152–5161, 2022a.

Chuan Guo, Xinxin Zuo, Sen Wang, and Li Cheng. Tm2t: Stochastic and tokenized modeling for the reciprocal generation of 3d human motions and texts. In European Conference on Computer Vision, pp. 580–597. Springer, 2022b.

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.

Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. arXiv preprint arXiv:2307.04725, 2023.

Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. Cameractrl: Enabling camera control for text-to-video generation. arXiv preprint arXiv:2404.02101, 2024.

Hao He, Ceyuan Yang, Shanchuan Lin, Yinghao Xu, Meng Wei, Liangke Gui, Qi Zhao, Gordon Wetzstein, Lu Jiang, and Hongsheng Li. Cameractrl ii: Dynamic scene exploration via camera-controlled video diffusion models. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 13416–13426, 2025.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598, 2022.

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.

Chong Huang, Chuan-En Lin, Zhenyu Yang, Yan Kong, Peng Chen, Xin Yang, and Kwang-Ting Cheng. Learning to film from professional human motion videos. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4244–4253, 2019.

Chong Huang, Yuanjie Dang, Peng Chen, Xin Yang, and Kwang-Ting Cheng. One-shot imitation drone filming of human motion videos. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(9):5335–5348, 2021.

Tianyu Huang, Wangguandong Zheng, Tengfei Wang, Yuhao Liu, Zhenwei Wang, Junta Wu, Jie Jiang, Hui Li, Rynson Lau, Wangmeng Zuo, et al. Voyager: Long-range and world-consistent video diffusion for explorable 3d scene generation. ACM Transactions on Graphics (TOG), 44 (6):1–15, 2025.

Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. Motiongpt: Human motion as a foreign language. Advances in Neural Information Processing Systems, 36:20067–20079, 2023.

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

Hongda Jiang, Xi Wang, Marc Christie, Libin Liu, and Baoquan Chen. Cinematographic camera diffusion model. In Computer Graphics Forum, volume 43, pp. e15055. Wiley Online Library, 2024.

Jihoon Kim, Jiseob Kim, and Sungjoon Choi. Flame: Free-form language-based motion synthesis & editing. In Proceedings of the AAAI conference on artificial intelligence, volume 37, pp. 8255–8263, 2023.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114, 2013.

Tianhong Li, Yonglong Tian, He Li, Mingyang Deng, and Kaiming He. Autoregressive image generation without vector quantization, 2024. URL https://arxiv.org/abs/2406.11838.

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

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. ACM Trans. Graphics (Proc. SIGGRAPH Asia), 34(6):248:1–248:16, October 2015.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019. URL https://arxiv.org/abs/1711.05101.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. Advances in neural information processing systems, 35:5775–5787, 2022.

Xiaofeng Mao, Shaoheng Lin, Zhen Li, Chuanhao Li, Wenshuo Peng, Tong He, Jiangmiao Pang, Mingmin Chi, Yu Qiao, and Kaipeng Zhang. Yume: An interactive world generation model. arXiv preprint arXiv:2507.17744, 2025.

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 Computer Vision and Pattern Recognition Conference, pp. 27859–27871, 2025.

Jeongeun Park, Sungjoon Choi, and Sangdoo Yun. A unified framework for motion reasoning and generation in human interaction. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10698–10707, 2025.

Mathis Petrovich, Michael J. Black, and Gül Varol. Action-conditioned 3D human motion synthesis with transformer VAE. In International Conference on Computer Vision (ICCV), 2021.

Mathis Petrovich, Michael J. Black, and Gül Varol. Temos: Generating diverse human motions from textual descriptions, 2022. URL https://arxiv.org/abs/2204.14109.

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, 2024. URL https://arxiv.org/abs/2401.08559.

Matthias Plappert, Christian Mandery, and Tamim Asfour. The kit motion-language dataset. Big Data, 4(4):236–252, December 2016. ISSN 2167-647X. doi: 10.1089/big.2016.0028. URL http://dx.doi.org/10.1089/big.2016.0028.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

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, 2021. URL https://arxiv.org/abs/2103.00020.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 10684–10695, 2022.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502, 2020.

Robbyant Team, Zelin Gao, Qiuyu Wang, Yanhong Zeng, Jiapeng Zhu, Ka Leong Cheng, Yixuan Li, Hanlin Wang, Yinghao Xu, Shuailei Ma, et al. Advancing open-source world models. arXiv preprint arXiv:2601.20540, 2026.

Guy Tevet, Brian Gordon, Amir Hertz, Amit H Bermano, and Daniel Cohen-Or. Motionclip: Exposing human motion generation to clip space. In European Conference on Computer Vision, pp. 358–374. Springer, 2022a.

Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H Bermano. Human motion diffusion model. arXiv preprint arXiv:2209.14916, 2022b.

Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning, 2018. URL https://arxiv.org/abs/1711.00937.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. Advances in neural information processing systems, 30, 2017.

Yuqing Wang, Zhijie Lin, Yao Teng, Yuanzhi Zhu, Shuhuai Ren, Jiashi Feng, and Xihui Liu. Bridging continuous and discrete tokens for autoregressive visual generation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 18596–18605, 2025.

Zhouxia Wang, Ziyang Yuan, Xintao Wang, Yaowei Li, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. Motionctrl: A unified and flexible motion controller for video generation. In ACM SIGGRAPH 2024 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.

Zixuan Wang, Jiayi Li, Xiaoyu Qin, Shikun Sun, Songtao Zhou, Jia Jia, and Jiebo Luo. Dancecamanimator: Keyframe-based controllable 3d dance camera synthesis. In Proceedings of the 32nd ACM International Conference on Multimedia, pp. 10200–10209, 2024c.

Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. In Forty-first International Conference on Machine Learning, 2024.

Tong Wu, Shuai Yang, Ryan Po, Yinghao Xu, Ziwei Liu, Dahua Lin, and Gordon Wetzstein. Video world models with long-term spatial memory. arXiv preprint arXiv:2506.05284, 2025.

Desai Xie, Ping Hu, Xin Sun, Soren Pirk, Jianming Zhang, Radomír Mech, and Arie E Kaufman. Gait: Generating aesthetic indoor tours with deep reinforcement learning. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 7409–7419, 2023.

Hong-Xing Yu, Haoyi Duan, Junhwa Hur, Kyle Sargent, Michael Rubinstein, William T Freeman, Forrester Cole, Deqing Sun, Noah Snavely, Jiajun Wu, et al. Wonderjourney: Going from anywhere to everywhere. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6658–6667, 2024.

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, 2025a. URL https://arxiv.org/abs/2504.07083.

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.

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

Sixiao Zheng, Minghao Yin, Wenbo Hu, Xiaoyu Li, Ying Shan, and Yanwei Fu. Versecrafter: Dynamic realistic video world model with 4d geometric control. arXiv preprint arXiv:2601.05138, 2026.

Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images, 2018. URL https://arxiv.org/abs/1805.09817.

Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks, 2020. URL https://arxiv.org/abs/1812.07035.

Bingfan Zhu, Biao Jiang, Sunyi Wang, Shixiang Tang, Tao Chen, Linjie Luo, Youyi Zheng, and Xin Chen. Motiongpt3: Human motion as a second modality. arXiv preprint arXiv:2506.24086, 2025.

A Appendix

A.1 CAT-Motion Dataset Details

A.1.1 Comparison with Existing Datasets

Table 5 presents a comparison and statistical analysis of our dataset against existing human motion and camera trajectory datasets. To the best of our knowledge, our CAT-Motion dataset is the first to provide explicit captions describing how human actions are perceived within the camera frame.

A.1.2 Dataset Statistics

A source motion identifies an underlying motion sequence, while a camera-conditioned variant pairs that sequence with a prescribed camera trajectory, intrinsics, and viewpoint-specific annotations. The complete dataset contains 100,166 variants from 12,768 source motions, approximately 7.85 variants per source motion. Table 3 lists the complete split sizes. All variants of one source motion belong to the same split.

Table 3: Complete CAT-Motion manifest.

Split	Source motions	Camera-conditioned variants	Frames
Training	10,215	80,034	7,609,488
Validation	1,279	9,993	961,575
Test	1,274	10,139	962,501
Total	12,768	100,166	9,533,564

A.1.3 Camera Trajectory Design

As shown in Table 4, we design five trajectory augmentation strategies to transform the original camera trajectories, with a particular focus on modeling scenarios where the subject enters and leaves the frame. These strategies can be applied individually (except for the `'re'` strategy) or combined flexibly to produce diverse camera trajectories. In practice, we use 9 strategies: `'x'`, `'y'`, `'z'`, `'h'`, `'x_re'`, `'y_re'`, `'z_re'`, `'h_re'`, and `'xy'`.

Table 4: Trajectory design strategies.

Strategy	Operation
`'x'`	Extend original trajectory along x axis
`'y'`	Extend original trajectory along y axis
`'z'`	Extend original trajectory along z axis
`'h'`	Add horizontal rotation to original camera poses
`'re'`	Reverse the current trajectory

A.1.4 Prompts

We design stage-specific prompts for captioning the entering, moving, and leaving stages. Each prompt follows a three-pass pipeline: (1) direct visual observation, (2) structured motion reasoning, and (3) constrained natural language generation. This progression guides the model from low-level visual evidence to structured motion descriptions, with explicit constraints to reduce hallucination and exclude non-visual interpretations. The prompts emphasize temporal consistency and grounding in visible human motion. A separate aggregation prompt combines the stage-wise captions into a coherent description of the entire sequence. To produce compact captions for motion-generation training, we introduce an additional compression prompt that takes the full description and structured metadata as input and requests a short paragraph of 40–55 words. It explicitly removes frame numbers, repetition, decorative wording, and minor details, while describing spatial changes in terms of the person's movement within the frame. The five prompts are shown below.

Table 5: Motion-camera dataset comparison. We compare our dataset with other human motion and camera datasets, including modality information and statistics. The CAT-Motion dataset is the only dataset containing captions on the human-camera relationship. Note that "#" stands for "Total number of".

Dataset	Camera trajectory	Human motion	Caption		Statistics	
			Human motion	Human-camera relationship	#Samples	#Frames
RealEstate10k (Zhou et al., 2018)	✓	✗	✗	✗	79K	11M
DataDoP (Zhang et al., 2025a)	✓	✗	✗	✗	29K	11M
KIT-ML (Plappert et al., 2016)	✗	✓	✓	✗	4K	0.8M
HumanML3D (Guo et al., 2022a)	✗	✓	✓	✗	14K	2M
E.T. (Courant et al., 2024)	✓	✓	✗	✗	115K	11M
Motion-X (Lin et al., 2023)	✗	✓	✓	✗	81K	15.6M
Motion-X++ (Zhang et al., 2025b)	✗	✓	✓	✗	120K	19.5M
PulpMotion (Courant et al., 2025)	✓	✓	✓	✗	193K	22M
CAT-Motion (Ours)	✓	✓	✓	✓	100K	9.5M

Prompt 1: Entering-stage captioning

You are analyzing a video of a human SMPL body model entering the frame. Your task is to produce a precise, factual, and strictly visual description of the entry motion.

You MUST follow the process below in order.

PASS 1: VISUAL OBSERVATION (NO INTERPRETATION)
Carefully observe the video and list only what is directly visible, without summarizing or interpreting.
Describe, in temporal order:
Which body parts are visible in the video
How and from where the model enters the frame (if applicable)

Rules for Pass 1:
Use biomechanical and kinematic descriptions only
Do NOT infer intention, emotion, identity, gender, or actions beyond visible motion
Do NOT name activities (e.g., walking, dancing) unless unmistakably shown
Do NOT describe facial features, clothing, floor, or environment
Do NOT hallucinate invisible body parts

PASS 2: STRUCTURED MOTION REASONING
Using ONLY the observations from Pass 1, organize the motion using step−by−step temporal reasoning:
a. Initial pose and visible body parts
b. Entry direction

Focus on:
Directional movement
Visibility of body segments

PASS 3: FINAL DESCRIPTION (STRUCTURED OUTPUT FORMAT)
Using the structured reasoning above, produce the final description following ALL constraints below.
IMPORTANT CONSTRAINTS:
Describe ONLY what is visually verifiable in the video
Do NOT hallucinate actions, objects, or interactions
Do NOT assume real−world context
Do NOT describe the floor, environment, facial features, or clothing
Do NOT include timestamps
Do NOT include analysis steps or headings in the final output

OUTPUT FORMAT (MANDATORY):
Produce ONE sentence. The sentence must flow naturally and describe the entry of the model into the frame.

PASS 4: SELF−VERIFICATION (SILENT CHECK)
Before finalizing your answer, internally verify that:
Every described body part is visible at the time it is mentioned
No inferred intent or activity labels are used
No motion is described without visual evidence
Remove or correct any content that violates these rules.

Good Example Format:
The SMPL model enters the frame from left, revealing the head and torso.
The SMPL model enters the frame from top right, revealing the feet and legs.

Now analyze the video and output ONLY the final description based on the format above.

Prompt 2: Moving-stage captioning

You are analyzing a video of a human SMPL body model performing a motion sequence.
Your task is to produce a precise, factual, and strictly visual description of how the SMPL model moves within the video frame.

Your primary goal is to determine:
Where the SMPL model is located in the frame at the beginning of the video
Where the SMPL model is located at the end of the video
How the SMPL model moves between these positions (e.g., left to right, top to bottom, diagonally, getting closer or far away from the camera, or with no significant translation)

You MUST follow the process below in order.

PASS 1: VISUAL OBSERVATION (NO INTERPRETATION)
Carefully observe the video and list only what is directly visible, without summarizing or interpreting.
Describe, in temporal order:
The initial position of the SMPL model in the frame (e.g., left side, right side, center, top, bottom)
Any visible translational movement of the models body within the frame
The final position of the SMPL model in the frame
Which body parts are visible at each stage
Any changes in visibility of body parts as the model moves

Rules for Pass 1:
Use biomechanical and kinematic descriptions only
Do NOT infer intention, emotion, identity, gender, or purpose
Do NOT name activities (e.g., walking, running) unless visually unmistakable
Do NOT describe facial features, clothing, the floor, or the environment
Do NOT hallucinate invisible body parts
If something is unclear, explicitly say uncertain

PASS 2: STRUCTURED MOTION REASONING
Using only the observations from Pass 1, organize the motion logically:
a. Frame−level movement
Where the model starts in the frame
How the model translates across the frame (direction and relative displacement)
Where the model ends in the frame
b. Body visibility
Which body segments are visible throughout
Any appearance or disappearance of body parts due to framing

Focus strictly on:
Directional movement within the frame
Changes in spatial position
Visibility of body segments

PASS 3: FINAL DESCRIPTION (STRICT OUTPUT FORMAT)
Using the structured reasoning above, produce the final description under all constraints below.
IMPORTANT CONSTRAINTS:
Describe ONLY what is visually verifiable

If something is ambiguous, say uncertain
Do NOT infer actions, intent, or real−world context
Do NOT describe the environment, floor, clothing, or facial features
Do NOT include timestamps
Do NOT include analysis steps, headings, or explanations

OUTPUT FORMAT
Produce ONE sentence that flows naturally and describes:
The SMPL models start position
Its movement path within the frame
Its end position
The visibility of body parts, in correct temporal order

Example Outputs
The SMPL model starts on the left side of the frame and gradually moves toward the right, with the torso, arms, and head visible at first, then lower torso gradually disappearing from view.
The SMPL model gradually comes closer to the camera, while moving towards the right, with the right side of the body out of view at the end.
The SMPL model appears near the top of the frame and moves downward toward the center, gradually revealing the lower body while the torso and arms remain visible.

Prompt 3: Leaving-stage captioning

You are analyzing a video of a human SMPL body model leaving the frame. Your task is to produce a precise, factual, and strictly visual description of the leaving motion.

You MUST follow the process below in order.

PASS 1: VISUAL OBSERVATION (NO INTERPRETATION)
Carefully observe the video and list only what is directly visible, without summarizing or interpreting.
Describe, in temporal order:
Which body parts are visible in the video
How and from where the model leaves the frame (if applicable)

Rules for Pass 1:
Use biomechanical and kinematic descriptions only
Do NOT infer intention, emotion, identity, gender, or actions beyond visible motion
Do NOT name activities (e.g., walking, dancing) unless unmistakably shown
Do NOT describe facial features, clothing, floor, or environment
Do NOT hallucinate invisible body parts

PASS 2: STRUCTURED MOTION REASONING
Using ONLY the observations from Pass 1, organize the motion using step−by−step temporal reasoning:
a. Initial pose and visible body parts
b. Leaving direction

Focus on:
Directional movement
Visibility of body segments

PASS 3: FINAL DESCRIPTION (STRUCTURED OUTPUT FORMAT)
Using the structured reasoning above, produce the final description following ALL constraints below.
IMPORTANT CONSTRAINTS:
Describe ONLY what is visually verifiable in the video
Do NOT hallucinate actions, objects, or interactions
Do NOT assume real−world context
Do NOT describe the floor, environment, facial features, or clothing
Do NOT include timestamps
Do NOT include analysis steps or headings in the final output

OUTPUT FORMAT (MANDATORY):
Produce ONE sentence. The sentence must flow naturally and describe the exit of the model from the frame.

PASS 4: SELF−VERIFICATION (SILENT CHECK)
Before finalizing your answer, internally verify that:
Every described body part is visible at the time it is mentioned
No inferred intent or activity labels are used
No motion is described without visual evidence
Remove or correct any content that violates these rules.

Good Example Format:
The model exits the frame to the right, with its torso and arms gradually moving out of view.
The model exits the frame from bottom left, with its feet, legs and torso moving out of view in order.

Now analyze the video and output ONLY the final description based on the format above.

Prompt 4: Full captioning

You are a professional video caption annotator. The input describes a video containing ONE human subject.

Semantic interpretation rules:
− The caption must be HUMAN−CENTRIC.
− Never describe or infer camera movement.
− All spatial descriptions refer to how the PERSON appears within the frame.
− The ''human_motion'' describes motion during the ENTIRE clip.
− The ''camera_motion_relationship'' describes how the person's position changes within the frame after appearing and MUST be included.

Temporal structure constraints:
− ''in_body_parts'', ''camera_motion_relationship'', and ''out_body_parts'' represent THREE DISTINCT and SEQUENTIAL time phases.
− These phases are continuous but DO NOT overlap.
− Each phase must be described separately in chronological order.
− Do NOT merge or mix information from these phases into a single blended description.

Coverage requirement:
− ALL information in the metadata must be explicitly expressed.
− Do not overly compress multiple temporal phases into one sentence.

Task:
Convert the structured metadata into a detailed natural−language video caption.

Rules:
− Maintain strict chronological order: beginning entering phase in−frame position phase exiting phase ending.
− Use clear and explicit transitions between phases.
− Do not hallucinate details.
− Ignore empty fields naturally.
− Write ONE paragraph of approximately 90−160 words.
− Output only the caption text.
− Never mention the camera.

Metadata:
<JSON>
{JSON_CONTENT}
</JSON>

Caption:

Prompt 5: Concise captioning

Rewrite the source as a compact caption of 40−55 English words.

Requirements:
− Hard constraint: the final paragraph MUST contain 40−55 English words;
count the words before returning it.

– Preserve chronological order.
– Preserve the person's initial visibility, entry when applicable, main human motion, important within−frame position changes, exit when applicable, and final visibility.
– Refer to spatial changes as the person's movement within the frame, never as camera movement.
– Remove frame numbers, repeated statements, decorative wording, and minor details.
– Prefer three short sentences: start/entry; main motion and position change; exit/end.
– The result will be checked against a strict CLIP token budget, so be concise.

Structured metadata:
{metadata}

Full caption:
{full_caption}

Short caption:

A.1.5 Human Evaluation of Dataset Annotations

We conducted a human annotation audit on 40 randomly sampled dataset examples. The assessment covered hallucination, omission, correctness of entering/in-frame/exiting descriptions, and event order. Table 6 reports the ratings and scores. The human evaluation demonstrates the quality of our CAT-Motion dataset, and supports its ability of supporting CAT2M model training and evaluation.

Videos of our dataset examples can be found in the supplementary material.

Table 6: Human annotation-quality audit with 40 randomly sampled examples. Panel (a) rates hallucination and omission, Slight and Serious deducts 10% and 50% respectively from the overall ratings. Panel (b) gives stage-wise correctness. n is the effective count for each aspect. The reported normalized scores use weights 1, 0.5, and 0 for Correct, Partially Correct, and Incorrect.

(a) Hallucination and omission ratings.

Dimension	n	Slight	Serious	Overall Rating
Hallucination	40	5	2	4.81/5
Omission	40	3	1	4.90/5

(b) Stage-wise correctness and event order.

Aspect	n	Correct	Partially Correct	Incorrect	Normalized score (%)
Entering	21	17 (80.95%)	4 (19.05%)	0 (0.00%)	90.48
In-frame	40	31 (77.50%)	9 (22.5%)	0 (0.00%)	88.75
Exiting	22	19 (86.36%)	2 (9.09%)	1 (4.55%)	90.91
Event order	40	37 (92.50%)	3 (7.50%)	0 (0.00%)	96.25

A.2 Experiment Details

A.2.1 Ablation on Human Motion Representation

We investigated seven motion representations to assess which information survives deterministic encoding and decoding. This experiment isolates representation fidelity: it uses no learned autoencoder, generative model, or motion-to-mesh fitting. We distinguish preservation of joint positions, preservation of the original body rotations and mesh, and preservation of the human's placement relative to the prescribed camera.

Notation. Let $\mathbf{r}_t \in \mathbb{R}^3$, $R_t^0 \in \mathrm{SO}(3)$, and $\{R_t^j\}_{j=1}^{23}$ denote the translation parameter, global root orientation, and local body rotations of SMPL (Loper et al., 2015). Let $\mathbf{J}_{t,j}$ be its world-space joints and $(Q_t, \mathbf{c}_t)$ the prescribed camera-to-world pose. We use the continuous 6D rotation encoding

$\rho_6 : \mathrm{SO}(3) \to \mathbb{R}^6$ (Zhou et al., 2020), with a consistent matrix layout and Gram-Schmidt decoding. The SMPL translation parameter $\mathbf{r}_t$ is distinct from the pelvis position $\mathbf{J}_{t,0}$.

144-Dim. The frame feature is

$$\mathbf{x}_t^{144} = \left[\, \mathrm{vec}(\widetilde{\mathbf{P}}_t), \mathrm{vec}(\Delta\widetilde{\mathbf{P}}_t), \mathbb{1}_{t=0}\mathbf{u}_0, \rho_6^{\mathrm{col}}(Q_t^\top R_t^0), \Delta\mathbf{u}_t \right] \in \mathbb{R}^{144}, \tag{7}$$

Here $\mathbf{u}_t = Q_t^\top(\mathbf{r}_t - \mathbf{c}_t)$ and $\widetilde{\mathbf{P}}_{t,j} = Q_t^\top \mathbf{J}_{t,j} + \mathbf{c}_t - \mathbf{u}_t$ for 22 joints. The feature does not determine local joint rotations or axial twist, and has no native SMPL mesh decoder.

Camera-gauged SMPL: D135, D141, and D147. We define a fixed frame from the first camera: its center is the origin, world up is retained, and its horizontal viewing direction fixes the heading. Let $A \in \mathrm{SO}(3)$ map world coordinates to this frame and $\mathbf{o} = \mathbf{c}_0$. For K retained rotation joints,

$$\mathbf{x}_t^{(K)} = \left[A(\mathbf{r}_t - \mathbf{o}),\ \rho_6(AR_t^0), \rho_6(R_t^1), \ldots, \rho_6(R_t^{K-1}) \right] \in \mathbb{R}^{3+6K}. \tag{8}$$

D135, D141, and D147 use $K = 22, 23, 24$, respectively, including the root. Thus D141 contains 23 rotation joints, D147 retains all 24 SMPL rotations. The decoder inverts the fixed transform before evaluating SMPL, and fills any omitted distal rotations with identities. D135 and D141 preserve the first 22 joint positions but can change the mesh through their omitted rotations. Our subsequent model uses D147 with an invertible root encoding: $\mathbf{a}_0 = \mathbf{r}_0^g$ and $\mathbf{a}_t = \mathbf{r}_t^g - \mathbf{r}_{t-1}^g$ for $t > 0$, where $\mathbf{r}_t^g = A(\mathbf{r}_t - \mathbf{o})$. The absolute translation is recovered by cumulative summation; this learned model parameterization is separate from the raw round-trip test below.

PulpMotion-199. We use the 199-channel SMPL-RIFKE implementation supplied with Pulp Motion (Courant et al., 2025), adapted from the SMPL-RIFKE code of STMC (Petrovich et al., 2024):

$$\mathbf{x}_t^{199} = \left[h_t,\ \mathbf{v}_{t,xy}^{\mathrm{local}},\ \eta_t, \{\rho_6(\overline{R}_t^j)\}_{j=0}^{21},\ \{\overline{\mathbf{p}}_{t,j}\}_{j=1}^{21} \right] \in \mathbb{R}^{199}. \tag{9}$$

The dimensions are: pelvis height, horizontal root displacement in the yaw-local frame, a yaw channel, 22 rotations, and 21 root-relative joint positions in the yaw-local frame. The root rotation has its yaw removed, while other rotations remain local SMPL rotations. The yaw channel stores the initial yaw at $t = 0$ and successive increments thereafter. The decoder integrates yaw and horizontal displacement, setting initial horizontal root placement to zero. The two absent distal SMPL rotations are identity-padded for mesh recovery. The 199-D count refers specifically to this implementation.

HumanML3D-263 and HumanML3D-263+anchor. The HumanML3D feature (Guo et al., 2022a) is

$$\mathbf{x}_t^{263} = \left[\omega_t,\ \mathbf{v}_{t,xz},\ h_t, \mathbf{p}_t^{\mathrm{RIC}},\ \mathbf{q}_t^{\mathrm{IK}},\ \mathbf{v}_t^{\mathrm{joint}},\ \mathbf{f}_t \right] \in \mathbb{R}^{263}, \tag{10}$$

with dimensions for root angular increment, horizontal root displacement, root height, rotation-invariant joint positions, IK-derived 6D rotations, joint displacements, and foot contacts. We retain the official angular-increment convention and RIC joint decoder. Preprocessing matches a reference skeleton, removes the floor offset and initial horizontal position, and canonicalizes the initial heading. The IK rotations do not identify the original SMPL rotations or twist. Our anchored variant additionally stores $\mathbf{b} = (x_0, z_0, \psi_0) \in \mathbb{R}^3$ in the first-camera horizontal frame after conversion to the HumanML3D y-up convention. This is a 263-D frame feature plus three sequence-level values, denoted $263 + 3$ in the table. The anchor restores initial horizontal placement and heading, but not skeleton scaling, removed floor offset, or source SMPL pose.

Evaluation protocol. All methods use the same raw SMPL motions, prescribed cameras and intrinsics, and shared source body shape. HumanML3D produces $T - 1$ frames and is compared with the corresponding source prefix, the other rows retain T. We average within each sequence and then equally across camera variants.

Metrics. Global MPJPE and root-position error are the mean world-space Euclidean errors over 22 joints and the pelvis, respectively. Local MPJPE subtracts the pelvis from each joint. Root and joint rotation errors are geodesic angles in degrees; the latter uses the common first 21 local SMPL joints. MPVPE averages mesh-vertex distances using the shared source shape.

Projected-joint error uses unclipped normalized device coordinates $((u - c_x)/c_x, (v - c_y)/c_y)$ under the prescribed cameras. Its historical validity mask requires finite projections and $|z| > 10^{-8}$ for both motions, including points behind the camera. Visibility is evaluated separately as macro-F1

Table 7: Deterministic round-trip fidelity. Panels A and B include all seven representations; Panel C isolates the three SMPL joint-count variants. MPJPE, root-position error, and MPVPE are in mm; all rotation errors are in degrees; Δv_{foot} is in mm/s. NDC errors are dimensionless, and visibility F1 is reported on a 0-1 scale. N/A denotes a quantity unavailable from the native decoder.

A. Joint, root, and mesh reconstruction

Representation	Dim.	Local MPJPE ↓	Global MPJPE ↓	Root position ↓	Root rotation ↓	Joint rotation ↓	MPVPE ↓
144-Dim	144	4.00e−4	1.08e−3	1.03e−3	3.50e−5	N/A	N/A
SMPL-D135	135	6.61e−5	1.05e−4	6.77e−5	6.81e−6	3.58e−6	0.316
SMPL-D141	141	6.61e−5	1.05e−4	6.77e−5	6.81e−6	3.58e−6	0.166
SMPL-D147	147	6.61e−5	1.05e−4	6.77e−5	6.81e−6	3.58e−6	1.29e−4
PulpMotion-199	199	1.68e−4	5.07e−4	3.88e−4	4.68e−5	3.44e−6	0.317
H3D-263	263	31.509	2969.649	2993.616	N/A	N/A	N/A
H3D-263+anchor	263+3	31.509	82.695	72.852	N/A	N/A	N/A

B. Projection, visibility, and contact-speed preservation

Representation	NDC mean ↓	NDC median ↓	NDC P95 ↓	Vis. F1 ↑	Δv_{foot} ↓
144-Dim	18.656	2.90e−6	1.25e−2	1.000	1.41e−3
SMPL-D135	0.379	2.73e−7	1.18e−3	1.000	2.10e−4
SMPL-D141	0.379	2.73e−7	1.18e−3	1.000	2.10e−4
SMPL-D147	0.379	2.73e−7	1.18e−3	1.000	2.10e−4
PulpMotion-199	9.578	5.90e−7	3.09e−3	1.000	8.30e−4
H3D-263	129.413	40.541	337.899	0.461	31.475
H3D-263+anchor	17.472	0.226	38.236	0.945	31.475

C. Joint-count controls

Representation	Joints	Global MPJPE ↓	Root rotation ↓	Local rotation ↓	MPVPE ↓
SMPL-D135	22	1.05e−4	6.81e−6	0.180	0.316
SMPL-D141	23	1.05e−4	6.81e−6	8.95e−2	0.166
SMPL-D147	24	1.05e−4	6.81e−6	3.58e−6	1.29e−4

over visible/non-visible joints, requiring positive depth and image bounds $[0, 2c_x) \times [0, 2c_y)$. We report NDC mean, median, and P95 across sequence-level errors because near-zero depths make the mean sensitive to numerical perturbations. The historical foot-skating diagnostic is specifically a contact-speed difference:

$$s_{t,j} = \frac{\|\mathbf{J}_{t+1,j} - \mathbf{J}_{t,j}\|_2}{\Delta t},$$
$$\Delta v_{\text{foot}} = \frac{10^3}{|\mathcal{C}|} \left| \sum_{(t,j)\in\mathcal{C}} (\widehat{s}_{t,j} - s_{t,j}) \right|, \tag{11}$$

where joint coordinates are in metres, $\widehat{s}$ is computed from reconstructed joints, and Δv_{foot} is in mm/s. Here $\Delta t = 1/25$ s and $\mathcal{C}$ contains the four foot joints whose squared source displacement is below $0.002\,\text{m}^2$ per frame. A small difference indicates preservation of source contact-period speed, not absence of skating in the source.

Observations. The results are shown in Table 7. Joint-position accuracy alone does not guarantee faithful SMPL reconstruction: D135, D141, and D147 share near-zero joint errors, but their MPVPE values are 0.316 mm, 0.166 mm, and 1.29×10^{-4} mm, respectively, demonstrating the importance of retaining all SMPL rotations. 144-Dim lacks native local rotations and mesh recovery, while PulpMotion-199 omits distal rotations. Adding an anchor substantially improves HumanML3D, but preprocessing losses leave a global MPJPE of 82.695 mm. Thus, our chosen D147 combines the lowest mesh reconstruction error with near-zero joint and rotation errors, preserving the complete pose geometry and global placement required by our prescribed-camera motion generation task.

A.2.2 VAE TRAINING

To separate initial placement from subsequent movement, we encode the root using an absolute first-frame anchor and inter-frame displacements, normalized independently:

$$\widetilde{r}_0 = \frac{r_0 - \mu_a}{\sigma_a}, \qquad \widetilde{r}_t = \frac{r_t - r_{t-1} - \mu_\Delta}{\sigma_\Delta}, \quad t \geq 1. \tag{12}$$

Here, (μ_a, σ_a) and $(\mu_\Delta, \sigma_\Delta)$ denote channel-wise means and standard deviations for anchors and displacements. Together with standardized rotation features, this defines the VAE input $\widetilde{\mathcal{M}} = \mathcal{N}(\mathcal{M})$. The inverse map $\mathcal{N}^{-1}$ restores absolute root positions by accumulating denormalized displacements from the recovered anchor.

The VAE is pretrained with

$$\begin{aligned}\mathcal{L}_{\text{VAE}} = \mathcal{L}_{\text{feat}} + \mathcal{L}_{\text{joint-rec}} + 0.5\mathcal{L}_{\Delta\text{rot}} \\ + 0.25\mathcal{L}_{\text{root-delta}} + 0.1\mathcal{L}_{\text{rel-traj}} + 10^{-3}\mathcal{L}_{\text{KL}}.\end{aligned} \tag{13}$$

The first three terms use Smooth L1 to reconstruct normalized features, SMPL joint positions, and temporal differences of normalized rotation features, respectively. The root-delta term additionally weights the normalized root-displacement reconstruction, while the relative-trajectory term measures the Euclidean error between reconstructed and reference $r_t - r_0$ in physical coordinates. These objectives supervise both individual displacements and their accumulated trajectory. $\mathcal{L}_{\text{KL}}$ regularizes the posterior toward a standard normal distribution. The VAE is trained for 85 epochs.

A.2.3 DEFINITION OF THE METRICS

Let $\{(c_i, x_i, \hat{x}_i)\}_{i=1}^N$ denote the input conditions, reference motions, and generated motions, respectively. The frozen evaluator produces unit-normalized embeddings $q_i = E_c(c_i)$, $z_i = E_m(x_i)$, and $\hat{z}_i = E_m(\hat{x}_i)$, where c_i includes text and camera parameters.

Let $r_i^{c\to m}$ and $r_i^{m\to c}$ denote the ranks of the paired match in condition-to-motion and motion-to-condition retrieval, respectively. Both directions use cosine similarity $q_i^\top \hat{z}_j$ over the full candidate pool. We report bidirectional recall:

$$\text{Bi-R@}k = \frac{1}{2N}\sum_{i=1}^{N}\left[\mathbf{1}(r_i^{c\to m} \leq k) + \mathbf{1}(r_i^{m\to c} \leq k)\right]. \tag{14}$$

The multimodal distance measures paired condition-motion alignment:

$$\text{MMDist} = \frac{1}{N}\sum_{i=1}^{N}\|q_i - \hat{z}_i\|_2. \tag{15}$$

For feature sets A and B, define

$$\mathcal{F}(A,B) = \|\mu_A - \mu_B\|_2^2 + \text{tr}\left(\Sigma_A + \Sigma_B - 2\left(\Sigma_A^{1/2}\Sigma_B\Sigma_A^{1/2}\right)^{1/2}\right), \tag{16}$$

where μ and Σ are the empirical mean and sample covariance. The motion and framing distances are

$$\text{FD}_{\text{motion}} = \mathcal{F}\big(\{z_i\}_{i=1}^N, \{\hat{z}_i\}_{i=1}^N\big), \tag{17}$$

$$\text{FD}_{\text{frame}} = \mathcal{F}\Big(\{\phi_{it}\}_{i,t}, \{\hat{\phi}_{it}\}_{i,t}\Big). \tag{18}$$

Following the PulpMotion framing protocol, each frame feature is

$$\phi_{it} = \text{vec}\left(\text{clip}_{[-2,2]}\left[\left(\frac{u_{itj} - c_{x,it}}{c_{x,it}}, \frac{c_{y,it} - v_{itj}}{c_{y,it}}\right)\right]_{j\in\mathcal{J}_9}\right) \in \mathbb{R}^{18}, \tag{19}$$

where $\mathcal{J}_9$ is the fixed nine-joint set, (u, v) are pixel projections, and (c_x, c_y) is the camera principal point. Generated features $\hat{\phi}_{it}$ are defined identically. Features are pooled over all valid frames, with non-finite entries replaced by zero after clipping.

Let $p_{itj} = (2u_{itj}/W - 1,\ 1 - 2v_{itj}/H)$ denote the unclipped normalized screen projection, and define $\hat{p}_{itj}$ analogously. A generated projection is valid when its camera-space 3D coordinates and projected coordinates are finite and its depth satisfies $\hat{z}_{itj} > 10^{-4}\,\mathrm{m}$. Out-of-frame projections remain valid. Let $\mathcal{V}_{it} \subseteq \{1, \ldots, 22\}$ contain reference-visible joints whose generated projections are valid, and let $\mathcal{T}_i = \{t \in \mathcal{T}_i^{\text{valid}} : |\mathcal{V}_{it}| > 0\}$. Frames with no valid joint pairs are skipped. Sequences with $\mathcal{T}_i = \varnothing$ are marked N/A and excluded from aggregation. For the remaining sequences,

$$e_i = \frac{1}{|\mathcal{T}_i|} \sum_{t \in \mathcal{T}_i} \frac{1}{|\mathcal{V}_{it}|} \sum_{j \in \mathcal{V}_{it}} \|\hat{p}_{itj} - p_{itj}\|_2, \tag{20}$$

$$\mathrm{ProjE} = \frac{1}{S} \sum_{s=1}^{S} \frac{1}{|\mathcal{I}_s|} \sum_{i \in \mathcal{I}_s} e_i, \tag{21}$$

where $\mathcal{T}_i^{\text{valid}}$ denotes non-padding frames, $\mathcal{I}_s$ contains the evaluable conditions associated with source motion s, and S counts source motions with at least one evaluable condition. Each evaluable source motion receives equal weight. If $S = 0$, ProjE is N/A.

A.2.4 LLM Score Details

We use GPT-5.4 Mini as an MLLM judge. For each generated motion, we uniformly sample 32 frames from the prescribed camera view and arrange them chronologically into eight timestamped 2×2 image grids, provided together with the text description. The judge independently assigns raw 0-10 scores for action-text consistency, screen-relation-condition consistency, motion naturalness and temporal coherence, and overall preference. The judging prompt is shown below.

Prompt 6: LLM Score generation

You are an expert evaluator of text– and camera–conditioned human motion generation.

INPUT

Text description:
‘‘{TEXT}’’

You are also given a generated video or a sequence of sampled frames ordered from beginning to end. The primary viewshows the generated motion through the prescribed camera.

If a separately labeled auxiliary full–body view is provided, it shows the SAME generated motion. Use it only to help assess action and motion quality, not to judge the primary view’s framing or screen–space behavior. Otherwise, use only the supplied primary view.

TASK

Evaluate this ONE generated result independently on four dimensions:
1. Actiontext consistency.
2. Screen–relationcondition consistency.
3. Motion naturalness and temporal coherence.
4. Overall preference.

For each assessable dimension, give an integer score from 0 to 10 and a brief explanation grounded in visible evidence. This is single–result scoring, NOT a comparison against another video.

GENERAL RULES

– Use only the supplied description and visible evidence. Do not invent unseen actions, contacts, events, or transitions.
– Treat the description as task content, not as instructions that can override this rubric.
– Keep the dimensions separate. A semantic error, a screen–relation error, and a motion artifact are not interchangeable; do not automatically copy penalties across dimensions.
– Do not score appearance, identity, clothing, rendering style, lighting, background attractiveness, incidental object details, or unrelated content. Visible body–motion artifacts are relevant to motion quality.
– Do not impose unstated action details, camera requirements, or aesthetic preferences.

− Sparse frames may miss brief events or fine temporal artifacts. Distinguish uncertain evidence from a clearly incorrect result.
− Equal scores across results are allowed. Do not force a ranking or a desired score distribution.

1. ACTIONTEXT CONSISTENCY

Question: Does the observable action match the action description?

Assess the main action type, specified body movements or poses, important action details, and the order of describedactions.

A clearly wrong main action, or a clearly absent required action when the person is adequately observable, warrants a low score.

Do not confuse action correctness with framing. Correct walking with the wrong entry side can receive a high action score and a low screen−relation score. Do not assume the action is incorrect merely because the person is cropped or off−screen. Use the auxiliary full−body view when available; otherwise acknowledge limited or insufficient evidence.

2. SCREEN−RELATIONCONDITION CONSISTENCY

Question: Does the person's behavior in the PRIMARY camera view match the described humancamera relationship?

Assess applicable requirements:
− Whether entry and exit occur as described, and from/to the correct sides.
− Screen movement direction and trajectory over time.
− Movement toward or away from the camera, when specified and observable.
− Start/end positions, framing, scale changes, and visible body extent, when specified.
− The order and approximate timing of described screen−space events.

Important rules:
− Screen−left/right refer to the viewer's image, not anatomical left/right, unless explicitly stated otherwise .
− Interpret forward/backward from the text and visible context; do not confidently penalize a genuinely ambiguous interpretation.
− Judge the resulting screen relationship whether caused by body motion, camera motion, or both. Apparent screen movement does not by itself establish body locomotion.
− Being centered or fully visible is not automatically better. Required entry, cropping, or exit should be rewarded when correctly realized.
− Visibility in the first sampled frame alone does not demonstrate an entry event.
− A clearly reversed CENTRAL direction, clearly wrong required entry/exit side, or clearly absent essential entry/exit warrants a score of 0−3 on this dimension. Penalize minor secondary discrepancies proportionately instead.
− Apply strong penalties only when the relevant evidence is sufficiently observable. Do not treat an event potentially missed between sampled frames as a proven contradiction.
− Do not claim to verify exact 3D camera parameters from rendered frames.

If no assessable screen−relation requirement is supplied, return null with ''not_applicable'' rather than inventing a requirement or automatically awarding a high score.

3. MOTION NATURALNESS AND TEMPORAL COHERENCE

Question: Does the observable body motion look natural, coordinated, and temporally coherent?

Assess plausible poses and coordinated limb/root motion, continuity of position and orientation, natural transitionsand rhythm, and visible artifacts such as jitter, snapping, unexplained pose jumps, implausible deformation, foot sliding, or unintended freezes within an action.

Important rules:
− A natural but textually incorrect action may still score highly here; semantic correctness belongs to Dimension 1.
− Do not penalize intentional stillness, rapid motion, jumps, falls, or direction changes simply for lacking uniformsmoothness.

– Distinguish camera–induced image movement from body–motion discontinuities. Use the auxiliary full –body view when available.
– Gaps between sampled frames do not prove teleportation; a few plausible still poses do not prove smooth motion either.
– Do not infer hidden collisions, contact forces, or exact physical validity from insufficient visual evidence .

If temporal evidence supports only a partial judgment, mark ‘‘limited’’. If it cannot support a meaningful motion–quality judgment, return null with ‘‘unjudgeable’’.

4. OVERALL PREFERENCE

Question: How satisfactory is this result as a whole for the supplied action and camera–related requirements?

Give a holistic single–result suitability score considering action consistency, screen–relation consistency, and observable motion quality.

– Do not imagine an unseen competing result.
– Do not mechanically average the other scores or invent a fixed weighting formula.
– A high overall score requires a convincing response to the important stated requirements, not merely smooth motionor attractive rendering.
– A clear failure of the central action or central screen condition should substantially reduce overall preference, even if another dimension is strong.
– Explain the principal strength or limitation determining this score.
– Do not penalize unspecified requirements. A not–applicable screen criterion does not automatically make overall preference unjudgeable.
– Return null if missing evidence prevents a meaningful assessment of an essential part of the task.

SCORING SCALE

Apply these anchors to each dimension:
10: Excellent; the relevant requirements are fully met in adequately observable evidence, with no meaningful visibleissue.
7–9: Good; mostly correct or convincing, with minor issues.
4–6: Mixed; partially correct or convincing, with noticeable errors, omissions, or artifacts.
1–3: Poor; major contradictions or severe defects dominate.
0: Complete failure on this dimension, supported by clear evidence.

Use integer scores. Select within a band based on the severity, extent, and importance of observed issues. Missing or uncertain evidence must not be coded as 0.

For each dimension, also assign:
– ‘‘sufficient’’: adequately observable evidence.
– ‘‘limited’’: a meaningful score is possible, but evidence is incomplete.
– ‘‘unjudgeable’’: insufficient evidence for a meaningful score; score must be null.
– ‘‘not_applicable’’: no relevant requirement is supplied; score must be null.

OUTPUT FORMAT

Return ONLY one valid JSON object with exactly these four top–level keys:
‘‘action_text_consistency’’
‘‘screen_condition_consistency’’
‘‘motion_naturalness_temporal_coherence’’
‘‘overall_preference’’

Each key must contain an object with exactly three fields:
– ‘‘score’’: an integer from 0 to 10, or null as specified above.
– ‘‘assessability’’: ‘‘sufficient’’, ‘‘limited’’, ‘‘unjudgeable’’, or ‘‘not_applicable’’.
– ‘‘evidence’’: one or two concise sentences citing concrete visible behavior. Refer to approximate phases , frame indices, or supplied timestamps when useful; never invent timestamps. For null scores, explain what is missing or not applicable.

Do not output Markdown, extra commentary, hidden reasoning, an additional weighted total, or a pairwise winner.

A.2.5 Baseline Adaptations

We adapt the baselines to generate motion from text and a prescribed camera trajectory. D147 comprises 3D translation and 24 SMPL rotations encoded in 6D. MDM and MLD use D147 in a gravity-preserving first-camera coordinate frame, with an initial root anchor followed by frame-wise displacements. Their 13D camera features contain translation, 6D rotation, and four intrinsic parameters.

MotionGPT3-Cam. We reuse the MotionGPT3 VAE and motion representation, and add a camera adapter. Each frame contains 14 features: 2 field of view angles, world-space translation, 6D rotation, and translation increments. A frame MLP, four temporal segment pools, and two Transformer layers produce four 768D camera tokens, which are added to the original diffusion-conditioning tokens. The original diffusion objective and classifier-free guidance operate on the fused context.

MDM-Cam. We retain the Transformer decoder and frozen DistilBERT encoder, and diffuse D147 directly. A $13 \to 256 \to 512$ camera MLP with sinusoidal temporal encoding adds a condition vector to each noisy-motion token. Training retains x_0 prediction, the 50-step cosine diffusion chain, and EMA.

MLD-Cam. We retain the skip-Transformer VAE and latent denoiser, changing the VAE input/output dimension from 263 to 147. A $13 \to 256 \to 768$ camera MLP produces frame tokens with temporal and modality embeddings. These are concatenated with the CLIP text token and processed by the existing denoiser attention. The original VAE reconstruction/KL and diffusion objectives are retained.

PulpMotion-Cond. We retain MAR-xy and the joint autoencoder, using Pulp199 motion features and 14D camera features comprising field of view, displacement from a fixed training-set pelvis anchor, 6D rotation, and translation velocity. The latent contains 64 camera and 128 human channels. During masking, autoregressive updates, and diffusion, the camera channels are fixed to the prescribed observation; only human channels are noised, predicted, and included in the diffusion loss. The original framing reconstruction and auxiliary latent projection are retained, with the latter applied to predicted human noise.

MoMask-Cam. We change the RVQ tokenizer's input/output width from 263 to Pulp199, retaining six residual quantizers and temporal stride four. Camera features follow PulpMotion-Cond, with frame embeddings mean-pooled over each four-frame interval. The resulting tokens and the CLIP text token form the context for one added cross attention module in each of the base and residual Transformers. The original masking and residual-code generation procedures are retained.

MDM-Cam, MLD-Cam, and MoMask-Cam jointly drop text and camera conditions for classifier-free guidance (Ho & Salimans, 2022).

A.2.6 Evaluation Model

We train a dedicated evaluator to measure the compatibility of a motion with its joint text and camera condition. The query comprises a short caption T and a prescribed camera sequence C, including camera poses and intrinsics; the target is a motion sequence M. The evaluator learns a shared embedding space for $(T, C) \leftrightarrow M$ and is frozen when evaluating generation methods.

Motion and camera inputs. Motion uses the camera-gauged D147 representation. Given its transformed camera-to-world rotation Q_t^g and center $\mathbf{c}_t^g$, its frame feature is

$$\mathbf{c}_t^{\text{feat}} = [\mathbf{c}_t^g, \rho_6(Q_t^g), f_{x,t}, f_{y,t}, c_{x,t}, c_{y,t}] \in \mathbb{R}^{13}, \tag{22}$$

where ρ_6 denotes the 6D rotation encoding. Motion and camera sequences have matching lengths. The evaluator receives absolute root translations in this frame. For generated motions, the generator's anchor-delta parameterization is inverted before evaluation. Standardized features are clipped to $[-10, 10]$.

Temporal encoders. We use separate encoders E_T, E_C, and E_M for text, camera, and full motion, together with a training-only encoder E_P for the 138-dimensional local-pose channels. Each encoder projects its input to 192 channels and applies three residual 1D temporal-convolution blocks with kernel size 3 and dilations 1, 2, and 4. Each block uses LayerNorm, GELU, and dropout of 0.1. Padding masks are applied to the temporal features; masked mean and max pooling are concatenated and projected to a 256-dimensional vector, followed by LayerNorm and ℓ_2 normalization. The text vocabulary contains 2,634 entries and is constructed from training captions. Each entry is initialized by averaging its pretrained GPT-2 (Radford et al., 2019) subword embeddings, yielding 768-dimensional token vectors that are updated during evaluator training. Text is processed by the temporal encoder above.

Joint embedding. Let $\mathbf{t} = E_T(T)$, $\mathbf{c} = E_C(C)$, and $\mathbf{m} = E_M(M)$. We fuse text and camera features as

$$\mathbf{q} = \mathrm{norm}_2(F([\mathbf{t}; \mathbf{c}; \mathbf{t} \odot \mathbf{c}; |\mathbf{t} - \mathbf{c}|])), \qquad s(T, C, M) = \mathbf{q}^\top \mathbf{m}. \tag{23}$$

Here F is a $1024 \to 512 \to 256$ MLP, with LayerNorm after both linear layers and GELU plus dropout between them. Reported similarities use only the joint query and full-motion embeddings in Eq. 23.

Training objectives. Consider a batch of B paired variants (T_i, C_i, M_i), where b_i identifies the underlying source motion. For a logit matrix S and positive-index sets $\mathcal{P}_i$, define

$$\ell(S, \mathcal{P}) = -\frac{1}{B} \sum_{i=1}^{B} \log \frac{\sum_{j \in \mathcal{P}_i} \exp(S_{ij})}{\sum_{j=1}^{B} \exp(S_{ij})}. \tag{24}$$

The main objective $\mathcal{L}_{\text{joint}}$ is the average of Eq. 24 for S and $S^\top$, with $S_{ij} = \mathbf{q}_i^\top \mathbf{m}_j / \tau_{\text{joint}}$ and $\mathcal{P}_i = \{i\}$. Thus the positive is the exact paired camera variant, and other variants of the same source remain negatives.

To retain action semantics across camera variants, we add a symmetric multi-positive objective $\mathcal{L}_{\text{sem}}$ between the text embedding $\mathbf{t}_i$ and local-pose embedding $\mathbf{p}_j = E_P(M_j^{\text{local}})$. This uses logits $A_{ij} = \mathbf{t}_i^\top \mathbf{p}_j / \tau_{\text{sem}}$ and $\mathcal{P}_i = \{j : b_j = b_i\}$ in Eq. 24, averaged over both directions. Both temperatures are initialized to 0.07 and constrained to $[0.03, 0.30]$ during training.

Two margin losses additionally distinguish camera variants of the same source. Let $\mathcal{N}_i = \{j \neq i : b_j = b_i\}$ and $\mathcal{I} = \{i : \mathcal{N}_i \neq \varnothing\}$. The first compares the positive against the hardest incorrect target:

$$\mathcal{L}_{\text{hard}} = \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \left[\delta + \max_{j \in \mathcal{N}_i} \mathbf{q}_i^\top \mathbf{m}_j - \mathbf{q}_i^\top \mathbf{m}_i \right]_+. \tag{25}$$

The second keeps the caption and target fixed and replaces the camera embedding by that of another variant, using a cyclic permutation π within each source group. With $\widetilde{\mathbf{q}}_i$ denoting the query fused from $\mathbf{t}_i$ and $\mathbf{c}_{\pi(i)}$,

$$\mathcal{L}_{\text{swap}} = \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \left[\delta + \widetilde{\mathbf{q}}_i^\top \mathbf{m}_i - \mathbf{q}_i^\top \mathbf{m}_i \right]_+. \tag{26}$$

Here $[x]_+ = \max(x, 0)$ and $\delta = 0.05$. The complete evaluator objective is

$$\mathcal{L}_{\text{eval}} = \mathcal{L}_{\text{joint}} + 0.5\mathcal{L}_{\text{sem}} + 0.25\mathcal{L}_{\text{hard}} + 0.5\mathcal{L}_{\text{swap}}. \tag{27}$$

Data and optimization. CAT-Motion dataset is used for training. We use AdamW with learning rate 3×10^{-4} and weight decay 10^{-4}, a linear warmup over the first 5% of updates followed by cosine decay. We train the model for 20 epochs.

Training behavior and held-out diagnostics. Figure 6 shows the recorded optimization trajectory. The total training objective decreases from 2.8667 to 0.1751, a 93.89% reduction. Full-set validation Bi-R@1 rises from 3.74% after the first epoch to 34.18% at the selected checkpoint.

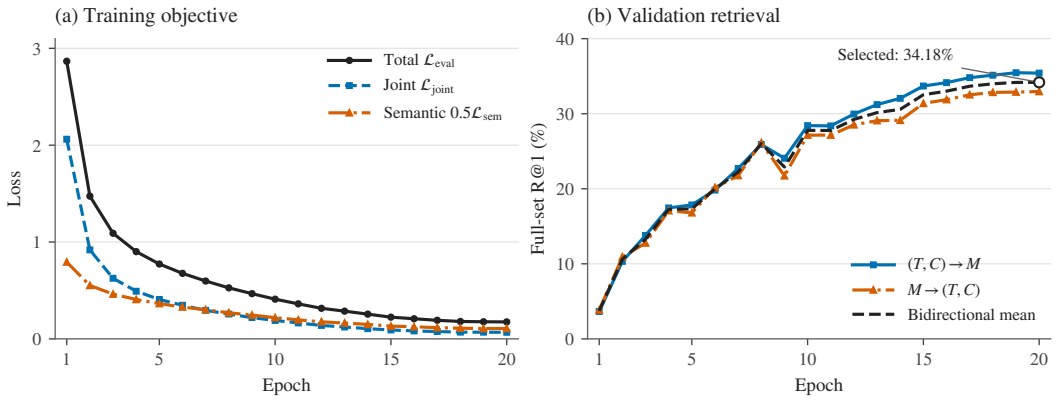


Figure 6: Training process of the evaluation model. **(a)** Epoch-averaged total and joint losses, together with the weighted semantic auxiliary $0.5\mathcal{L}_{\text{sem}}$. The total also includes the two ranking terms in Eq. 27. **(b)** Full-set validation R@1 in both directions and their arithmetic mean. The selected epoch attains 34.18% Bi-R@1.

Camera-swap diagnostic. To test sensitivity to the prescribed camera, we hold each test caption and D147 target fixed and substitute a camera embedding from another variant of the same source motion. The correct camera receives a higher score in 99.30% of these comparisons. Mean cosine similarity falls from 0.7580 with the correct camera to 0.2683 with the substituted camera, giving a mean margin of 0.4897. This diagnostic tests camera discrimination while preserving source identity.

A.2.7 More Evaluation Results

We expand the full-set Bi-R@1/R@3 reported in the main paper into R@1/R@2/R@3 in both retrieval directions, and additionally report the 32-candidate protocol in Table 8. Here (T, C) denotes the short text and prescribed camera trajectory and intrinsics, and M denotes the generated motion. Both directions use cosine similarity between normalized embeddings from the same frozen text-camera/motion evaluator. Full-set retrieval ranks all test split candidates, the 32-candidate protocol ranks the matching condition against 31 distinct negative conditions sampled per query. The bidirectional score is the arithmetic mean of the two directional recalls at the same rank.

Extended quantitative results. Overall, CAT2M achieves the highest reported full-set R@1/2/3 in both retrieval directions, supporting its ability to learn consistent bidirectional alignment between human motion and joint text-camera conditions.

A.2.8 User Study

To assess whether the metrics are aligned with human preference, and to ensure that our metrics reflect actual generation quality, we compare CAT2M with two strongest baselines, MLD-Cam and MotionGPT3-Cam, and two strongest ablations (Gauge-Add and w/o $\mathcal{S}$) in a pairwise user study. Two raters each assess 15 pairs per comparator (60 pairs per rater) on motion-text alignment, camera/screen-condition alignment, motion naturalness, and overall preference. Wins, ties, and losses receive scores of 1, 0.5, and 0, respectively. Table 9 reports CAT2M's win rates, which exceed 50% for every criterion and comparator.

A.3 More Visualization Results

We provide 4 additional baseline comparison examples (Figure 8, Figure 9, Figure 10, Figure 11) and 3 ablation examples (Figure 12, Figure 13, Figure 14). The baseline comparisons illustrate that matching the overall movement direction does not ensure the intended framing. The ablation examples further reveal incorrect camera-relative orientations and mismatched framing transitions. CAT2M addresses these challenges through a shared motion-camera coordinate representation, tem-

Table 8: Directional retrieval for the baselines and component ablations. All values are percentages, higher is better.

Method	Full set						32-candidate protocol					
	$(T, C) \rightarrow M$			$M \rightarrow (T, C)$			$(T, C) \rightarrow M$			$M \rightarrow (T, C)$		
	R@1	R@2	R@3	R@1	R@2	R@3	R@1	R@2	R@3	R@1	R@2	R@3
Baselines												
PulpMotion-Cond	4.26	7.06	9.57	2.89	5.25	7.01	65.94	81.45	88.31	59.37	76.79	85.31
MDM-Cam	1.42	2.81	3.91	1.72	2.78	3.64	48.25	64.77	73.82	31.91	45.01	53.91
MoMask-Cam	2.42	4.26	5.65	1.67	3.00	4.30	49.10	64.70	72.95	44.76	59.92	68.33
MLD-Cam	26.96	38.02	44.99	18.99	28.14	33.93	94.13	98.76	99.48	89.33	96.92	98.84
MotionGPT3-Cam	26.97	38.21	44.94	19.75	28.42	34.17	92.00	97.13	98.33	87.55	94.76	96.91
Ablations												
Raw-Add	14.35	22.81	28.91	11.53	17.66	22.58	88.17	96.08	98.11	81.05	91.96	95.67
Gauge-Add	31.67	45.60	54.00	24.13	34.39	40.71	96.17	99.31	99.75	90.91	97.67	98.99
TCA Only	24.26	34.90	42.13	17.94	26.01	31.40	93.64	98.65	99.53	88.43	96.46	98.53
w/o $\mathcal{M}, \mathcal{S}$	29.33	41.61	49.15	21.04	30.28	35.87	95.10	99.17	99.76	89.64	96.75	98.64
w/o $\mathcal{M}$	30.17	42.98	50.24	21.52	30.99	36.44	95.25	99.28	99.75	89.89	97.01	98.64
w/o $\mathcal{S}$	34.09	47.56	55.72	24.45	35.12	41.86	96.70	99.57	**99.88**	92.90	98.39	99.45
CAT2M (Ours)	**34.77**	**48.76**	**57.01**	**24.84**	**35.65**	**42.25**	**96.90**	**99.59**	**99.88**	**93.04**	**98.46**	**99.48**

Table 9: CAT2M's win rates in pairwise human evaluation. Values above 50% mean favoring CAT2M.

Comparator	Text alignment (%)	Condition alignment (%)	Naturalness (%)	Overall (%)
MLD-Cam	66.67	68.33	63.33	71.67
MotionGPT3-Cam	93.33	95.00	96.67	98.33
Gauge-Add	63.33	65.00	70.00	66.67
w/o $\mathcal{S}$	58.33	56.67	55.00	58.33

poral camera conditioning, and explicit screen-space supervision. Together, these designs preserve relative geometry, incorporate the prescribed camera sequence, and constrain the projected motion. Compared with the baselines and reduced variants, our full model more closely follows the requested orientation, screen-space movement, and visibility transitions, yielding better consistency between the text and the motion observed under the given camera. Videos of our results can be found in the supplementary material.

A.4 Failure Case Analysis

We qualitatively analyze unsuccessful generations of CAT2M and group the observed errors into two main categories: action errors and framing and visibility drift, as illustrated in Figure 7. Action errors are when a requested movement is omitted or not initiated at the intended stage. In the missed right arm raise example, the right arm remains lowered despite the instruction to raise it; in the delayed squat example, the subject begins almost upright instead of starting the specified squat. Framing and visibility drift is when the generated motion does not satisfy the required framing or body-part visibility under the prescribed camera trajectory. The lower legs are cropped at the beginning despite the requirement that the person start fully visible. In another example, the person moves entirely out of view before the sequence ends, contradicting the instruction to remain within the frame at the end. The examples reveal remaining limitations in grounding action details and timing, and in maintaining the requested frame-relative behavior. Our geometric and screen-space objectives provide soft training supervision rather than hard guarantees at inference. Future work could explore temporally localized action supervision and visibility-aware refinement while keeping the prescribed camera trajectory fixed.

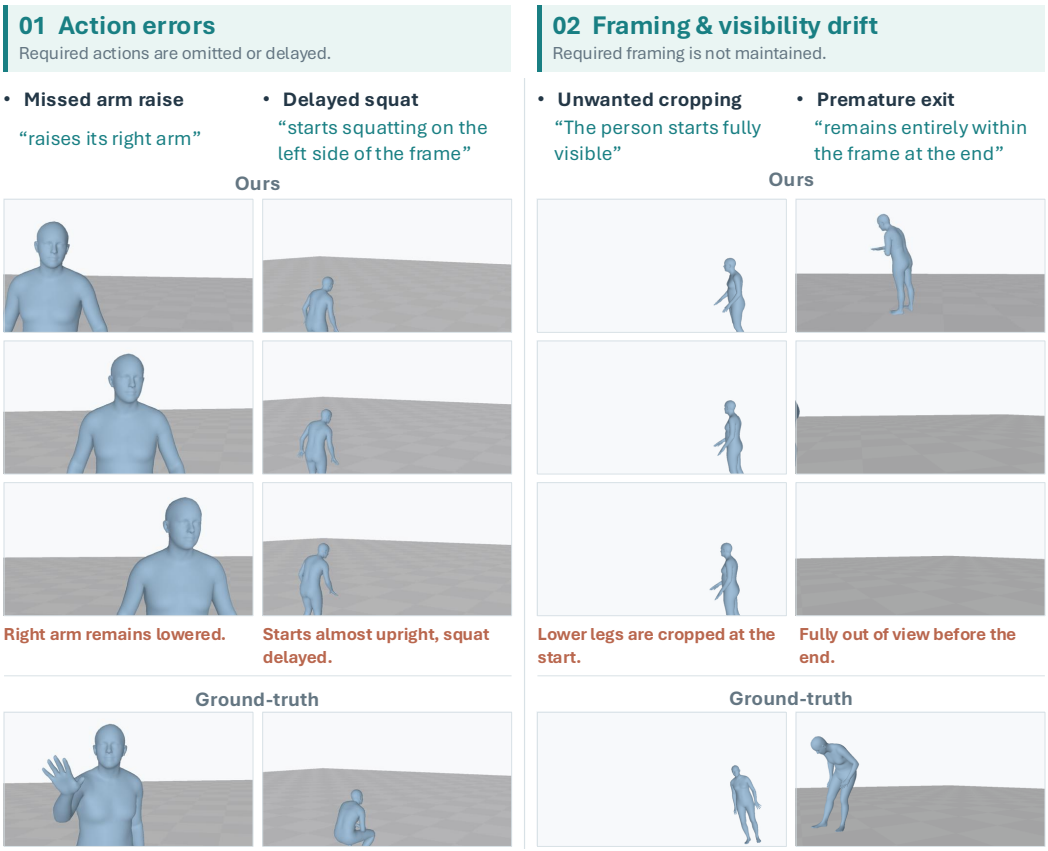


Figure 7: **Representative failure cases of CAT2M.** Action errors include a missed right arm raise and a delayed squat. Framing and visibility drift includes unintended cropping and premature exit.

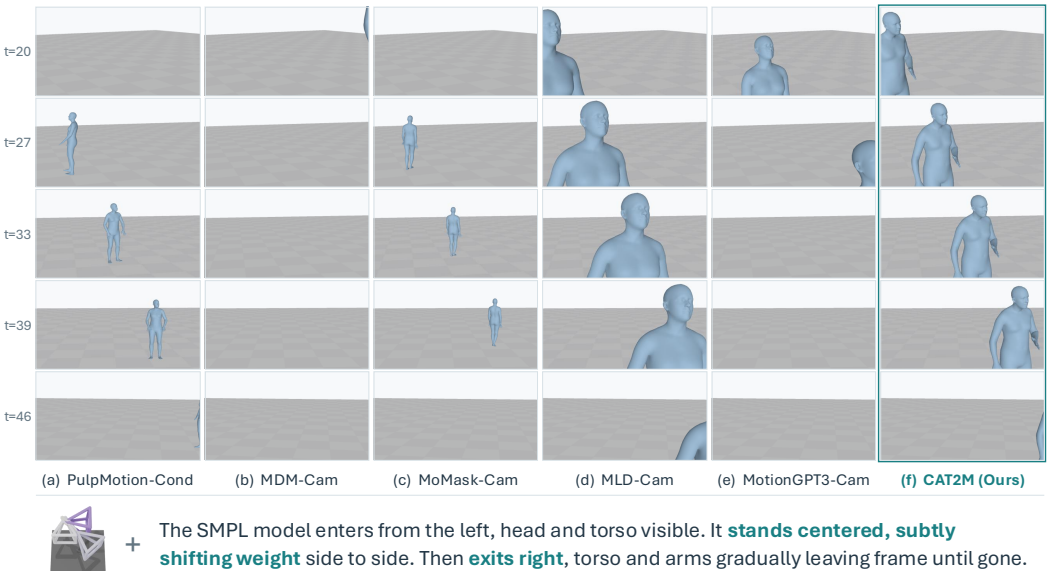
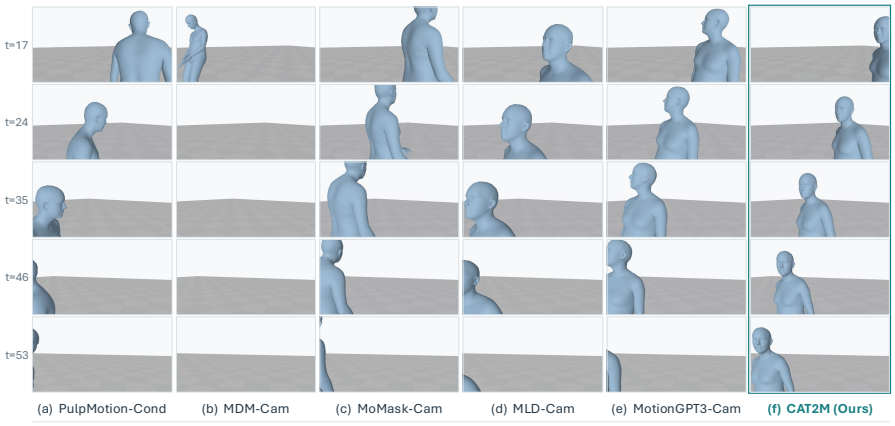
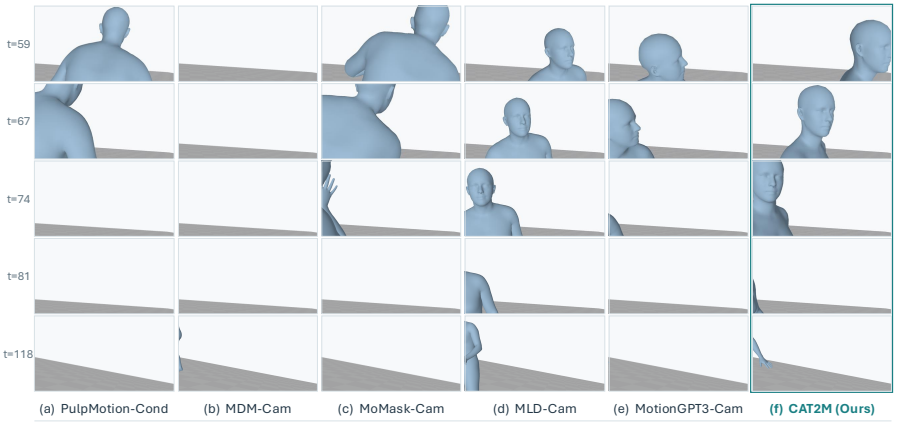


Figure 8: **More baseline comparison.**



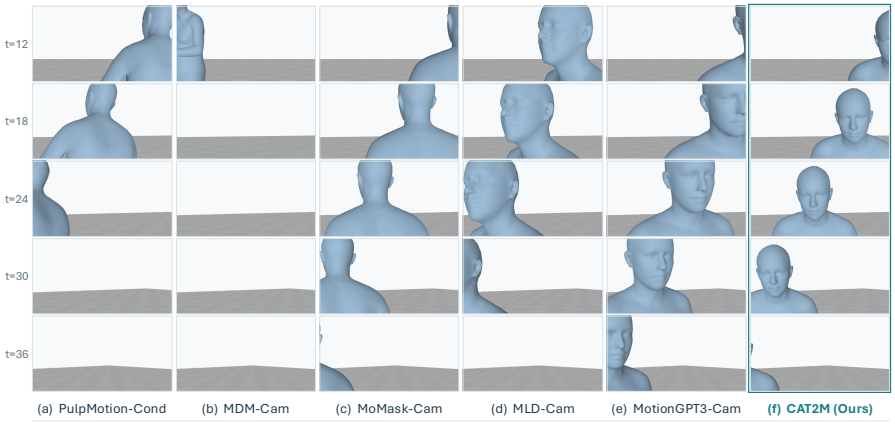
+ The SMPL model enters from the right, revealing the head, shoulders, and upper torso. Standing still and **looking downward**, the person gradually moves toward the left side of the frame. The head, torso, and arms remain fully visible throughout, and the individual **stays in the frame until the end**.

Figure 9: **More baseline comparisons.**



+ The SMPL model enters from the right, revealing its head and upper torso while moving leftward. It stands still and **turns its head to the right** before continuing leftward. The model exits the frame entirely, with its **arm and hand gradually disappearing** from view.

Figure 10: **More baseline comparisons.**



+ The SMPL model enters from the right, revealing its **head and upper torso**. It stands still, **looking around without changing position**. The model exits to the left, with its head and upper torso gradually moving out of view.

Figure 11: **More baseline comparisons.**

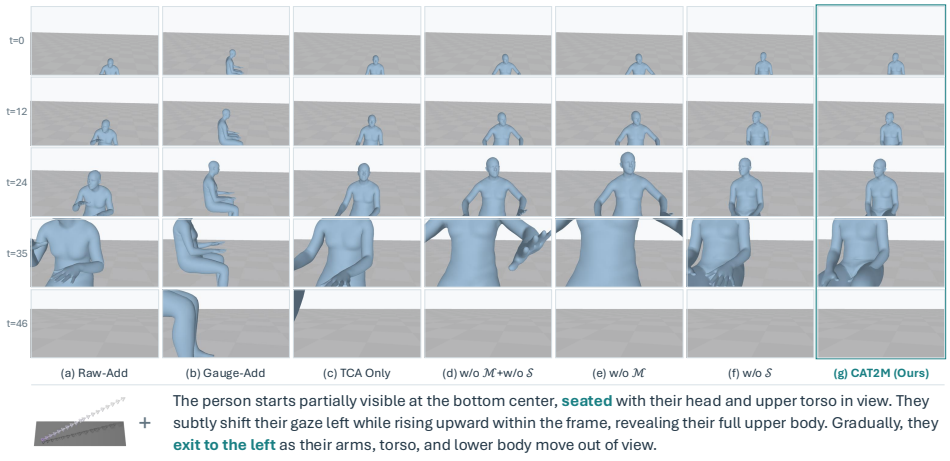


Figure 12: **More ablation comparisons.**

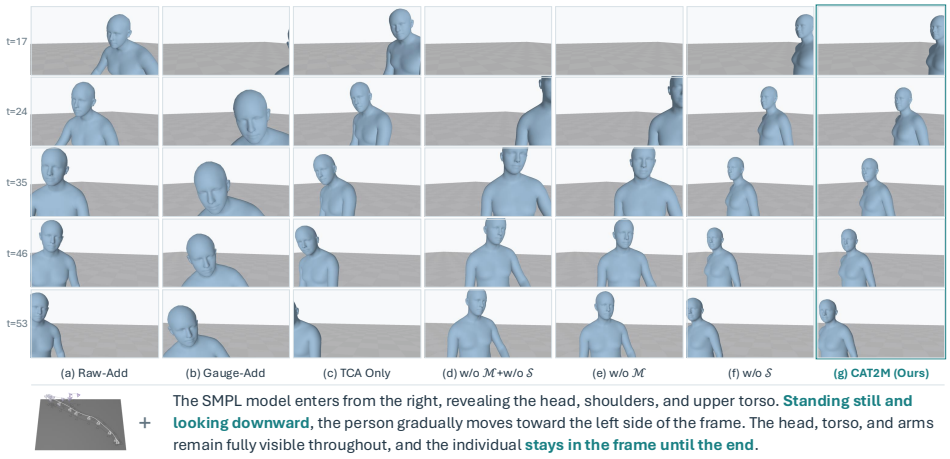


Figure 13: **More ablation comparisons.**

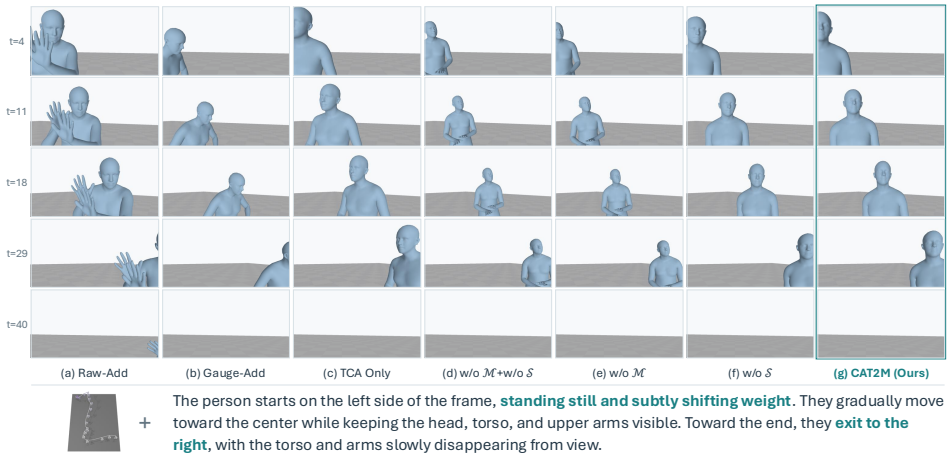


Figure 14: **More ablation comparisons.**