PULP MOTION: FRAMING-AWARE MULTIMODAL CAMERA AND HUMAN MOTION GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Treating human motion and camera trajectory generation separately overlooks a core principle of cinematography: the tight interplay between actor performance and camera work in the screen space. In this paper, we are the first to cast this task as a text-conditioned joint generation, aiming to maintain consistent on-screen framing while producing two heterogeneous, yet intrinsically linked, modalities: human motion and camera trajectories. We propose a simple, model-agnostic framework that enforces multimodal coherence via an auxiliary modality: the on-screen framing induced by projecting human joints onto the camera. This on-screen framing provides a natural and effective bridge between modalities, promoting consistency and leading to more precise joint distribution. We first design a joint autoencoder that learns a shared latent space, together with a lightweight linear mapping from the human and camera latents to a framing latent. We then introduce Auxiliary Sampling, which exploits this linear map to steer generation toward a coherent framing modality. To support this task, we also introduce the PulpMotion dataset, a camera-motion and human-motion dataset with rich captions, and high-quality human motions. Extensive experiments across DiT- and MAR-based architectures show the generality and effectiveness of our method in generating on-frame coherent camera-human motions, while also achieving gains on textual alignment for both modalities. Our qualitative results yield more cinematographically meaningful framings setting the new state of the art for this task. Code, models and data will be made publicly available.

1 INTRODUCTION

Cinematography is inherently a collaborative task, shaped by the joint relationship between the actor and the director. On the one hand, the director's camera seeks to frame the actors, adjusting to their movements to capture the desired performance on screen. On the other hand, the actor must also remain attentive to the presence of the camera, *e.g.* pausing at a marker until the camera arrives, before continuing a movement. Such motions are not spontaneous but rather intentional. These carefully crafted choices aim at enhancing the cinematic aesthetics ultimately presented on the silver screen. The pursuit of this synergy between actors and cameras, *i.e.* balancing naturalistic performance with the demands of visual on-screen composition, remains one of the central challenges in filmmaking.

Prior work has typically addressed only one side of this joint problem, treating them as standalone modalities: either human motion generation (Zhang et al., 2024; Tevet et al., 2023; Jiang et al., 2024a) or camera trajectory generation (Jiang et al., 2024b; Courant et al., 2024; Zhang et al., 2025a), but never both simultaneously. In this work, we introduce the text-conditioned task of jointly generating human motion and camera trajectories. This task is challenging, as any mismatch between motion and camera may lead to poor framing, how the characters are spatially presented on the screen, or even empty frames (*e.g.*, the subject moving out of view). The root problem of this joint generative task, referred to in computer vision as multimodal generation, is to produce high-quality outputs for each modality while maintaining multimodal coherence.

Multimodal generation has been widely studied in domains such as video–audio (Ruan et al., 2023; Hayakawa et al., 2025) and image–text (Li et al., 2025; Xu et al., 2023b). However, most approaches either rely solely on paired data to capture multimodal relationships (Xie et al., 2025; Li et al., 2025; Swerdlow et al., 2025), explicitly enforce correlations through architectural or algorithmic

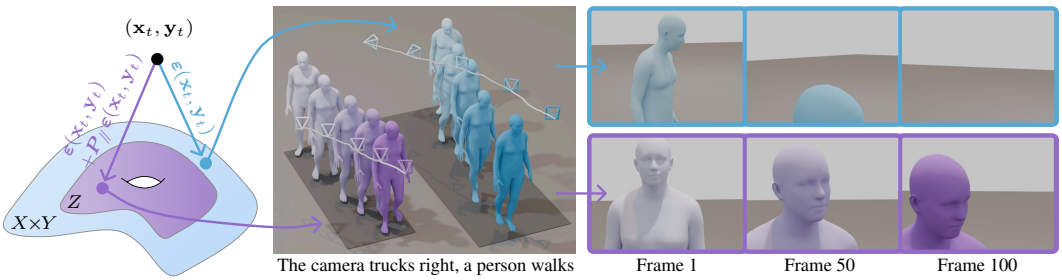


Figure 1: **Overview of our proposed auxiliary sampling.** We adapt the joint generation of $(\mathbf{x}, \mathbf{y})$ (camera trajectories and human motion) by leveraging an auxiliary modality $\mathbf{z}$ (on-screen human framing) to steer sampling toward of coherent joint generation via an orthogonal projection $\boldsymbol{P}_{/\!/}$. Specifically, our diffusion model predicts noise $\varepsilon(\mathbf{x}, \mathbf{y})$, which is then adjusted along the auxiliary guidance direction.

designs (Hu et al., 2023; Ruan et al., 2023; Xu et al., 2023b; Tang et al., 2023), or require training adaptations or sampling guided by models trained on external data (Bao et al., 2023; Hayakawa et al., 2025; Xing et al., 2024; Kouzelis et al., 2025).

Instead, in this work, we propose a multimodal generation framework that leverages an auxiliary modality as a natural bridge between generated modalities to enhance multimodal coherence. Concretely, we use the on-screen human framing within the camera as an auxiliary modality to enforce coherence between generated human motion and camera trajectories.

Our framework consists of two stages: (1) learning a joint latent space for human motion and camera trajectories, along with a linear transform mapping them into the bridging modality, *i.e.* the on-screen framing. This linear transform captures the relationship between the generated and auxiliary modalities directly in the latent space. (2) a sampling process augmented with an additional term derived from this linear relationship, steering generation towards coherent multimodal generation. For evaluation, we present PulpMotion, an extended version of a prior human-camera dataset, with more samples, motion captions, and higher-quality motion. We benchmark our approach on this dataset for both DiT-based (Peebles & Xie, 2023) and MAR-based (Li et al., 2024) architectures to demonstrate the generality and model-agnosticity of our approach. Our results show consistently improved coherence between generated motion and trajectories, yielding better framing quality and lower out-of-frame rates while preserving strong motion and trajectory generation performance.

Our contributions are: (1) a unified framework that jointly generates human motion and camera trajectories leveraging an auxiliary modality (on-screen framing) to enforce multimodal coherence during sampling, (2) the PulpMotion dataset, an extension of the prior human-camera dataset with more samples, motion captions, and higher-quality human motion, and (3) an extensive evaluation across multiple architectures demonstrating the method's generality and effectiveness.

2 Related Work

Human motion generation. Diffusion-based approaches have driven recent progress (Ho et al., 2020; Rombach et al., 2022; Tevet et al., 2023; Kim et al., 2023; Zhang et al., 2024) on human motion generation, with extensions for efficient latent spaces, fast sampling, stronger textual alignment, and leverage of external data (Chen et al., 2023; Dai et al., 2024; Andreou et al.; Zhang et al., 2023b). The newly proposed MAR architecture combines autoregressive and diffusion modeling and further pushes state-of-the-art performance (Li et al., 2024; Meng et al., 2024; Xiao et al., 2025).
However, most methods treat motion as a *single modality*; although interactions with objects, people, and scenes are increasingly modeled (Peng et al., 2025b; Geng et al., 2025; Liang et al., 2024; Fan et al., 2024; Shan et al., 2024; Wang et al., 2024b; Cen et al., 2024), joint human–camera generation remains largely unexplored. Existing efforts typically use camera parameters only as constraints or conditioning, rather than modeling their joint distribution with motion (Patel & Black, 2025; Ye et al., 2023; Wang et al., 2024a; Kocabas et al., 2024; Sun et al., 2023).

Camera Trajectory Generation. Camera control has evolved from handcrafted rules to learning-based methods that either learn cinematic from example videos or optimize trajectories in differen-

tiable 3D space (Blinn, 1988; Lino & Christie, 2015; Drucker et al., 1992; Jiang et al., 2020; 2021; Wang et al., 2023; Jiang et al., 2024d; Chen et al., 2024). To reduce reliance on exemplary data, RL is often applied on drones and indoor scenes (Huang et al., 2019; Bonatti et al., 2020; Xie et al., 2023), but remains environment-specific and style-limited. Diffusion-based camera generation, coupled with new datasets, further advances text-conditioned control and reduces reliance on curated reference videos (Jiang et al., 2024b; Courant et al., 2024; Wang et al., 2024e;d; Zhang et al., 2025a). However, similarly to human motion generation, camera generation is also often regarded as a *single modality* problem conditioned on motion, rather than modeling the joint motion–camera distribution. In this work, we bridge this gap by adding human motion into the camera trajectory generation pipeline, modelling the synergy between how and what to film.

Multimodal generation. Most multimodal generation works leverage paired data to implicitly capture joint distribution, e.g, text–image unified generation has been explored with different architectures: Dual Diffusion (Li et al., 2025) employs a DiT-based design, while Show-o (Xie et al., 2025) adopts an autoregressive backbone plus diffusion head framework. However, in practice, relying solely on paired data to learn implicit multimodal correlations often requires large datasets and still fails to fully capture multi-modal relationships.

Therefore, some works explicitly enforce multimodal correlations through architectural or algorithmic design. Hu et al. (2023) introduce a unified transition that compresses discrete representations across modalities under a discrete diffusion framework. MM Diffusion (Ruan et al., 2023) exploits a similar idea. It employs multimodal attention and random shifts to align multimodal information. Xu et al. (2023b) emphasizes architectural separation by disentangling context and data layers to encourage joint conditioning. Alternatively, CoDi (Tang et al., 2023) modifies cross-attention layers to emphasize a pre-aligned, modality-specific latent space, enabling any-to-any generation across multiple modalities. Despite their effectiveness on specific tasks, these approaches often depend on handcrafted architectures which limit their generality and adaptability across tasks and models.

Another line of work focuses on adapting only the training or the sampling process, avoiding architectural modifications. For instance, UniDiffuser (Bao et al., 2023) trains a single multimodal diffusion with independent timesteps for each modality and applies an adapted classifier-free guidance scheme (CFG) (Ho & Salimans, 2021) with modality-specific timesteps. MMDisco (Hayakawa et al., 2025), inspired by classifier guidance (Dhariwal & Nichol, 2021), enables video–audio generation by training a joint discriminator to construct a guidance term during sampling, which is also used as regularisation for finetuning. Meanwhile, some works leverage foundation models to implicitly exploit larger datasets and stronger representations. For example, Xing et al. (2024) use the pre-trained multimodal binder ImageBind (Girdhar et al., 2023) to align generations via classifier-like guidance. Similarly, Kouzelis et al. (2025) design a representation-guidance term based on a diffusion generator trained on paired DINOv2 (Oquab et al., 2024) and image data, preserving the joint distribution without requiring an explicit classifier. However, these approaches still rely on adapting training or using large external pre-trained models, such as Imagebind or DINOv2.

In this work, we leverage an auxiliary modality to bridge the target modalities, steering sampling toward coherent joint generation in an architecture-agnostic manner, without requiring training adaptations or pre-trained models on external data. See extended discussion in Appendix B.

3 METHOD

Problem formulation. We consider a sample as a pair of camera trajectory $\mathbf{x}$ and human motion $\mathbf{y}$; both are sequences of F frames. We aim to generate both modalities with respect to a textual description c, specifying the desired camera trajectory and human motion, *i.e.*, sampling from the joint camera–human distribution $p(\mathbf{x}, \mathbf{y}|\mathrm{c})$.

Our approach. Most related works capture cross-modal relationships by relying exclusively on paired data, crafting specific architectural or algorithmic designs, or introducing training adaptations or sampling strategies guided by external models. In contrast, our approach strengthens the connection between modalities using an auxiliary modality, $\mathbf{z}$, which explicitly bridges them. In our setting, $\mathbf{z}$ represents the on-screen human framing, *i.e.* the 2D projection of human joints in the camera view, a natural characteristic of the human-camera relationship.

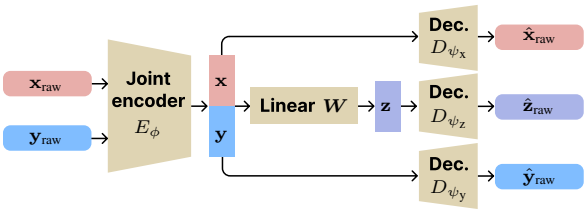


Figure 2: **Architecture of the multimodal autoencoder.** Camera motion $\mathbf{x}_{\text{raw}}$ and human trajectory $\mathbf{y}_{\text{raw}}$ are jointly encoded by E_ϕ, linearly transformed via $\boldsymbol{W}$ into an auxiliary on-screen framing latent $\mathbf{z}$. Three decoders D_{ψ_x}, D_{ψ_y}, and D_{ψ_z} reconstruct raw modalities: $\hat{\mathbf{x}}_{\text{raw}}, \hat{\mathbf{y}}_{\text{raw}}, \hat{\mathbf{z}}_{\text{raw}}$.

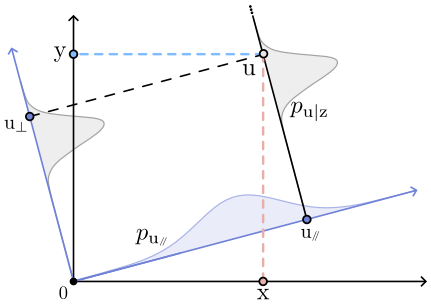


Figure 3: **Illustration of the decomposition.** $\mathbf{u} = [\mathbf{x}, \mathbf{y}]^\top$ decomposes onto two orthogonal components $\mathbf{u}_\perp + \mathbf{u}_{/\!/} \in \ker \boldsymbol{W} \oplus \ker(\boldsymbol{W})^\perp$. Our auxiliary sampler leverages this to encourage samples along $\mathbf{u}_{/\!/}$: the parallel term to the modality $\mathbf{z}$.

Next, we describe our multimodal latent space with a latent linear transform on the auxiliary modality (Section 3.1) and present our auxiliary sampling scheme, which leverages the relationship between generated modalities and the auxiliary modality (Section 3.2).

3.1 Multimodal Latent Space

In multimodal generation, different modalities often exhibit varying properties, such as scale or geometric structure, which makes direct generation in the raw modality space challenging (Tang et al., 2023; Xing et al., 2024). Moreover, operating directly in the raw space increases computational and memory costs (Rombach et al., 2022) and can destabilize some diffusion losses (Meng et al., 2024).

To address these challenges, we adopt a latent diffusion framework for our joint human-camera generation task. Our latent representation is designed with two key aspects: (1) instead of embedding modalities separately, all modalities are mapped into a shared latent space to align them, and (2) a lightweight learnable linear transform $\boldsymbol{W}$ that maps the latent camera and human representations into an on-screen framing latent, which bridges both modalities.

More specifically, we propose an autoencoder architecture, shown in Figure 2. The model first maps the camera $\mathbf{x}_{\text{raw}}$ and human $\mathbf{y}_{\text{raw}}$ with a joint encoder E_ϕ, producing latent embeddings $\mathbf{x}$ and $\mathbf{y}$. A learnable linear transform $\boldsymbol{W}$ then maps these embeddings into a on-screen framing latent $\mathbf{z}$:

$$\mathbf{z} = \boldsymbol{W}\,[\mathbf{x}, \mathbf{y}]^\top. \tag{1}$$

Finally, three *independent* decoders D_{ψ_x}, D_{ψ_y} and D_{ψ_z} reconstruct each raw modality $(\mathbf{x}_{\text{raw}}, \mathbf{y}_{\text{raw}}, \mathbf{z}_{\text{raw}})$ from its respective latent[1]. The model is trained end-to-end with the following reconstruction loss:

$$\begin{aligned}\mathcal{L}_{\text{AE}}(\phi, \psi_c, \psi_h, \psi_p) = \|D_{\psi_x}(E_\phi(\mathbf{x}_{\text{raw}}, \mathbf{y}_{\text{raw}})) - \mathbf{x}_{\text{raw}}\|^2 + \|D_{\psi_y}(E_\phi(\mathbf{x}_{\text{raw}}, \mathbf{y}_{\text{raw}})) - \mathbf{y}_{\text{raw}}\|^2 \\ + \|D_{\psi_z}(\boldsymbol{W}\,E_\phi(\mathbf{x}_{\text{raw}}, \mathbf{y}_{\text{raw}})) - \mathbf{z}_{\text{raw}}\|^2.\end{aligned} \tag{2}$$

Note that the on-screen framing is never directly encoded; it is learned exclusively via the linear transform from the camera and human latents and supervised only through its reconstruction loss.

3.2 Auxiliary Sampling

Given the multimodal latent space established in the previous section, we now introduce our multimodal latent diffusion framework, which incorporates an auxiliary sampling technique during inference to enhance cross-modal coherence.

[1]Recall $\mathbf{z}_{\text{raw}}$ is defined as the 2D projection of human joints in the camera view, see Section 5.1 for details for $\mathbf{x}_{\text{raw}}, \mathbf{y}_{\text{raw}}, \mathbf{z}_{\text{raw}}$.

We train a generative model to produce multimodal representations of human motion $\mathbf{x}$ and camera trajectories $\mathbf{y}$ from textual descriptions c. For this, we adopt the standard Denoising Diffusion Probabilistic Model (DDPM) framework (Ho et al., 2020):

$$\mathcal{L}_{\text{noise}}(\theta) = \mathbb{E}_{t,\boldsymbol{\varepsilon}_{\mathbf{xy}}}\left[\|\boldsymbol{\varepsilon}_{\mathbf{xy}} - \boldsymbol{\varepsilon}_\theta\left(\mathbf{x}_t, \mathbf{y}_t, \mathrm{c}\right)\|^2\right]. \tag{3}$$

Auxiliary sampling. Controllability in diffusion models is typically achieved via classifier-free guidance (CFG) (Ho & Salimans, 2021) over a conditioning signal $\boldsymbol{c}$:

$$\begin{aligned}\nabla_{\mathbf{x}_t,\mathbf{y}_t} \log \tilde{p}(\mathbf{x}_t, \mathbf{y}_t|\mathrm{c}) = \nabla_{\mathbf{x}_t,\mathbf{y}_t} \log p(\mathbf{x}_t, \mathbf{y}_t) \\ + w_c(\nabla_{\mathbf{x}_t,\mathbf{y}_t} \log p(\mathbf{x}_t, \mathbf{y}_t|\mathrm{c}) - \nabla_{\mathbf{x}_t,\mathbf{y}_t} \log p(\mathbf{x}_t, \mathbf{y}_t)).\end{aligned} \tag{4}$$

Here, CFG explicitly splits the score into an unconditional term and a conditional term, with the latter scaled by w_c. In our case, we aim to control the cross-modal coherence between $\mathbf{x}$ and $\mathbf{y}$ through $\mathbf{z}$. Following the CFG strategy, we split the unconditional score term $\nabla_{\mathbf{x}_t,\mathbf{y}_t} \log p(\mathbf{x}_t, \mathbf{y}_t)$ in Equation (4) into a new unconditional component and an additional term in $\mathbf{z}$. To this end, we leverage the relationship in Equation (1) that links $\mathbf{z}$ to $\mathbf{x}$ and $\mathbf{y}$.
Let $\mathbf{u} = [\mathbf{x}, \mathbf{y}]^\top$. Since $\mathbf{z}$ is a compressed representation of $\mathbf{u}$ (*i.e.*, $\dim(\mathbf{z}) < \dim(\mathbf{u})$), it cannot fully capture all information in $\mathbf{u}$. We therefore decompose $\mathbf{u}$ into a $\mathbf{z}$-dependent component $\mathbf{u}_{/\!/}$ and a complementary orthogonal component $\mathbf{u}_\perp$, such that $\mathbf{u} = \mathbf{u}_\perp + \mathbf{u}_{/\!/}$ as illustrated in Figure 3. This decomposition is precisely what we aim for: the component $\mathbf{u}_{/\!/}$ characterized by $\mathbf{z}$, steers the sampling toward a coherent $\mathbf{u}$, while the complementary component acts as an "unconditional" term.

Lemma 3.1. *Let $\boldsymbol{P}_{/\!/}$ denote the projection onto the orthogonal space of* $\ker(\boldsymbol{W})$. *Then, we have:*

$$\mathbf{u}_\perp := (\boldsymbol{I} - \boldsymbol{P}_{/\!/})\mathbf{u} \sim \mathcal{N}((\boldsymbol{I} - \boldsymbol{P}_{/\!/})\boldsymbol{\mu}, \sigma^2(\boldsymbol{I} - \boldsymbol{P}_{/\!/})) \quad \textit{and} \quad \mathbf{u}_{/\!/} := \boldsymbol{P}_{/\!/}\mathbf{u} \sim \mathcal{N}(\boldsymbol{P}_{/\!/}\boldsymbol{\mu}, \sigma^2\boldsymbol{P}_{/\!/}), \tag{5}$$

and the density of $\mathbf{u}$ *decomposes as*

$$p(\mathbf{u}) = p(\mathbf{u}_\perp)\, p(\mathbf{u}_{/\!/}). \tag{6}$$

Since $\boldsymbol{W}^\top\boldsymbol{W}$ is invertible in our setting, $\boldsymbol{P}_{/\!/}$ can be expressed as $\boldsymbol{P}_{/\!/} = \boldsymbol{W}(\boldsymbol{W}^\top\boldsymbol{W})^{-1}\boldsymbol{W}^\top$. See Figure 3 for illustration and Section C.2 for complete development and proof.

Thus, using Equation (6), the first term in Equation (4) can be split into an *unconditional* component over $(\mathbf{x}_t, \mathbf{y}_t)$ and a $\mathbf{z}$-dependent component weighted by w_z that guides sampling toward $\mathbf{z}_t$:

$$\begin{aligned}\nabla_{\mathbf{x}_t,\mathbf{y}_t} \log \tilde{p}(\mathbf{x}_t, \mathbf{y}_t|\mathrm{c}) = \nabla_{\mathbf{x}_t,\mathbf{y}_t} \log p(\mathbf{u}_\perp) + (1 + w_z)\nabla_{\mathbf{x}_t,\mathbf{y}_t} \log p(\mathbf{u}_{/\!/}) \\ + w_c(\nabla_{\mathbf{x}_t,\mathbf{y}_t} \log p(\mathbf{x}_t, \mathbf{y}_t|\mathrm{c}) - \nabla_{\mathbf{x}_t,\mathbf{y}_t} \log p(\mathbf{x}_t, \mathbf{y}_t)).\end{aligned} \tag{7}$$

Finally, we perform sampling using the following linear combination of the model's predictions, recalling that $\boldsymbol{\varepsilon}(\mathbf{x}) \propto \nabla_{\mathbf{x}} \log p(\mathbf{x})$, with more detailed derivation included in Section C.4:

$$\begin{aligned}\boldsymbol{\varepsilon}\left(\mathbf{x}_t, \mathbf{y}_t, \mathrm{c}, t\right) = \boldsymbol{\varepsilon}\left(\mathbf{x}_t, \mathbf{y}_t, \emptyset, t\right) \\ + w_z\boldsymbol{P}_{/\!/}\, \boldsymbol{\varepsilon}\left(\mathbf{x}_t, \mathbf{y}_t, \emptyset, t\right) \\ + w_c(\boldsymbol{\varepsilon}\left(\mathbf{x}_t, \mathbf{y}_t, \mathrm{c}, t\right) - \boldsymbol{\varepsilon}\left(\mathbf{x}_t, \mathbf{y}_t, \emptyset, t\right)).\end{aligned} \tag{8}$$

Since the final sampling does not explicitly condition on $\mathbf{z}_t$, there is no need to include the bridging modality during training. This reduces training cost and yields a more general approach.

4 PULPMOTION DATASET

Training a joint human–camera model requires paired data of human motions and camera trajectories. However, as shown in Table 1, most prior works provide only one modality, focusing either on human (e.g., HumanML3D (Guo et al., 2022a)) or on camera (e.g., RealEstate10k (Zhou et al., 2018)). More recently, the E.T. dataset (Courant et al., 2024) provides paired samples, but it prioritizes camera aspect, with lower-quality human motions and missing rich textual captions, making it inappropriate to train human motion models. This motivates us to introduce the PulpMotion dataset, a joint human-camera dataset with good-quality human motions along with motion captions.

Section 4.1 is the overview of the PulpMotion and Section 4.2 describes the extraction pipeline.

Table 1: **Comparison of human and camera datasets.** We compare PulpMotion with existing human motion and/or camera trajectory datasets. We summarize modality coverage, available captions, dataset size (hours, frames, samples), sample length statistics (median, mean, std), and vocabulary size.

Dataset	Camera		Human		#Hours	#Frames	#Samples	Sample lengths (frames)			#Vocabulary
	Traj	Caption	Motion	Caption				Median	Mean	Std	
RealEstate10k (Zhou et al., 2018)	✓	✗	✗	✗	121	11M	79K	115	136.9	80.0	-
CamVid-30K (Zhao et al.)	✓	✗	✗	✗	-	-	30K	-	-	-	-
DynPose100k (Rockwell et al., 2025)	✓	✗	✗	✗	157	6.8M	100K	63	67.97	17.91	-
CameraBench (Lin et al., 2025)	✗	✓	✗	✗	-	-	4K	-	-	-	-
CCD (Jiang et al., 2024b)	✓	✓	✗	✗	50	4.5M	25K	189	180.4	69.6	48
DataDoP (Zhang et al., 2025a)	✓	✓	✗	✗	113	11M	29K	-	424.8	-	8,698
KIT-ML (Plappert et al., 2016)	✗	✗	✓	✓	12	0.8M	4K	71	99.0	99.6	1,623
HumanML3D (Guo et al., 2022a)	✗	✗	✓	✓	29	2M	14K	147	140.0	57.50	5,371
Motion-X++ (Zhang et al., 2025b)	✗	✗	✓	✓	181	19.5M	120K	152	167.9	125.33	8,116
E.T. (Courant et al., 2024)	✓	✓	✓	✗	120	11M	115K	75	93.9	73.8	1,790
PulpMotion(Ours)	✓	✓	✓	✓	314	22M	193K	107	117.3	63.6	4,599

Table 2: **Motion refinement and text-motion alignment.** We report metrics on the PulpMotion dataset, comparing raw extracted motions (Wang et al., 2024a) with our refined motions. Captions are generated either from human motions using m2t model (Jiang et al., 2024a) or from RGB frames using our VLM-based approach (Bai et al., 2025).

Motion	Caption	TMR-Score ↑	R1 ↑	R2 ↑	R3 ↑
Extracted	M2T	4.08	1.16	2.47	3.63
Extracted	VLM	8.06	3.65	6.64	9.20
Refined	M2T	8.54	2.29	4.24	5.77
Refined	VLM	**16.22**	**4.84**	**8.86**	**12.34**

Table 3: **Motion refinement and motion quality.** We compare PulpMotion motion samples with HumanML3D (Guo et al., 2022a), evaluating raw extracted motions (Wang et al., 2024a) against our refined motions, using either m2t captions from human motions (Jiang et al., 2024a) or our VLM-based captions from RGB frames (Bai et al., 2025).

Motion	Caption	FD_{TMR} ↓	P ↑	R ↑	D ↑	C ↑
Extracted	-	595.39	0.53	0.13	0.32	0.15
Refined	M2T	505.45	0.50	0.19	0.30	0.17
Refined	VLM	**447.69**	**0.55**	**0.21**	**0.37**	**0.21**

4.1 Dataset description and comparison

Table 1 compares PulpMotion with existing human and camera datasets. Our dataset stands out by providing all modalities, camera trajectories and captions, human motions and captions, while most prior datasets cover only a subset. With 193K samples and 314 hours, PulpMotion is also the largest, nearly doubling the number of samples in E.T. (115K) and surpassing other motion-centric datasets such as Motion-X++ (120K). In terms of temporal coverage, PulpMotion exhibits longer sequences, with a median length of 107 frames, a mean of 117.3 frames, and a standard deviation of 63.6, indicating both richer and more diverse motion content compared to previous datasets.

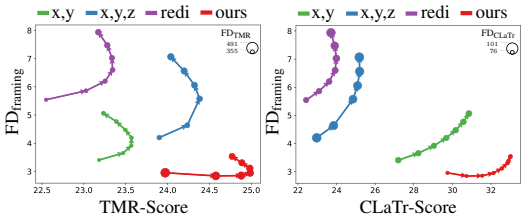
4.2 Extraction pipeline

Human-camera pair extraction. Following the E.T. extraction pipeline, we use TRAM (Wang et al., 2024a) to obtain 3D human-camera poses from videos of the CondensedMovies dataset (Bain et al., 2020) and apply the same post-processing steps (filtering, smoothing and cropping to a maximum length of 300 frames). For PulpMotion, we replace SLAHMR (Ye et al., 2023) with TRAM because it is significantly faster (∼1 fps vs. <0.1 fps), enabling large-scale processing. Moreover, TRAM provides higher-quality estimates, allowing us to keep trajectories that the E.T. pipeline previously filtered.

Human-camera captions generation. We generate detailed human motion captions inspired by HumanML3D (Guo et al., 2022a) using the Qwen2.5-VL (Bai et al., 2025) vision-language model, prompted with video clips and bounding boxes of the target character. The VLM follows annotation guidelines similar to HumanML3D. To assess the captioning quality we compute the motion-text alignment metrics (*i.e.* cosine similarity and retrieval recall) based on the TMR features Petrovich et al. (2023). As shown in Table 2, our method achieves higher text-motion alignment than existing motion-to-text models Jiang et al. (2024a), attaining a TMR-Score of 8.06 against 4.08.

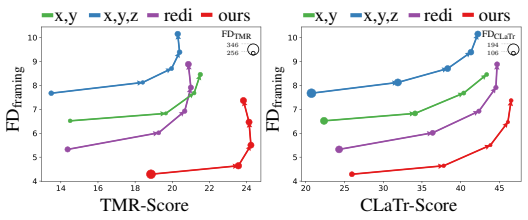
For camera captions, we follow the E.T. methodology: performing motion tagging and inputting it to a large language model (LLM) to produce user-friendly descriptions.

Human motion refinement. TRAM's output often contains lower-quality human motion compared to mocap-based datasets like HumanML3D. To address this, we introduce a refinement step to en-



(a) $FD_{framing}$-TMR-Score (b) $FD_{framing}$-CLaTr-Score

Figure 4: **Comparison in DiT on the mixed set.** Framing quality and modality-text alignment for c guidance ranges from 5 to 11. The optimal region is at the bottom-right (low framing FD, high alignment).



(a) $FD_{framing}$-TMR-Score (b) $FD_{framing}$-CLaTr-Score

Figure 5: **Comparison in MAR on the mixed set.** Framing quality and modality-text alignment for c guidance ranges from 1 to 5. The optimal region is at the bottom-right (low framing FD, high alignment).

hance motion quality. The main source of error in TRAM arises from partial observations; e.g., close-up shots capturing only the upper body. Therefore, to improve overall motion quality, we detect out-of-frame body parts via camera reprojection and refine these regions using the RePaint editing method (Lugmayr et al., 2022) with a HumanML3D-pretrained diffusion model. To assess the captioning quality we compute the motion quality metrics (*i.e.* Fréchet distance and PRDC Naeem et al. (2020)) based on the TMR features Petrovich et al. (2023). As shown in Table 3, this step significantly reduces the FD_{TMR} score from 595.39 to 447.69.
We provide additional details on the dataset extraction pipeline in the Appendix D.1.

5 Experiments

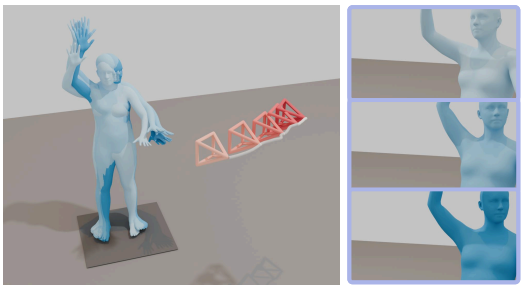
5.1 Experimental Setup

Data representation.
Framing features z_{raw}. We use the 2D Normalized Device Coordinates (NDC): screen-projected coordinates normalized to the range $[-1, 1]$, of nine key human joints (ankles, pelvis, spine, head, shoulders, and wrists). For a sequence of F frames, this yields $\boldsymbol{X}_{\text{framing}} \in \mathbb{R}^{F\times 18}$.
Camera feature x_{raw}. We extend the features from Courant et al. (2024) by notably adding the intrinsics. For a trajectory of F frames: $\boldsymbol{X}_{\text{cam}} = (\boldsymbol{R}, \dot{\boldsymbol{T}}, \boldsymbol{D}, \boldsymbol{F}) \in \mathbb{R}^{F\times 14}$ where $\boldsymbol{R} \in \mathbb{R}^{F\times 6}$ denotes rotation using the 6D continuous representation (Zhou et al., 2019), $\dot{\boldsymbol{T}} \in \mathbb{R}^{F\times 3}$ is the linear velocity, $\boldsymbol{D} \in \mathbb{R}^{F\times 3}$ is the relative distance bewteen the camera and the human, and $\boldsymbol{F} \in \mathbb{R}^{F\times 2}$ encodes the horizontal and vertical fields of view (assuming the principal point lies at the image center).
Human features y_{raw}. We use the features introduced in Petrovich et al. (2024), for a motion of F frames: $\boldsymbol{X}_{\text{human}} = (\boldsymbol{r}_z, \dot{\boldsymbol{r}}_x, \dot{\boldsymbol{r}}_y, \dot{\boldsymbol{\alpha}}, \boldsymbol{\Theta}, \boldsymbol{J}) \in \mathbb{R}^{F\times 199}$ where $\boldsymbol{r}_z \in \mathbb{R}^F$ is the Z (up) coordinate of the pelvis, $\dot{\boldsymbol{r}}_x \in \mathbb{R}^F$ and $\dot{\boldsymbol{r}}_y \in \mathbb{R}^F$ are the linear velocities of the pelvis, $\dot{\boldsymbol{\alpha}} \in \mathbb{R}^F$ is the angular velocity of the Z angle of the body, $\boldsymbol{\Theta} \in \mathbb{R}^{F\times 132}$ are the 22 first SMPL (Loper et al., 2023) pose parameters (6D representation (Zhou et al., 2019)), and $\boldsymbol{J} \in \mathbb{R}^{F\times 63}$ are the 22 joints positions (pelvis excluded).

Metrics.
Framing metrics. Since no existing metrics assess framing quality, we propose two metrics based on the 9-joint NDC representation introduced above. First, the Fréchet distance $FD_{framing}$ measures how well the on-screen framing of the generated camera and human matches a reference distribution. Second, the *Out-rate* is the fraction of frames where none of the 9 joints appear on-screen.
Camera metrics. We use the metrics introduced in Courant et al. (2024). To evaluate the camera trajectory quality, we report the FD_{CLaTr} and CLaTr-based coverage (Naeem et al., 2020); to evaluate the camera trajectory coherence, we report the CLaTr-Score and segmentation F1.
Human metrics. We use the standard text-to-motion metrics (Guo et al., 2020) using the TMR (Petrovich et al., 2023) feature space. We then report FD_{TMR} and TMR-Score, and TMR-based R-precision. In addition, to evaluate how well generated samples span the variety of real data we compute the TMR-based coverage (Naeem et al., 2020).

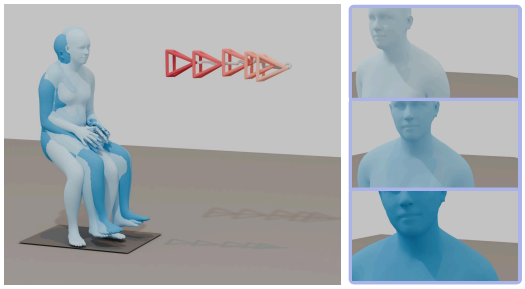
Table 4: **State-of-the-art comparison on the mixed subset.** We compare four baselines: independent modality generation $(\mathbf{x})(\mathbf{y})$, dual-modality generation $(\mathbf{x}, \mathbf{y})$, triplet-modality generation $(\mathbf{x}, \mathbf{y}, \mathbf{z})$, and ReDi (Kouzelis et al., 2025), along with our auxiliary sampling (Aux). Results are reported for DiT (Peebles & Xie, 2023) and MAR (Li et al., 2024). Superscript $\pm$ denotes the 95% confidence interval over 10 samplings.

Methods	Framing		Human				Camera			
	$FD_{framing}$ ↓	Out-rate ↓	FD_{TMR} ↓	TMR-Score ↑	R3 ↑	Coverage ↑	FD_{CLaTr} ↓	CLaTr-Score ↑	F1 ↑	Coverage ↑
Ground-truth	0.00	0.89	0.00	17.72	22.00	1.00	0.00	68.88	87.43	1.00
Auto-encoder	0.23	4.61	124.78	18.16	21.81	85.30	15.64	57.98	67.04	86.64
DiT										
$(\mathbf{x})(\mathbf{y})$	$11.21^{\pm 0.12}$	$48.02^{\pm 0.24}$	$357.99^{\pm 0.52}$	$25.03^{\pm 0.06}$	$4.34^{\pm 0.12}$	$10.55^{\pm 0.17}$	$67.76^{\pm 0.20}$	$46.74^{\pm 0.11}$	$46.71^{\pm 0.24}$	$53.66^{\pm 0.36}$
$(\mathbf{x})(\mathbf{y})$+Aux (ours)	$8.24^{\pm 0.07}$	$41.24^{\pm 0.24}$	$422.45^{\pm 0.78}$	$26.46^{\pm 0.07}$	$4.64^{\pm 0.10}$	$9.08^{\pm 0.15}$	$56.41^{\pm 0.32}$	$50.69^{\pm 0.10}$	$50.72^{\pm 0.14}$	$51.39^{\pm 0.22}$
$(\mathbf{x}, \mathbf{y})$	$4.90^{\pm 0.05}$	$25.98^{\pm 0.24}$	$372.61^{\pm 0.90}$	$23.50^{\pm 0.07}$	$3.67^{\pm 0.08}$	$10.72^{\pm 0.15}$	$87.07^{\pm 0.87}$	$30.75^{\pm 0.17}$	$34.28^{\pm 0.27}$	$51.62^{\pm 0.40}$
$(\mathbf{x}, \mathbf{y}, \mathbf{z})$	$4.18^{\pm 0.03}$	$23.88^{\pm 0.19}$	$390.08^{\pm 1.20}$	$23.88^{\pm 0.12}$	$3.22^{\pm 0.11}$	$11.58^{\pm 0.13}$	$97.45^{\pm 0.61}$	$23.34^{\pm 0.16}$	$27.40^{\pm 0.18}$	$50.80^{\pm 0.44}$
ReDi	$5.57^{\pm 0.04}$	$28.99^{\pm 0.22}$	$360.07^{\pm 1.26}$	$22.48^{\pm 0.06}$	$5.68^{\pm 0.18}$	$12.83^{\pm 0.16}$	$83.66^{\pm 1.05}$	$22.73^{\pm 0.22}$	$26.53^{\pm 0.20}$	$55.24^{\pm 0.41}$
$(\mathbf{x}, \mathbf{y})$+Aux (ours)	$3.37^{\pm 0.02}$	$16.76^{\pm 0.19}$	$431.54^{\pm 1.15}$	$25.05^{\pm 0.07}$	$3.89^{\pm 0.14}$	$8.91^{\pm 0.13}$	$80.08^{\pm 0.76}$	$32.81^{\pm 0.19}$	$36.06^{\pm 0.25}$	$48.68^{\pm 0.20}$
MAR										
$(\mathbf{x})(\mathbf{y})$	$11.59^{\pm 0.08}$	$51.05^{\pm 0.24}$	$296.01^{\pm 0.73}$	$21.71^{\pm 0.09}$	$11.69^{\pm 0.12}$	$17.48^{\pm 0.23}$	$111.42^{\pm 0.75}$	$51.96^{\pm 0.11}$	$51.69^{\pm 0.11}$	$49.85^{\pm 0.41}$
$(\mathbf{x})(\mathbf{y})$+Aux (ours)	$9.13^{\pm 0.07}$	$47.30^{\pm 0.22}$	$308.90^{\pm 0.65}$	$24.12^{\pm 0.08}$	$12.07^{\pm 0.16}$	$14.64^{\pm 0.19}$	$91.85^{\pm 0.69}$	$55.75^{\pm 0.10}$	$54.78^{\pm 0.18}$	$48.67^{\pm 0.21}$
$(\mathbf{x}, \mathbf{y})$	$8.51^{\pm 0.07}$	$40.75^{\pm 0.28}$	$275.30^{\pm 0.55}$	$21.68^{\pm 0.06}$	$10.60^{\pm 0.19}$	$17.10^{\pm 0.28}$	$117.77^{\pm 0.63}$	$42.84^{\pm 0.14}$	$42.69^{\pm 0.23}$	$54.89^{\pm 0.37}$
$(\mathbf{x}, \mathbf{y}, \mathbf{z})$	$8.66^{\pm 0.09}$	$37.50^{\pm 0.17}$	$268.41^{\pm 0.71}$	$20.13^{\pm 0.08}$	$10.59^{\pm 0.11}$	$19.83^{\pm 0.33}$	$148.12^{\pm 0.96}$	$38.58^{\pm 0.10}$	$38.34^{\pm 0.09}$	$51.74^{\pm 0.41}$
ReDi	$6.96^{\pm 0.07}$	$32.25^{\pm 0.18}$	$275.58^{\pm 0.66}$	$20.84^{\pm 0.08}$	$10.90^{\pm 0.11}$	$18.41^{\pm 0.32}$	$122.40^{\pm 0.77}$	$42.60^{\pm 0.15}$	$42.70^{\pm 0.21}$	$54.96^{\pm 0.51}$
$(\mathbf{x}, \mathbf{y})$+Aux (ours)	$6.42^{\pm 0.04}$	$33.65^{\pm 0.23}$	$301.39^{\pm 0.25}$	$24.46^{\pm 0.07}$	$11.28^{\pm 0.09}$	$14.14^{\pm 0.14}$	$108.74^{\pm 0.46}$	$45.96^{\pm 0.14}$	$45.39^{\pm 0.22}$	$53.67^{\pm 0.38}$



Human: A person raising their right arm.
Camera: The camera performs a trucking right.

Figure 6: **Example with DiT on the mixed subset.**



Human: A person sitting.
Camera: The camera performs a push in.

Figure 7: **Example with MAR on the mixed subset.**

5.2 Comparison to the state of the art

In this section, we compare our **auxiliary sampling** (Aux) method against several baselines: (1) **independent modality generation** $(\mathbf{x})(\mathbf{y})$: two separate models for each modality; (2) **dual-modality generation** $(\mathbf{x}, \mathbf{y})$: a single model generates both modalities without the auxiliary modality; (3) **triplet-modality generation** $(\mathbf{x}, \mathbf{y}, \mathbf{z})$: a single model generates both modalities and the auxiliary modality; and (4) **ReDi** (Kouzelis et al., 2025): a single model for both modalities and the auxiliary modality, with representation sampling leveraging the auxiliary modality. For all baselines and our method, we evaluate using both DiT-based (Peebles & Xie, 2023) and MAR (Li et al., 2024) architectures (see Section E.2.1 for more details on architectures).

Quantitative results. Table 4 reports a comparison of our auxiliary sampling (Aux) method against state-of-the-art baselines across both DiT and MAR architectures on the mixed subset. We summarise our experimental observations as follows:
(i) Auxiliary sampling improves coherence. With our sampling, framing consistently improves: $FD_{framing}$ drops from $11.21 \rightarrow 3.37$ (DiT) and $9.13 \rightarrow 6.42$ (MAR). The out rate is best among baselines for DiT (16.76). We observe that Aux enhances both $(\mathbf{x})(\mathbf{y})$ and $(\mathbf{x}, \mathbf{y})$ settings.
(ii) Strong per-modality performance. Relative to ReDi, Aux improves text–modality alignment: TMR-Score increases from $22.48 \rightarrow 25.05$ (DiT) and $20.84 \rightarrow 24.46$ (MAR), and CLaTr-Score from $22.73 \rightarrow 32.81$ (DiT) and $42.70 \rightarrow 45.39$ (MAR), indicating stronger human–text and camera–text alignment. In terms of fidelity, camera quality improves (FD_{CLaTr}: $97.45 \rightarrow 80.08$ on DiT; $148.12 \rightarrow 108.74$ on MAR), with a minor trade-off in human fidelity (FD_{TMR}: $372.61 \rightarrow 431.54$ on DiT; $275.30 \rightarrow 301.39$ on MAR). Overall, we see Aux enhances multimodal coherence and on-screen framing quality while preserving strong per-modality performance.
Moreover, we compare our method with baselines for DiT and MAR in Figures 4 and 5, showing the trade-off between framing quality ($FD_{framing}$) and modality-text alignment (TMR for human, CLaTr for camera) across different textual guidance values (w_c in Equation 8). The optimal point lies in

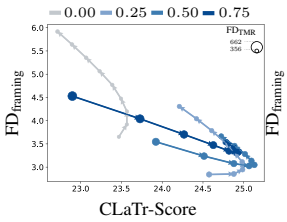
the bottom-right corner of each plot (low $\text{FD}_{\text{framing}}$, high modality scores). Across both architectures and both modalities, our auxiliary sampling achieves the best performance, improving both framing quality and textual alignment, showing its effectiveness and generality.

Qualitative results. Figures 6 and 7 show qualitative results with Aux sampling for DiT and MAR, respectively. In both cases, the generated human motion is precise—for example, for DiT, the person raises the *right* arm as specified. The camera trajectories are also coherent with the prompt, and accurately following the human motion while maintaining correct on-screen framing, with the subject's head consistently in view. These results highlight that Aux produces humans and cameras that are well aligned with the input prompts, achieving precise motion and coherent framing across different architectures. Further examples are provided in the Appendix E.2 we also provide an ***anonymous online gallery*** with additional qualitative videos.

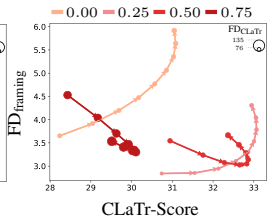
5.3 Ablation Study

Table 5: **Auxiliary guidance ablation on the mixed subset.** We vary the auxiliary guidance weight w_z to evaluate its effect on the framing, camera and human metrics. Results are reported for DiT (Peebles & Xie, 2023) and MAR (Li et al., 2024). Superscript $\pm$ denotes the 95% confidence interval over 10 samplings.

w_z	Framing		Human				Camera			
	$\text{FD}_{\text{framing}}\downarrow$	Out-rate $\downarrow$	$\text{FD}_{\text{TMR}}\downarrow$	TMR-Score $\uparrow$	R3 $\uparrow$	Coverage $\uparrow$	$\text{FD}_{\text{CLaTr}}\downarrow$	CLaTr-Score $\uparrow$	F1 $\uparrow$	Coverage $\uparrow$
DiT										
0.00	$4.90^{\pm 0.05}$	$25.98^{\pm 0.24}$	$372.61^{\pm 0.90}$	$23.50^{\pm 0.07}$	$3.67^{\pm 0.08}$	$10.72^{\pm 0.15}$	$87.07^{\pm 0.87}$	$30.75^{\pm 0.17}$	$34.28^{\pm 0.27}$	$51.62^{\pm 0.40}$
0.25	$3.37^{\pm 0.02}$	$16.76^{\pm 0.19}$	$431.54^{\pm 1.15}$	$25.05^{\pm 0.07}$	$3.89^{\pm 0.14}$	$8.91^{\pm 0.13}$	$80.08^{\pm 0.76}$	$32.81^{\pm 0.19}$	$36.06^{\pm 0.25}$	$48.68^{\pm 0.20}$
0.50	$3.09^{\pm 0.02}$	$11.99^{\pm 0.16}$	$493.53^{\pm 1.64}$	$25.30^{\pm 0.07}$	$7.23^{\pm 0.10}$	$7.09^{\pm 0.17}$	$90.06^{\pm 0.58}$	$32.45^{\pm 0.13}$	$35.98^{\pm 0.09}$	$44.98^{\pm 0.31}$
0.75	$3.37^{\pm 0.02}$	$9.66^{\pm 0.12}$	$548.60^{\pm 1.62}$	$24.99^{\pm 0.07}$	$7.08^{\pm 0.09}$	$5.63^{\pm 0.15}$	$123.16^{\pm 0.75}$	$29.58^{\pm 0.12}$	$30.88^{\pm 0.17}$	$38.88^{\pm 0.34}$
MAR										
0.00	$8.51^{\pm 0.07}$	$40.75^{\pm 0.28}$	$275.30^{\pm 0.55}$	$21.68^{\pm 0.06}$	$10.60^{\pm 0.19}$	$17.10^{\pm 0.28}$	$117.77^{\pm 0.63}$	$42.84^{\pm 0.14}$	$42.69^{\pm 0.23}$	$54.89^{\pm 0.37}$
0.50	$6.42^{\pm 0.04}$	$33.65^{\pm 0.23}$	$301.39^{\pm 0.25}$	$24.46^{\pm 0.07}$	$11.28^{\pm 0.09}$	$14.14^{\pm 0.14}$	$108.74^{\pm 0.46}$	$45.96^{\pm 0.14}$	$45.39^{\pm 0.22}$	$53.67^{\pm 0.38}$
1.00	$5.93^{\pm 0.02}$	$32.09^{\pm 0.17}$	$326.29^{\pm 0.31}$	$25.42^{\pm 0.07}$	$11.35^{\pm 0.14}$	$12.57^{\pm 0.12}$	$144.02^{\pm 0.60}$	$44.05^{\pm 0.16}$	$40.19^{\pm 0.15}$	$47.26^{\pm 0.34}$
1.50	$6.04^{\pm 0.02}$	$32.77^{\pm 0.15}$	$346.14^{\pm 0.42}$	$25.65^{\pm 0.05}$	$11.35^{\pm 0.20}$	$11.63^{\pm 0.19}$	$193.14^{\pm 0.46}$	$40.61^{\pm 0.11}$	$36.64^{\pm 0.17}$	$38.21^{\pm 0.41}$

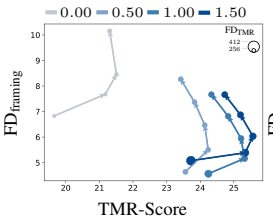


(a) $\text{FD}_{\text{framing}}$-TMR-Score

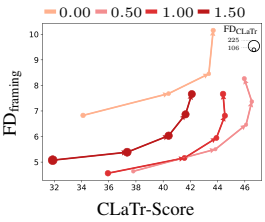


(b) $\text{FD}_{\text{framing}}$-CLaTr-Score

Figure 8: w_z **ablation in DiT on the mixed set.** Framing quality and modality-text alignment for $\boldsymbol{c}$ guidance ranges from 4 to 12. The optimal region is at the bottom-right (low framing FD, high alignment).



(a) $\text{FD}_{\text{framing}}$-TMR-Score



(b) $\text{FD}_{\text{framing}}$-CLaTr-Score

Figure 9: w_z **ablation in MAR on the mixed set.** Framing quality and modality-text alignment for $\boldsymbol{c}$ guidance ranges from 1 to 5. The optimal region is at the bottom-right (low framing FD, high alignment).

To assess controllability and effectiveness of Aux, we ablate in Table 5 the auxiliary guidance weight w_z (Eq (8)) on both DiT and MAR. We see that a (1) moderate guidance weight improves framing and text–modality alignment. On DiT, increasing w_z from 0.00 to 0.25 reduces $\text{FD}_{\text{framing}}$ $4.90 \rightarrow 3.37$ and Out-rate $25.98 \rightarrow 16.76$; on MAR, w_z=0.50 lowers them $8.51 \rightarrow 6.42$ and $40.75 \rightarrow 33.65$. (2) Pushing w_z further continues to improve framing but degrades fidelity: FD_{TMR} and FD_{CLaTr} rise (DiT $431.54 \rightarrow 493.53$, MAR $301.39 \rightarrow 326.29$). (3) At high weights, the trend becomes unstable (w_z=0.75 for DiT, 1.50 for MAR), with FD_{TMR} spiking to 548.60 and FD_{CLaTr} to 193.14.
We then illustrate Figures 8 and 9 for the trade-off between framing quality ($\text{FD}_{\text{framing}}$, lower is better) and text–modality alignment (TMR, CLaTr; higher is better) as the Aux guidance weight w_z varies. The optimum lies near the bottom-right of each plot. Across both architectures, we see: (1) introducing guidance yields a large gain: w_z:0 $\rightarrow$ 0.25 (DiT) and 0.50 (MAR) shift points toward the bottom-right; (2) further increases, 0.50 (DiT), 1.0 (MAR), continue to improve framing but begin to reduce fidelity, reflected by larger markers (higher Fréchet distances); and (3) at very high weights, 0.75 (DiT), 1.50 (MAR), performance degrades on both axes.

Summary of findings. From our experiments and ablations, we find: (1) our proposed Aux sampling consistently improves human–camera coherence (better framing, fewer empty frames) while preserving strong per-modality performance; (2) the gains generalize across architectures, though absolute performance depends on tuning the guidance weight (as with CFG);

6 CONCLUSION

In this paper, we presented a unified framework for joint generation of human motion and camera trajectories, enforcing multimodal coherence via on-screen framing. Extensive evaluations on the proposed PulpMotion dataset demonstrate the generality and effectiveness of our autoencoder with auxiliary sampling. Future work includes extending the auxiliary-modality approach to other domains and enabling finer-grained framing (e.g., targeting specific body parts).

REFERENCES

Nefeli Andreou, Xi Wang, Victoria Fernández Abrevaya, Marie-Paule Cani, Yiorgos Chrysanthou, and Vicky Kalogeiton. Lead: Latent realignment for human motion diffusion. In *Computer Graphics Forum*. Wiley Online Library.

Sherwin Bahmani, Ivan Skorokhodov, Aliaksandr Siarohin, Willi Menapace, Guocheng Qian, Michael Vasilkovsky, Hsin-Ying Lee, Chaoyang Wang, Jiaxu Zou, Andrea Tagliasacchi, et al. Vd3d: Taming large video diffusion transformers for 3d camera control. *arXiv preprint arXiv:2407.12781*, 2024.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Max Bain, Arsha Nagrani, Andrew Brown, and Andrew Zisserman. Condensed movies: Story based retrieval with contextual embeddings. In *ACCV*, 2020.

Fan Bao, Shen Nie, Kaiwen Xue, Chongxuan Li, Shi Pu, Yaole Wang, Gang Yue, Yue Cao, Hang Su, and Jun Zhu. One transformer fits all distributions in multi-modal diffusion at scale. In *ICLR*, 2023.

Jim Blinn. Where am I? what am I looking at? (cinematography). *IEEE Computer Graphics and Applications*, 1988.

R. Bonatti, W. Wang, C. Ho, A. Ahuja, M. Gschwindt, E. Camci, E. Kayacan, S. Choudhury, and S. Scherer. Autonomous aerial cinematography in unstructured environments with learned artistic decision-making. *J. Field Robotics.*, 2020.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *NeurIPS*, 2020.

Zhi Cen, Huaijin Pi, Sida Peng, Zehong Shen, Minghui Yang, Shuai Zhu, Hujun Bao, and Xiaowei Zhou. Generating human motion in 3d scenes from text descriptions. In *CVPR*, 2024.

Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. Maskgit: Masked generative image transformer. In *CVPR*, 2022.

Weiliang Chen, Fangfu Liu, Diankun Wu, Haowen Sun, Haixu Song, and Yueqi Duan. Dreamcinema: Cinematic transfer with free camera and 3d character. *arXiv preprint arXiv:2408.12601*, 2024.

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 *CVPR*, 2023.

Soon Yau Cheong, Duygu Ceylan, Armin Mustafa, Andrew Gilbert, and Chun-Hao Paul Huang. Boosting camera motion control for video diffusion transformers. *arXiv preprint arXiv:2410.10802*, 2024.

W. G. Cochran. The distribution of quadratic forms in a normal system, with applications to the analysis of covariance. *Mathematical Proceedings of the Cambridge Philosophical Society*, 30 (2):178–191, April 1934. ISSN 0305-0041, 1469-8064. doi: 10.1017/S0305004100016595.

Robin Courant, Nicolas Dufour, Xi Wang, Marc Christie, and Vicky Kalogeiton. E.T. the Exceptional Trajectories: text-to-camera-trajectory generation with character awareness. In *ECCV*, 2024.

Wenxun Dai, Ling-Hao Chen, Jingbo Wang, Jinpeng Liu, Bo Dai, and Yansong Tang. Motionlcm: Real-time controllable motion generation via latent consistency model. In *ECCV*. Springer, 2024.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. In *NeurIPS*, 2021.

Steven M Drucker, Tinsley A Galyean, and David Zeltzer. Cinema: A system for procedural camera movements. In *Symposium on Interactive 3D graphics*, 1992.

Ke Fan, Junshu Tang, Weijian Cao, Ran Yi, Moran Li, Jingyu Gong, Jiangning Zhang, Yabiao Wang, Chengjie Wang, and Lizhuang Ma. Freemotion: A unified framework for number-free text-to-motion synthesis. In *ECCV*, 2024.

Zichen Geng, Zeeshan Hayder, Wei Liu, and Ajmal Saeed Mian. Auto-regressive diffusion for generating 3d human-object interactions. In *AAAI*, 2025.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15180–15190, 2023.

Chuan Guo, Xinxin Zuo, Sen Wang, Shihao Zou, Qingyao Sun, Annan Deng, Minglun Gong, and Li Cheng. Action2motion: Conditioned generation of 3d human motions. In *International Conference on Multimedia*, 2020.

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 *CVPR*, 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 *ECCV*, 2022b.

Chuan Guo, Yuxuan Mu, Muhammad Gohar Javed, Sen Wang, and Li Cheng. Momask: Generative masked modeling of 3d human motions. In *CVPR*, 2024.

Akio Hayakawa, Masato Ishii, Takashi Shibuya, and Yuki Mitsufuji. Discriminator-guided cooperative diffusion for joint audio and video generation. *ICLR*, 2025.

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.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS-W*, 2021.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 2020.

Seong-Eun Hong, Soobin Lim, Juyeong Hwang, Minwook Chang, and Hyeongyeop Kang. Bipo: Bidirectional partial occlusion network for text-to-motion synthesis. *arXiv preprint arXiv:2412.00112*, 2024.

Minghui Hu, Chuanxia Zheng, Heliang Zheng, Tat-Jen Cham, Chaoyue Wang, Zuopeng Yang, Dacheng Tao, and Ponnuthurai N Suganthan. Unified discrete diffusion for simultaneous vision-language generation. *ICLR*, 2023.

C. Huang, C. Lin, Z. Yang, Y. Kong, P. Chen, X. Yang, and K. Cheng. Learning to film from professional human motion videos. In *CVPR*, 2019.

Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. Motiongpt: Human motion as a foreign language. 36, 2024a.

H. Jiang, B. Wang, X. Wang, M. Christie, and B. Chen. Example-driven virtual cinematography by learning camera behaviors. *ACM TOG*, 2020.

H. Jiang, M. Christie, X. Wang, L. Liu, B. Wang, and B. Chen. Camera keyframing with style and control. *ACM TOG*, 2021.

Hongda Jiang, Xi Wang, Marc Christie, Libin Liu, and Baoquan Chen. Cinematographic camera diffusion model. *Computer Graphics Forum*, 2024b.

Nan Jiang, Zimo He, Zi Wang, Hongjie Li, Yixin Chen, Siyuan Huang, and Yixin Zhu. Autonomous character-scene interaction synthesis from text instruction. In *SIGGRAPH Asia*, 2024c.

Xuekun Jiang, Anyi Rao, Jingbo Wang, Dahua Lin, and Bo Dai. Cinematic behavior transfer via nerf-based differentiable filming. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6723–6732, 2024d.

Tero Karras, Miika Aittala, Tuomas Kynkäänniemi, Jaakko Lehtinen, Timo Aila, and Samuli Laine. Guiding a diffusion model with a bad version of itself. *NeurIPS*, 37:52996–53021, 2024.

Jihoon Kim, Jiseob Kim, and Sungjoon Choi. Flame: Free-form language-based motion synthesis & editing. In *AAAI*, 2023.

Muhammed Kocabas, Ye Yuan, Pavlo Molchanov, Yunrong Guo, Michael J Black, Otmar Hilliges, Jan Kautz, and Umar Iqbal. Pace: Human and camera motion estimation from in-the-wild videos. In *2024 International Conference on 3D Vision (3DV)*, pp. 397–408. IEEE, 2024.

Theodoros Kouzelis, Efstathios Karypidis, Ioannis Kakogeorgiou, Spyros Gidaris, and Nikos Komodakis. Boosting generative image modeling via joint image-feature synthesis. *arXiv preprint arXiv:2504.16064*, 2025.

Tianhong Li, Yonglong Tian, He Li, Mingyang Deng, and Kaiming He. Autoregressive image generation without vector quantization. *NeurIPS*, 2024.

Zijie Li, Henry Li, Yichun Shi, Amir Barati Farimani, Yuval Kluger, Linjie Yang, and Peng Wang. Dual diffusion for unified image generation and understanding. In *CVPR*, pp. 2779–2790, 2025.

Han Liang, Wenqian Zhang, Wenxuan Li, Jingyi Yu, and Lan Xu. Intergen: Diffusion-based multi-human motion generation under complex interactions. *IJCV*, 2024.

Zhiqiu Lin, Siyuan Cen, Daniel Jiang, Jay Karhade, Hewei Wang, Chancharik Mitra, Tiffany Ling, Yuhan Huang, Sifan Liu, Mingyu Chen, et al. Towards understanding camera motions in any video. *arXiv preprint arXiv:2504.15376*, 2025.

Christophe Lino and Marc Christie. Intuitive and efficient camera control with the toric space. *ACM TOG*, 2015.

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM TOG*, 2023.

Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *CVPR*, 2022.

Zichong Meng, Yiming Xie, Xiaogang Peng, Zeyu Han, and Huaizu Jiang. Rethinking diffusion for text-driven human motion generation. *arXiv preprint arXiv:2411.16575*, 2024.

Muhammad Ferjad Naeem, Seong Joon Oh, Youngjung Uh, Yunjey Choi, and Jaejun Yoo. Reliable fidelity and diversity metrics for generative models. In *ICML*, 2020.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research Journal*, 2024.

Priyanka Patel and Michael J. Black. Camerahmr: Aligning people with perspective. In *2025 International Conference on 3D Vision (3DV)*, pp. 1562–1571, 2025. doi: 10.1109/3DV66043.2025.00146.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, 2023.

Xiangyu Peng, Zangwei Zheng, Chenhui Shen, Tom Young, Xinying Guo, Binluo Wang, Hang Xu, Hongxin Liu, Mingyan Jiang, Wenjun Li, Yuhui Wang, Anbang Ye, Gang Ren, Qianran Ma, Wanying Liang, Xiang Lian, Xiwen Wu, Yuting Zhong, Zhuangyan Li, Chaoyu Gong, Guojun Lei, Leijun Cheng, Limin Zhang, Minghao Li, Ruijie Zhang, Silan Hu, Shijie Huang, Xiaokang Wang, Yuanheng Zhao, Yuqi Wang, Ziang Wei, and Yang You. Open-sora 2.0: Training a commercial-level video generation model in $200k. *arXiv preprint arXiv:2503.09642*, 2025a.

Xiaogang Peng, Yiming Xie, Zizhao Wu, Varun Jampani, Deqing Sun, and Huaizu Jiang. Hoidiff: Text-driven synthesis of 3d human-object interactions using diffusion models. In *CVPR-W*, 2025b.

Mathis Petrovich, Michael J Black, and Gül Varol. TMR: Text-to-motion retrieval using contrastive 3d human motion synthesis. In *ICCV*, 2023.

Mathis Petrovich, Or Litany, Umar Iqbal, Michael J Black, Gul Varol, Xue Bin Peng, and Davis Rempe. Multi-track timeline control for text-driven 3d human motion generation. In *CVPR-W*, 2024.

Ekkasit Pinyoanuntapong, Pu Wang, Minwoo Lee, and Chen Chen. Mmm: Generative masked motion model. In *CVPR*, 2024.

Matthias Plappert, Christian Mandery, and Tamim Asfour. The KIT motion-language dataset. *Big Data*, 2016.

Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, et al. Movie gen: A cast of media foundation models. *arXiv preprint arXiv:2410.13720*, 2024.

Chris Rockwell, Joseph Tung, Tsung-Yi Lin, Ming-Yu Liu, David F Fouhey, and Chen-Hsuan Lin. Dynamic camera poses and where to find them. In *CVPR*, 2025.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.

Ludan Ruan, Yiyang Ma, Huan Yang, Huiguo He, Bei Liu, Jianlong Fu, Nicholas Jing Yuan, Qin Jin, and Baining Guo. Mm-diffusion: Learning multi-modal diffusion models for joint audio and video generation. In *CVPR*, 2023.

Mengyi Shan, Lu Dong, Yutao Han, Yuan Yao, Tao Liu, Ifeoma Nwogu, Guo-Jun Qi, and Mitch Hill. Towards open domain text-driven synthesis of multi-person motions. In *ECCV*, 2024.

Yu Sun, Qian Bao, Wu Liu, Tao Mei, and Michael J. Black. Trace: 5d temporal regression of avatars with dynamic cameras in 3d environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8856–8866, June 2023.

Alexander Swerdlow, Mihir Prabhudesai, Siddharth Gandhi, Deepak Pathak, and Katerina Fragkiadaki. Unified multimodal discrete diffusion. *arXiv preprint arXiv:2503.20853*, 2025.

Zineng Tang, Ziyi Yang, Chenguang Zhu, Michael Zeng, and Mohit Bansal. Any-to-any generation via composable diffusion. *NeurIPS*, 2023.

Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H Bermano. Human motion diffusion model. In *ICLR*, 2023.

Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025a.

Xi Wang, Robin Courant, Jinglei Shi, Eric Marchand, and Marc Christie. JAWS: Just A Wild Shot for cinematic transfer in neural radiance fields. In *CVPR*, 2023.

Xi Wang, Robin Courant, Marc Christie, and Vicky Kalogeiton. Akira: Augmentation kit on rays for optical video generation. In *CVPR*, 2025b.

Yufu Wang, Ziyun Wang, Lingjie Liu, and Kostas Daniilidis. Tram: Global trajectory and motion of 3d humans from in-the-wild videos. In *ECCV*, 2024a.

Zan Wang, Yixin Chen, Baoxiong Jia, Puhao Li, Jinlu Zhang, Jingze Zhang, Tengyu Liu, Yixin Zhu, Wei Liang, and Siyuan Huang. Move as you say interact as you can: Language-guided human motion generation with scene affordance. In *CVPR*, 2024b.

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 *SIGGRAPH*, 2024c.

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 *CVPR*, pp. 7892–7901, 2024d.

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, 2024e.

Lixing Xiao, Shunlin Lu, Huaijin Pi, Ke Fan, Liang Pan, Yueer Zhou, Ziyong Feng, Xiaowei Zhou, Sida Peng, and Jingbo Wang. Motionstreamer: Streaming motion generation via diffusion-based autoregressive model in causal latent space. *arXiv preprint arXiv:2503.15451*, 2025.

Desai Xie, Ping Hu, Xin Sun, Soren Pirk, Jianming Zhang, Radomir Mech, and Arie E. Kaufman. GAIT: Generating aesthetic indoor tours with deep reinforcement learning. In *ICCV*, 2023.

Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. In *ICLR*, 2025.

Yazhou Xing, Yingqing He, Zeyue Tian, Xintao Wang, and Qifeng Chen. Seeing and hearing: Open-domain visual-audio generation with diffusion latent aligners. In *CVPR*, 2024.

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

Sirui Xu, Zhengyuan Li, Yu-Xiong Wang, and Liang-Yan Gui. Interdiff: Generating 3d human-object interactions with physics-informed diffusion. In *ICCV*, 2023a.

Xingqian Xu, Zhangyang Wang, Gong Zhang, Kai Wang, and Humphrey Shi. Versatile diffusion: Text, images and variations all in one diffusion model. In *ICCV*, 2023b.

Vickie Ye, Georgios Pavlakos, Jitendra Malik, and Angjoo Kanazawa. Decoupling human and camera motion from videos in the wild. In *CVPR*, 2023.

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 *CVPR*, 2023a.

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`.

Mingyuan Zhang, Xinying Guo, Liang Pan, Zhongang Cai, Fangzhou Hong, Huirong Li, Lei Yang, and Ziwei Liu. Remodiffuse: Retrieval-augmented motion diffusion model. In *CVPR*, 2023b.

Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *IEEE TPAMI*, 2024.

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.

Yuyang Zhao, Chung-Ching Lin, Kevin Lin, Zhiwen Yan, Linjie Li, Zhengyuan Yang, Jianfeng Wang, Gim Hee Lee, and Lijuan Wang. Genxd: Generating any 3d and 4d scenes. *ICLR*.

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.

Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *ACM TOG*, 2018.

Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *CVPR*, 2019.

A Use of Large Language Models

We used large language models solely for text polishing and grammar correction during manuscript preparation. No LLMs were involved in the conception or design of the method, experiments, or analysis. All technical content, results, and conclusions have been independently and carefully verified and validated by the authors.

B Detailed related work

Detailed Human motion generation Related Work. Inspired by the success of denoising diffusion models in image generation (Ho et al., 2020; Rombach et al., 2022), several pioneering works (Tevet et al., 2023; Kim et al., 2023; Zhang et al., 2024) adapt diffusion processes to human motion generation. These models are then followed by extensions that leverage pre-trained latent spaces for efficiency, apply consistency distillation for faster sampling, improve caption–motion alignment, and exploit external databases for higher-quality motion (Chen et al., 2023; Dai et al., 2024; Andreou et al.; Zhang et al., 2023b).

While diffusion-based approaches typically represent motion data as raw joint positions and orientations, or continuous latent vectors, another line of work adopts Vector Quantization (VQ) with discrete motion tokens. TM2T (Guo et al., 2022b) first introduce VQ into text-to-motion generation, followed by T2M-GPT (Zhang et al., 2023a), which employed a GPT-style autoregressive model (Brown et al., 2020). More recently, MMM (Pinyoanuntapong et al., 2024), MoMask (Guo et al., 2024), and BiPO (Hong et al., 2024) propose to apply bidirectional attention-based masked generation, inspired by MaskGIT (Chang et al., 2022).

Recently, the masked autoregressive architecture (MAR) (Li et al., 2024) has been proposed to combine the strengths of autoregressive and diffusion models. It has drawn significant attention in the human motion community (Li et al., 2024; Meng et al., 2024; Xiao et al., 2025), as it leverages an autoregressive transformer to handle temporal dynamics while retaining the high generation quality of diffusion models, enabling new state-of-the-art performances.

Nevertheless, most motion generation methods treat motion as an isolated modality, which oversimplifies real-world scenarios where humans continuously interact with their surroundings. Consequently, recent research has begun modeling human interactions with objects (Xu et al., 2023a; Peng et al., 2025b; Geng et al., 2025), other humans (Liang et al., 2024; Fan et al., 2024; Shan et al., 2024), and scenes (Wang et al., 2024b; Cen et al., 2024; Jiang et al., 2024c). However, while recent studies have considered human–camera interaction in motion estimation (Patel & Black, 2025; Ye et al., 2023; Wang et al., 2024a; Kocabas et al., 2024; Sun et al., 2023), motion generation remains largely unexplored, with existing efforts treating camera parameters merely as constraints or conditioning signals rather than modeling their joint distribution with motion.

Detailed Generative Camera Trajectory generation Related Work. Over the past two decades, camera control and generation have progressed from handcrafted, rule-based geometric design (Blinn, 1988; Lino & Christie, 2015; Drucker et al., 1992) to deep learning methods that exploit the descriptive and fitting capacity of neural networks: approaches that either learn cinematic rules from example-based references (Jiang et al., 2020; 2021) or leverage the differentiability of deep models to optimize camera trajectories in real-data-supported 3D environments (Wang et al., 2023; Jiang et al., 2024d; Chen et al., 2024).

However, these example-based methods often rely on carefully curated reference videos, and in some cases even synthetic annotation pairs, either to train discriminative models or to optimize trajectories. To mitigate this dependency, other works explore reinforcement learning (RL). In drone cinematography (Huang et al., 2019; Bonatti et al., 2020), RL is guided by human pose and optical flow, while in indoor environments, Xie et al. (2023) propose to use an aesthetic model as the reward function. Though effective within specific environments, both example-based and RL-based methods often collapse into limited trajectory styles and require environment-specific training, resulting in poor generalization.

Yet, with the rapid progress of image and video generative models (Polyak et al., 2024; Peng et al., 2025a; Wang et al., 2025a), a notable direction is to bypass explicit 3D representations and instead treat the model as a universal renderer. This has enabled direct camera-controlled video generation (Wang et al., 2024c; He et al., 2024; Xu et al., 2024; Bahmani et al., 2024; Zheng et al., 2024; Cheong et al., 2024; Wang et al., 2025b). While showing great potential, this line of work faces several limitations: (1) it overlooks scene semantics (e.g., character performance); (2) it still relies on manually designed, complex camera trajectories, which remain challenging for users; and (3) given the relatively low quality of current video generators, the outputs are hard to use directly in production, while the end-to-end nature of these models prevents artists from accessing intermediate assets (e.g., meshes, trajectories, lighting conditions).

To achieve geometric controllability and provide intermediate assets without requiring expert-designed trajectories, Jiang et al. (2024b) introduced the first diffusion-based approach for camera generation. Although limited to synthetic data, their key idea of deriving camera behavior from semantic prompts opens a new direction. Subsequently, Courant et al. (2024) proposed E.T., a large-scale dataset of realistic camera trajectories with human motion from real films, together with evaluation metrics and novel architectural designs. DanceCamAnimator Wang et al. (2024e) and DanceCamera3D Wang et al. (2024d) also focus on dance-specific camera control conditioned on music. More recently, GenDoP (Zhang et al., 2025a) constructed an object-wise, interaction-centric

dataset and employed an autoregressive model to generate trajectories from textual descriptions and visual inputs.

Similarly to human motion generation, most camera generation works condition on human motion but rely solely on global trajectories, which restricts the interaction between camera and subject and overlooks the intrinsic joint distribution problem. In this work, we aim to bridge this gap by addind human motion into the camera trajectory generation pipeline, modelling the symbiosis between how and what to film.

C Detailed method

C.1 Background

C.1.1 Moore-Penrose Pseudo-Inverse and induced projections

We provide the background defining the Moore-Penrose pseudo-inverse of an $m \times n$ matrix W. Let $k := \min\{n, m\}$.

Consider a singular value decomposition of W, given by $W = UDV^\top$, where:

1. U is an $m \times m$ orthogonal matrix, i.e., $U^\top U = \boldsymbol{I}_m$,
2. V is an $n \times n$ orthogonal matrix, i.e., $V^\top V = \boldsymbol{I}_n$,
3. $D = \mathrm{diag}(d_1, \ldots, d_k)$ is a $k \times k$ diagonal matrix with non-negative, non-increasing diagonal entries.

The Moore-Penrose pseudo-inverse $W^\dagger$ is then given by:

$$W^\dagger = VD^\dagger U^\top,$$

where $D^\dagger$ is the $k \times k$ diagonal matrix with entries $D^\dagger_{i,i} = d_i^{-1}$ if $d_i > 0$, and 0 otherwise.

Note that if the $k \times k$ matrix $W^\top W$ is invertible, then $D^\dagger = D^{-1}$ and hence $W^\dagger = W^\top (W^\top W)^{-1}$.

Define $P := D^\dagger D$, a diagonal matrix with entries:

$$P_{i,i} = \begin{cases} 1 & \text{if } d_i \neq 0, \\ 0 & \text{otherwise.} \end{cases}$$

We end up with the following properties:

1. *Projection:* $W^\dagger W = VPV^\top$, and $(W^\dagger W)^2 = W^\dagger W$.
2. *Symmetry:* $(W^\dagger W)^\top = W^\dagger W$.
3. *Projection image:* $\ker W = \mathrm{im}(\boldsymbol{I}_n - W^\dagger W)$.

Thus, $P_\perp := \boldsymbol{I}_n - W^\dagger W$ is the orthogonal projection onto $\ker W$, and $P_{/\!/} := W^\dagger W$ is the orthogonal projection onto the orthogonal of $\ker W$: the coimage of W.

C.1.2 Statistics

We start by a well-known special case of the main theorem from Cochran (1934), that we apply to projection matrices.

Theorem C.1 (Cochran)**.** *Let $X \sim \mathcal{N}(\mu, \sigma^2 \boldsymbol{I}_n)$ be an isotropic Gaussian random vector, and let $F \subseteq \mathbb{R}^n$ be a linear subspace. If P_F and $P_{F^\perp}$ are the respective orthogonal projections onto F and its orthogonal complement $F^\perp$, then:*

1. *$P_F X \sim \mathcal{N}(P_F \mu, \sigma^2 P_F)$ and $P_{F^\perp} X \sim \mathcal{N}(P_{F^\perp} \mu, \sigma^2 P_{F^\perp})$ are (possibly degenerate) Gaussian random vectors.*
2. *$P_F X$ and $P_{F^\perp} X$ are independent.*

For completeness, we provide a succinct proof below.

Proof. Let $P = \begin{bmatrix} P_F \\ P_{F^\perp} \end{bmatrix}$. By the properties of Gaussian vectors, PX is a Gaussian vector with covariance matrix:

$$\Sigma = \sigma^2 \begin{bmatrix} P_F \circ P_F^\top & P_F \circ P_{F^\top}^\top \\ P_{F^\top} \circ P_F^\top & P_{F^\top} \circ P_{F^\top}^\top \end{bmatrix} = \sigma^2 \begin{bmatrix} P_F & 0 \\ 0 & P_{F^\top} \end{bmatrix}, \tag{9}$$

where the last inequality follows from the properties of the orthogonal projections P_F and $P_{F^\perp}$, namely: $P_F = P_F^\top$, $P_F^2 = P_F$, and $P_F P_{F^\perp} = 0$. Similarly, since the multiplication by P is linear, $\mathbb{E}(PX) = \begin{bmatrix} P_F \mu \\ P_{F^\perp} \mu \end{bmatrix}$. Item 1 follows from the block decomposition, and Item 2 follows from Equation (9) and the fact that uncorrelated Gaussian vectors are independent. □

C.2 METHOD

Let $\boldsymbol{W}^\dagger$ denote the Moore-Penrose pseudo-inverse of $\boldsymbol{W}$ (see Section C.1.1). Using this notation, the matrix $\boldsymbol{P}_\perp := \boldsymbol{I} - \boldsymbol{W}^\dagger \boldsymbol{W}$ (resp. $\boldsymbol{P}_{/\!/} := \boldsymbol{W}^\dagger \boldsymbol{W}$) can be then identified as the orthogonal projection onto $\ker \boldsymbol{W}$ (resp. $\ker (\boldsymbol{W})^\perp$). As $\mathbf{u} = [\mathbf{x}, \mathbf{y}]^\top \sim \mathcal{N}(\boldsymbol{\mu}, \sigma^2 \boldsymbol{I}_n)$, Cochran's theorem (Theorem C.1) then guarantees that the corresponding projections $\mathbf{u}$:

$$\mathbf{u}_\perp := \boldsymbol{P}_\perp \mathbf{u} \sim \mathcal{N}(\boldsymbol{P}_\perp \boldsymbol{\mu}, \sigma^2 \boldsymbol{P}_\perp) \quad \text{and} \quad \mathbf{u}_{/\!/} := \boldsymbol{P}_{/\!/} \mathbf{u} \sim \mathcal{N}(\boldsymbol{P}_{/\!/} \boldsymbol{\mu}, \sigma^2 \boldsymbol{P}_{/\!/}), \tag{10}$$

are independent (possibly degenerate) Gaussian vectors.

Observe that $\mathbf{z} = \boldsymbol{W}\mathbf{u} = \boldsymbol{W}\mathbf{u}_{/\!/}$, so $\mathbf{z}$ is $\mathbf{u}_{/\!/}$-measurable and thus independent of $\mathbf{u}_\perp$. Moreover, $\mathbf{u}_{/\!/} = \boldsymbol{W}^\dagger \boldsymbol{W} \mathbf{u} = \boldsymbol{W}^\dagger \mathbf{z}$, which shows that $\mathbf{u}_{/\!/}$ is a measurable function of $\mathbf{z}$. Therefore,

$$\mathbb{E}[\mathbf{u} \mid \mathbf{z}] = \mathbb{E}\left[\mathbf{u}_\perp + \boldsymbol{W}^\dagger \mathbf{z} \mid \mathbf{z}\right] \overset{\mathbf{u}_\perp \text{ indep. } \mathbf{z}}{=} \mathbb{E}(\mathbf{u}_\perp) + \mathbb{E}[\boldsymbol{W}^\dagger \mathbf{z} \mid \mathbf{z}] \overset{\mathbf{z}\text{-meas.}}{=} \boldsymbol{P}_\perp \boldsymbol{\mu} + \boldsymbol{W}^\dagger \mathbf{z}. \tag{11}$$

We have the following decomposition of $\mathbf{u}$ into two independent variables: $\mathbf{u} = \boldsymbol{P}_\perp \mathbf{u} + \boldsymbol{P}_{/\!/} \mathbf{u} = \boldsymbol{P}_\perp \mathbf{u} + \boldsymbol{W}^\dagger \mathbf{z}$. This induces a decomposition of the density $p_{\mathbf{u}}$ of $\mathbf{u}$ into two density functions[2] $p_{\mathbf{u}_\perp}$ and $p_{\mathbf{u}_\parallel} = p_{\boldsymbol{W}^\dagger \mathbf{z}}$ of $\mathbf{u}_\perp$ and $\mathbf{u}_{/\!/} = \boldsymbol{W}^\dagger \mathbf{z}$ respectively. Given point $u \in \mathbb{R}^n$:

$$p_{\mathbf{u}}(u) = p_{(\mathbf{u}_\perp, \mathbf{u}_\parallel)}(u_\perp, u_{/\!/}) \overset{\text{indep.}}{=} p_{\mathbf{u}_\perp}(u_\perp) p_{\mathbf{u}_{/\!/}}(u_{/\!/}), \tag{12}$$

where the functions $p_{\mathbf{u}_\perp}$ and $p_{\mathbf{u}_{/\!/}}$ are given by:

$$v \mapsto p_{\mathbf{u}_\perp}(v) = \frac{\mathbb{1}_{\ker \boldsymbol{W}}(v)}{\sqrt{2\pi\sigma^2}^{\dim \ker \boldsymbol{W}}} \exp\left[-\frac{1}{2\sigma^2}(v - \boldsymbol{\mu})^\top (v - \boldsymbol{\mu})\right], \text{ and}$$
$$v \mapsto p_{\mathbf{u}_{/\!/}}(v) = \frac{\mathbb{1}_{\ker (\boldsymbol{W})^\perp}(v)}{\sqrt{2\pi\sigma^2}^{\dim \ker (\boldsymbol{W})^\perp}} \exp\left[-\frac{1}{2\sigma^2}(v - \boldsymbol{\mu})^\top (v - \boldsymbol{\mu})\right].$$

C.3 MORE DETAILS ON THE DENSITY DECOMPOSITION

We give here more details on the possible decompositions of the density $p_{\mathbf{u}}$ of $\mathbf{u}$. The decomposition given in Equation (12) from ensures that we have for a given point $u \in \mathbb{R}^n$:

$$p_{\mathbf{u}}(u) = p_{\mathbf{u}_\perp}((\boldsymbol{I} - \boldsymbol{W}^\dagger \boldsymbol{W})u) p_{\mathbf{u}_{/\!/}}(\boldsymbol{W}^\dagger \boldsymbol{W} u). \tag{13}$$

Following the argumentation of Equation (11), given a point $z \in \mathbb{R}^m$, the conditional distribution $\mathbf{u} \mid \mathbf{z} = z$ of $\mathbf{u}$ conditionally to the event $\{\mathbf{z} = z\}$ is given by $\mathbf{u}_\perp + \boldsymbol{W}^\dagger z$: a translation of $\mathbf{u}_\perp$ by the constant $\boldsymbol{W}^\dagger z$ and thus admits a density $p_{\mathbf{u}|\mathbf{z}=z}$, given by:

$$(u, z) \longmapsto p_{\mathbf{u}|\mathbf{z}=z}(u) = p_{\mathbf{u}_\perp + \boldsymbol{W}^\dagger z}(u) = p_{\mathbf{u}_\perp}(u - \boldsymbol{W}^\dagger z).$$

[2] Note that $\mathbf{u}_\perp$ and $\mathbf{u}_{/\!/}$ do not admit densities w.r.t. the Lebesgue measure on $\mathbb{R}^n$, since they are supported on the non-full-dimensional vector spaces $\ker \boldsymbol{W}$ and $\ker (\boldsymbol{W})^\perp$ respectively. Nevertheless, they admit densities $p_{\mathbf{u}_\perp}$ and $p_{\mathbf{u}_{/\!/}}$ w.r.t. the respective Lebesgue measures on these subspaces. See Section C.3 for more details.

This can be directly shown by the change of variable $u = u_\perp + \boldsymbol{W}^\dagger z = u_\perp + \boldsymbol{W}^\dagger \boldsymbol{W} u_{/\!/}$ in Equation (13), since on one hand, we have $\mathbf{z} = \boldsymbol{W}\mathbf{u}$ almost surely, hence

$$p_{\mathbf{u}}(u) = p_{(\mathbf{u}_\perp, \mathbf{u}_\parallel)}(u_\perp, u_{/\!/}) \overset{\text{indep.}}{=} p_{\mathbf{u}_\perp}(u_\perp) p_{\mathbf{u}_{/\!/}}(u_{/\!/}) \overset{\text{shift}}{=} p_{\mathbf{u}_\perp + u_{/\!/}}(u_\perp + u_{/\!/}) p_{\mathbf{u}_{/\!/}}(u_{/\!/}).$$

Furthermore, on the other hand, we have $\mathbf{u}_{/\!/} = \boldsymbol{W}^\dagger \mathbf{z}$ and $\mathbf{z} = \boldsymbol{W}\mathbf{u}$, hence $p_{\mathbf{z}}(\cdot) \propto p_{\mathbf{u}_{/\!/}}(\boldsymbol{W}\cdot)$, and therefore, for any point u and $z = \boldsymbol{W}u$:

$$p_{\mathbf{u}}(u) \overset{z=\boldsymbol{W}u}{=} p_{\mathbf{u}_\perp + \boldsymbol{W}^\dagger z}(u) p_{\mathbf{u}_\parallel}(\boldsymbol{W}^\dagger z) \overset{z=\boldsymbol{W}u}{\propto} p_{\mathbf{u}_\perp + \boldsymbol{W}^\dagger z}(u) p_{\mathbf{z}}(z).$$

This equality being true for (almost) every u and $z = \boldsymbol{W}u$, we conclude that $p_{\mathbf{u}|\mathbf{z}=z} \propto p_{\mathbf{u}_\perp + \boldsymbol{W}^\dagger z}$ almost everywhere. This property can be shown more directly by invoking the fact that the σ-algebra generated by $\mathbf{z}$ and the one of $\mathbf{u}_{/\!/}$ are the same.

C.4 Auxiliary sampling derivation

In this section, we detail the derivation from Equation (14) to Equation (8). Starting from Equation (4) and applying the decomposition in Equation (6), we have:

$$\begin{aligned}\nabla_{\mathbf{x}_t, \mathbf{y}_t} \log \tilde{p}(\mathbf{x}_t, \mathbf{y}_t | \mathrm{c}) = \nabla_{\mathbf{x}_t, \mathbf{y}_t} \log p(\mathbf{u}_\perp) + (1 + w_z) \nabla_{\mathbf{x}_t, \mathbf{y}_t} \log p(\mathbf{u}_{/\!/}) \\ + w_c (\nabla_{\mathbf{x}_t, \mathbf{y}_t} \log p(\mathbf{x}_t, \mathbf{y}_t | \mathrm{c}) - \nabla_{\mathbf{x}_t, \mathbf{y}_t} \log p(\mathbf{x}_t, \mathbf{y}_t)).\end{aligned} \tag{14}$$

Therefore, since $[\mathbf{x}, \mathbf{y}]^\top \sim \mathcal{N}(\boldsymbol{\mu}, \sigma^2 \boldsymbol{I})$ we obtain:

$$\begin{aligned}\nabla_{\mathbf{x}_t, \mathbf{y}_t} \log p(\mathbf{x}_t, \mathbf{y}_t | \mathrm{c}) &= -\frac{1}{\sigma^2}(\boldsymbol{I} - \boldsymbol{P}_{/\!/}) \left([\mathbf{x}_t, \mathbf{y}_t]^\top - \boldsymbol{\mu} \right) - \frac{1}{\sigma^2}(1 + w_z) \boldsymbol{P}_{/\!/} \left([\mathbf{x}_t, \mathbf{y}_t]^\top - \boldsymbol{\mu} \right) \\ &\quad - \frac{1}{\sigma^2} w_c \left((\mathbf{x}_c - \boldsymbol{\mu}_c) - ([\mathbf{x}_t, \mathbf{y}_t]^\top - \boldsymbol{\mu}) \right) \\ &= -\frac{1}{\sigma^2} \left([\mathbf{x}_t, \mathbf{y}_t]^\top - \boldsymbol{\mu} \right) \\ &\quad - \frac{1}{\sigma^2} w_z \boldsymbol{P}_{/\!/} \left([\mathbf{x}_t, \mathbf{y}_t]^\top - \boldsymbol{\mu} \right) \\ &\quad - \frac{1}{\sigma^2} w_c \left(([\mathbf{x}_t^c, \mathbf{y}_t^c]^\top - \boldsymbol{\mu}_c) - ([\mathbf{x}_t, \mathbf{y}_t]^\top - \boldsymbol{\mu}) \right).\end{aligned} \tag{15}$$

Finally, recalling that $\boldsymbol{\varepsilon} = -\frac{1}{\sigma}(\mathbf{x} - \boldsymbol{\mu})$, we can express the sampling equation with noise prediction as:

$$\begin{aligned}\boldsymbol{\varepsilon}(\mathbf{x}_t, \mathbf{y}_t, \mathrm{c}, t) = \boldsymbol{\varepsilon}(\mathbf{x}_t, \mathbf{y}_t, \emptyset, t) \\ + w_z \boldsymbol{P}_{/\!/} \, \boldsymbol{\varepsilon}(\mathbf{x}_t, \mathbf{y}_t, \emptyset, t) \\ + w_c (\boldsymbol{\varepsilon}(\mathbf{x}_t, \mathbf{y}_t, \mathrm{c}, t) - \boldsymbol{\varepsilon}(\mathbf{x}_t, \mathbf{y}_t, \emptyset, t)).\end{aligned} \tag{16}$$

D Detailed dataset

D.1 Detailed pipeline

In this section, we describe the construction of the PulpMotion dataset, illustrated in Figure 10. We first apply an off-the-shelf camera-human pose estimator (Wang et al., 2024a) to infer both camera and human poses from video clips of the CondensedMovies dataset (Bain et al., 2020). As noted in the main manuscript, a key challenge of video-based pose estimation is handling occluded or unseen body parts, which are often inaccurately predicted.

To address this, we first identify poorly estimated regions by reprojecting visible joints. We then use a vision-language model (VLM) to generate captions describing human motion, providing bounding boxes around the target person to guide the model's focus. We show an example of human motion caption generation in Figure 26 with input prompt and VLM response.

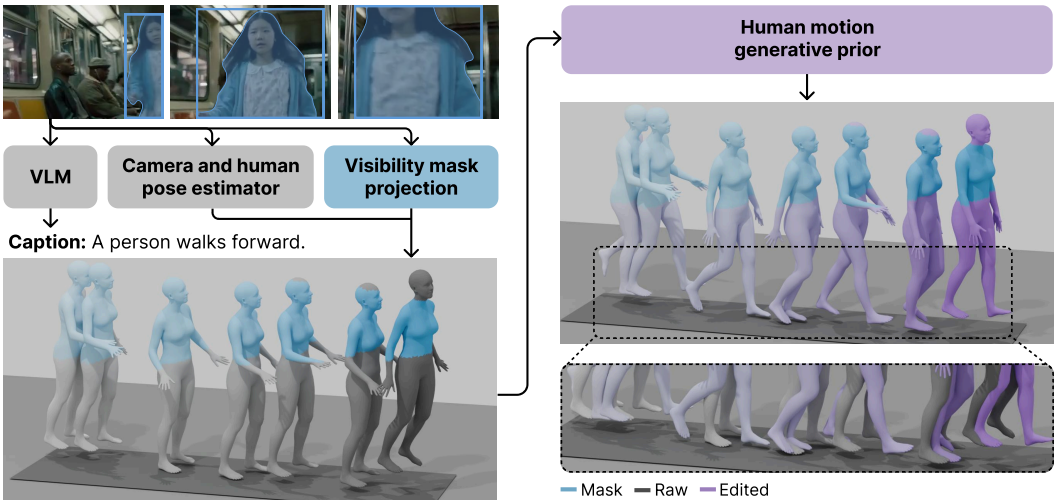


Figure 10: **Dataset refinement pipeline.** Given RGB frames from a video, we first estimate the camera and human pose. We then identify the out-of-screen body parts by reprojection. Finally, we refine the out-of-screen parts using a generative prior.

Table 6: **Comparison of PulpMotion and E.T. (Courant et al., 2024) datasets.** We compare the full (*all*), *pure*, and *mixed* subsets of PulpMotion with the E.T.. We summarize modality coverage, available captions, dataset size (hours, frames, samples), sample length statistics (median, mean, std), and vocabulary size.

Dataset	Camera		Human		#Hours	#Frames	#Samples	Sample lengths (frames)			#Vocabulary
	Traj	Caption	Motion	Caption				Median	Mean	Std	
E.T. Courant et al. (2024)											
all	✓	✓	✓	✗	120	11M	115K	75	93.9	73.8	1,790
pure					20	1.8M	30K	46	59.5	49.1	941
mixed					67	6M	65K	72	92.9	75.06	1,579
PulpMotion (Ours)											
all	✓	✓	✓	✓	314	22M	193K	107	117.3	63.6	4,599
pure					51	3.7M	41K	70	91.1	59.2	2,831
mixed					170	12M	105K	108	116.44	60.44	4,143

Next, in the right part of Figure 10, we refine the occluded regions using a diffusion-based editing method (Lugmayr et al., 2022) with a model pretrained on HumanML3D (Guo et al., 2022a). To avoid artifacts caused by naive editing, we refine the entire sub-kinematic chain of each occluded joint rather than modifying joints in isolation. Since visible parts remain largely unchanged, projection consistency between the reconstructed body and RGB frames is preserved.

D.2 Detailed statistics

Table 6 compares our PulpMotion dataset with E.T. Courant et al. (2024) across several dimensions.

Overall, PulpMotion significantly increases the dataset size, containing 314 hours and 22M frames compared to E.T.'s 120 hours and 11M frames. Our dataset also provides longer samples (median 107 frames vs. 75). The "pure" and "mixed" subsets follow the same trends, demonstrating consistent improvements.

Thanks to our refinement pipeline, as shown in the main manuscript, PulpMotion ensures higher-quality human motions. Additionally, PulpMotion includes HumanML3D-style human motion captions, which are not available in E.T.

E Detailed experiments

E.1 Auto-encoder

E.1.1 Detailed experimental setup

Implementation details. We adopt the ResNet-based autoencoder from MARDM (Meng et al., 2024) with ReLU activations. The joint encoder and three modality-specific decoders have temporal down-/up-sampling by a factor of 4, and each consist of two 1D-ResNet blocks. Latent dimensions are set to 64 for the camera, 128 for the human, and 64 for the projection. The model is trained for 325 epochs on the full Pulp Motion dataset with 64-frame samples (and evaluated on 300-frame samples), using AdamW with a learning rate of 1.9×10^{-4}, a batch size of 128, on a single A100 GPU. A linear warmup of 1K steps is applied, followed by decay by 0.1 after 4K steps.

E.1.2 Detailed performances

Table 7: **Reconstruction evaluation of autoencoder.** We report reconstruction metrics for *pure* and *mixed* subsets. Metrics span projection accuracy (MPJProjE, $FD_{framing}$), human pose quality (MPJPE, FDTMR, TMR-Score), and camera alignment (APE, FD_{CLaTr}, CLaTr-Score).

Methods	Framing		Human			Camera		
	MPJProjPE ↓	$FD_{framing}$ ↓	MPJPE ↓	FD_{TMR} ↓	TMR-Score ↑	APE ↓	FD_{CLaTr} ↓	CLaTr-Score ↑
pure								
Ground truth	0.00	0.00	0.00	0.00	16.47	0.00	0.00	70.25
AE	0.09	0.14	3.26	105.57	15.93	0.15	19.26	60.45
mixed								
Ground truth	0.00	0.00	0.00	0.00	17.72	0.00	0.00	68.88
AE	0.08	0.23	5.63	124.78	18.16	0.18	15.64	57.98

We evaluate the reconstruction quality of the autoencoder introduced in Section 3.1 using modality-specific errors: mean per-joint projected error (MPJProjPE) and mean per-joint error (MPJPE) for human motion, and absolute pose error (APE) for camera. Additionally, we compute reconstruction Fréchet distances and modality–text alignment metrics from Section 5.1.

As shown in Table 7, the autoencoder achieves low MPJProjPE (0.08–0.09), indicating reliable 2D frame reconstruction across both subsets. MPJPE reveals discrepancies in 3D pose recovery, particularly in the mixed setting (5.63 vs. 3.26). APE remains low ($\leq$ 0.18) but shows slight degradation in the mixed case, consistent with the observed drop in CLaTr-Score.

E.2 Generation

E.2.1 Detailed experimental setup

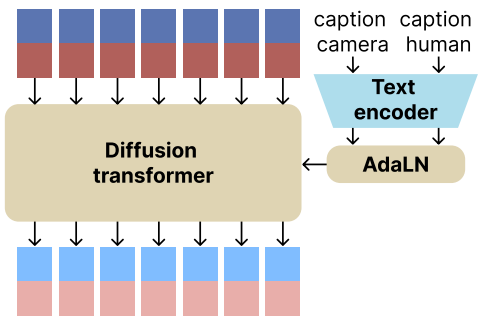


Figure 11: **Overview of the DiT architecture.**

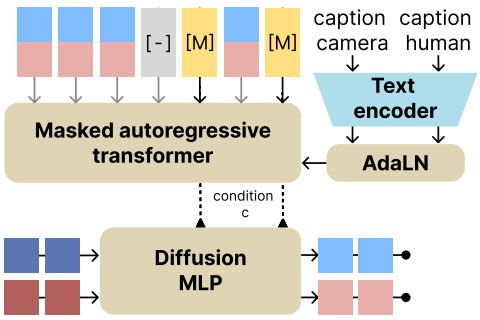
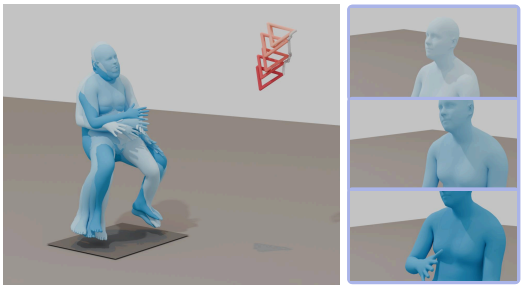


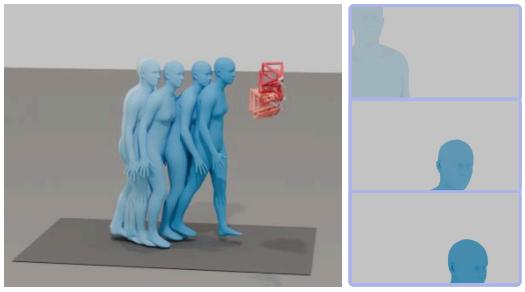
Figure 12: **Overview of the MAR architecture.**

We illustrate in Figures 12 and 11 both DiT Peebles & Xie (2023) and MAR Li et al. (2024) architectures used in this work.



Human: A person sitting.
Camera: The camera performs a boom bottom.

Figure 13: **Example with DiT on the mixed set.**



Human: A person walks while turning head to right.
Camera: The camera booms up.

Figure 14: **Example with MAR on the mixed set.**

Implementation details. We evaluate two architectures: a DiT-based model with in-context conditioning Courant et al. (2024) and a MAR-based model with AdaLN conditioning Meng et al. (2024). To ensure fairness, both are scaled to $\sim$ 28.3M parameters. The DiT model has 8 layers with a hidden dimension 532 and 14 attention heads. The MAR model uses a single-layer autoregressive transformer (hidden dimension 512, 8 heads) and a diffusion head with 3 MLP layers of width 1024. Both models are trained for 93k steps on the *pure* subset and 330K steps on the *mixed* subset on the pure and mixed subsets of Pulp Motion with 300-frame samples, using AdamW with a learning rate of 3×10^{-4}, a batch size of 128, on a single A100 GPU. A linear warmup of 2K steps is applied, followed by decay by 0.1 after 50K steps. For inference, we perform 50 DDPM sampling steps.

E.2.2 More qualitative results on mixed dataset

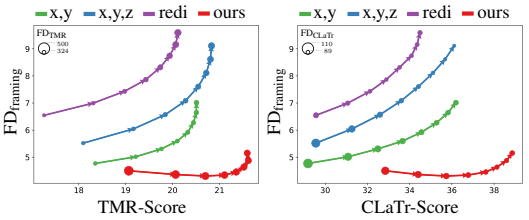
We show in Figure 13 and Figure 14 additional qualitative results on the *mixed* subset. We also provide an ***anonymous online gallery*** with additional qualitative videos, including generated samples, comparisons to baselines, dataset refinement before/after, and representative dataset samples.

The gallery includes:

- **Generation examples**, comparing our method to baselines on both DiT and MAR architectures. We observe that our approach achieves more stable and consistent framing, reliably keeping the human on screen, whereas baselines either ignore the subject entirely or fail to maintain framing over the full sequence.
- **Dataset pipeline**, illustrating the extraction of camera, human, and textual information as well as the refinement step, which noticeably improves motion naturalness (e.g., converting sliding artifacts into realistic walking).
- **Dataset examples**, highlighting the diversity of camera trajectories, human motions, and textual captions. The link is also included in the supplementary material (Section E.2.2).

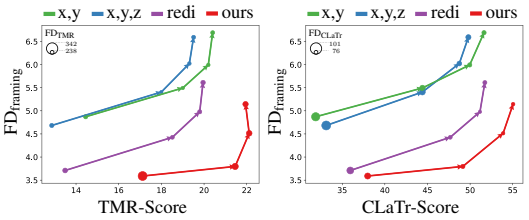
E.2.3 Comparison to the state of the art on pure dataset

Quantitative results. Table 8 reports a comparison of our auxiliary sampling (Aux) method against state-of-the-art baselines across both DiT and MAR architectures on the pure subset. We summarize our experimental observations as follows:
(i) Auxiliary sampling improves coherence. Applying Aux consistently lowers framing error and out rates: for DiT, $\text{FD}_{\text{framing}}$ drops from $10.24 \rightarrow 7.88$ $(\mathbf{x})(\mathbf{y})$ and $6.78 \rightarrow 5.03$ $(\mathbf{x}, \mathbf{y})$; for MAR, it decreases from $11.22 \rightarrow 9.32$ $(\mathbf{x})(\mathbf{y})$ and $6.55 \rightarrow 4.90$ $(\mathbf{x}, \mathbf{y})$. Out rates similarly improve, reaching the best values among baselines (DiT 24.92%, MAR 24.28%). These results show that Aux enhances multimodal coherence and framing even when using independent modality or dual-modality settings.
(ii) Strong per-modality performance. Relative to ReDi, Aux improves text–modality alignment: TMR-Score increases from $23.72 \rightarrow 24.67$ (DiT, $(\mathbf{x})(\mathbf{y})$) and $21.66 \rightarrow 23.25$ (MAR, $(\mathbf{x})(\mathbf{y})$), while CLaTr-Score rises from $57.74 \rightarrow 62.75$ (DiT) and $57.18 \rightarrow 60.74$ (MAR). Camera fidelity also improves (DiT FD_{CLaTr} : $86.06 \rightarrow 77.78$; MAR $120.50 \rightarrow 108.43$) with only minor trade-offs in



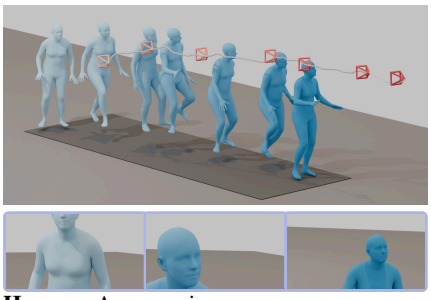
(a) DiT TMR (b) DiT CLaTr

Figure 15: **State-of-the-art comparison in DiT on the pure subset.** Trade-off between framing quality and modality-text alignment for textual guidance ranges from 5 to 12. The optimal region is at the bottom-right (low framing error, high alignment).



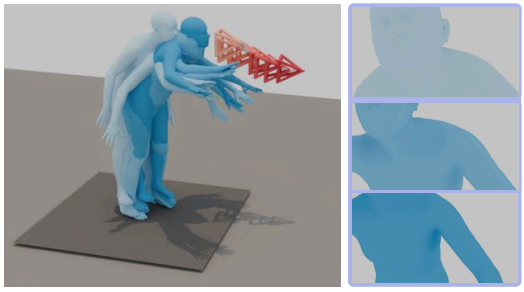
(a) MAR TMR (b) MAR CLaTr

Figure 16: **State-of-the-art comparison in MAR on the pure subset.** Trade-off between framing quality and modality-text alignment for textual guidance values ranges from 2 to 5. The optimal region is at the bottom-right (low framing error, high alignment).



Human: A person jumps.
Camera: The camera performs a pull out.

Figure 17: **Example with DiT on the pure subset.**



Human: A person leans forward.
Camera: The camera performs a pull out.

Figure 18: **Example with MAR on the pure subset.**

human fidelity. Overall, Aux enhances framing and multimodal coherence while maintaining strong per-modality alignment on the pure subset.

Table 8: **State-of-the-art comparison on the pure subset.** We compare four baselines: independent modality generation $(\mathbf{x})(\mathbf{y})$, dual-modality generation $(\mathbf{x}, \mathbf{y})$, triplet-modality generation $(\mathbf{x}, \mathbf{y}, \mathbf{z})$, and ReDi Kouzelis et al. (2025), along with our auxiliary sampling (Aux) method. Results are reported for DiT (Peebles & Xie, 2023) and MAR (Li et al., 2024). Superscript $\pm$ denotes the 95% confidence interval over 10 samplings.

Methods	Framing		Human				Camera			
	$FD_{framing}$ ↓	Out-rate ↓	FD_{TMR} ↓	TMR-Score ↑	R3 ↑	Coverage ↑	FD_{CLaTr} ↓	CLaTr-Score ↑	F1 ↑	Coverage ↑
Ground-truth	0.00	0.71	0.00	16.47	19.79	1.00	0.00	70.25	94.52	1.0
Auto-encoder	0.14	3.46	105.57	15.93	20.21	89.00	19.26	60.45	77.51	78.96
DiT										
$(\mathbf{x})(\mathbf{y})$	$10.24^{\pm 0.08}$	$41.70^{\pm 0.40}$	$384.31^{\pm 0.62}$	$23.72^{\pm 0.10}$	$20.63^{\pm 0.35}$	$10.46^{\pm 0.18}$	$86.06^{\pm 0.38}$	$57.74^{\pm 0.29}$	$75.53^{\pm 0.23}$	$31.83^{\pm 0.21}$
$(\mathbf{x})(\mathbf{y})$+Aux (ours)	$7.88^{\pm 0.07}$	$36.03^{\pm 0.36}$	$443.65^{\pm 0.89}$	$24.67^{\pm 0.09}$	$21.41^{\pm 0.35}$	$8.63^{\pm 0.19}$	$77.78^{\pm 0.57}$	$62.75^{\pm 0.31}$	$83.00^{\pm 0.35}$	$29.53^{\pm 0.28}$
$(\mathbf{x}, \mathbf{y})$	$6.78^{\pm 0.06}$	$36.25^{\pm 0.36}$	$372.75^{\pm 0.94}$	$20.74^{\pm 0.10}$	$18.16^{\pm 0.25}$	$12.73^{\pm 0.31}$	$93.37^{\pm 0.78}$	$35.99^{\pm 0.20}$	$48.82^{\pm 0.39}$	$44.56^{\pm 0.33}$
$(\mathbf{x}, \mathbf{y}, \mathbf{z})$	$5.56^{\pm 0.06}$	$29.81^{\pm 0.27}$	$334.29^{\pm 1.10}$	$18.04^{\pm 0.19}$	$15.52^{\pm 0.22}$	$17.46^{\pm 0.25}$	$108.05^{\pm 1.21}$	$28.62^{\pm 0.35}$	$41.91^{\pm 0.41}$	$45.83^{\pm 0.37}$
ReDi	$6.53^{\pm 0.05}$	$33.34^{\pm 0.27}$	$323.53^{\pm 0.74}$	$17.13^{\pm 0.19}$	$15.21^{\pm 0.30}$	$17.81^{\pm 0.19}$	$99.60^{\pm 0.92}$	$28.65^{\pm 0.36}$	$40.49^{\pm 0.41}$	$48.60^{\pm 0.33}$
$(\mathbf{x}, \mathbf{y})$+Aux (ours)	$5.03^{\pm 0.03}$	$24.92^{\pm 0.28}$	$424.81^{\pm 1.07}$	$21.80^{\pm 0.12}$	$18.32^{\pm 0.15}$	$11.69^{\pm 0.19}$	$91.36^{\pm 0.81}$	$38.42^{\pm 0.31}$	$51.61^{\pm 0.47}$	$40.94^{\pm 0.19}$
MAR										
$(\mathbf{x})(\mathbf{y})$	$11.22^{\pm 0.04}$	$45.39^{\pm 0.48}$	$261.20^{\pm 0.78}$	$21.66^{\pm 0.10}$	$27.59^{\pm 0.40}$	$18.89^{\pm 0.23}$	$120.50^{\pm 0.81}$	$57.18^{\pm 0.23}$	$68.09^{\pm 0.24}$	$38.97^{\pm 0.54}$
$(\mathbf{x})(\mathbf{y})$+Aux (ours)	$9.32^{\pm 0.07}$	$41.54^{\pm 0.36}$	$280.36^{\pm 0.84}$	$23.25^{\pm 0.06}$	$28.56^{\pm 0.27}$	$15.79^{\pm 0.17}$	$108.43^{\pm 0.66}$	$60.74^{\pm 0.12}$	$71.08^{\pm 0.48}$	$34.60^{\pm 0.32}$
$(\mathbf{x}, \mathbf{y})$	$6.55^{\pm 0.10}$	$30.19^{\pm 0.34}$	$251.94^{\pm 1.46}$	$20.16^{\pm 0.13}$	$25.48^{\pm 0.29}$	$28.25^{\pm 0.43}$	$108.28^{\pm 1.83}$	$52.17^{\pm 0.32}$	$67.31^{\pm 0.49}$	$55.48^{\pm 0.52}$
$(\mathbf{x}, \mathbf{y}, \mathbf{z})$	$6.10^{\pm 0.11}$	$30.11^{\pm 0.32}$	$242.81^{\pm 0.91}$	$19.23^{\pm 0.10}$	$25.17^{\pm 0.46}$	$30.33^{\pm 0.48}$	$116.75^{\pm 1.04}$	$49.52^{\pm 0.23}$	$63.14^{\pm 0.37}$	$55.81^{\pm 0.50}$
ReDi	$5.07^{\pm 0.10}$	$25.84^{\pm 0.31}$	$252.58^{\pm 0.92}$	$19.73^{\pm 0.11}$	$25.50^{\pm 0.41}$	$28.34^{\pm 0.37}$	$103.13^{\pm 1.14}$	$51.99^{\pm 0.27}$	$66.95^{\pm 0.38}$	$56.29^{\pm 0.59}$
$(\mathbf{x}, \mathbf{y})$+Aux (ours)	$4.90^{\pm 0.06}$	$24.28^{\pm 0.31}$	$281.39^{\pm 0.83}$	$21.90^{\pm 0.17}$	$26.43^{\pm 0.40}$	$17.48^{\pm 0.26}$	$100.66^{\pm 0.92}$	$55.43^{\pm 0.27}$	$69.76^{\pm 0.50}$	$47.87^{\pm 0.44}$

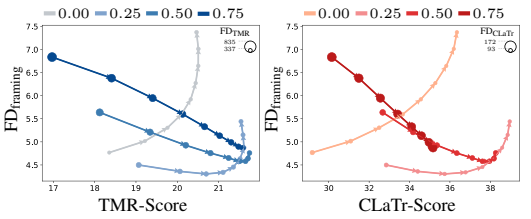
Moreover, we compare our method with baselines for DiT and MAR in Figures 15 and 16, showing the trade-off between framing quality ($FD_{framing}$) and modality-text alignment (TMR for human, CLaTr for camera) across different textual guidance values (w_c in Equation 8). The optimal point lies in the bottom-right corner of each plot (low $FD_{framing}$, high modality scores). Across both architectures and modalities, our auxiliary sampling (Aux) method achieves the best performance. It highlights its effectiveness in improving both framing quality and textual alignment on both architectures and for both modalities.

Table 9: **Auxiliary guidance ablation on the pure subset.** We vary the auxiliary guidance weight w_z to evaluate its effect on the framing, camera and human metrics. Results are reported for DiT (Peebles & Xie, 2023) and MAR (Li et al., 2024). Superscript $\pm$ denotes the 95% confidence interval over 10 samplings.

w_z	Framing		Human				Camera			
	$FD_{framing}$ ↓	Out-rate ↓	FD_{TMR} ↓	TMR-Score ↑	R3 ↑	Coverage ↑	FD_{CLaTr} ↓	CLaTr-Score ↑	F1 ↑	Coverage ↑
DiT										
0.00	$6.78^{\pm 0.06}$	$36.25^{\pm 0.36}$	$372.75^{\pm 0.94}$	$20.74^{\pm 0.10}$	$18.16^{\pm 0.25}$	$12.73^{\pm 0.31}$	$93.37^{\pm 0.78}$	$35.99^{\pm 0.20}$	$48.82^{\pm 0.39}$	$44.56^{\pm 0.33}$
0.25	$5.03^{\pm 0.03}$	$24.92^{\pm 0.28}$	$424.81^{\pm 1.07}$	$21.80^{\pm 0.12}$	$18.32^{\pm 0.15}$	$11.69^{\pm 0.19}$	$91.36^{\pm 0.81}$	$38.42^{\pm 0.31}$	$51.61^{\pm 0.47}$	$40.94^{\pm 0.19}$
0.50	$4.27^{\pm 0.03}$	$17.15^{\pm 0.30}$	$460.87^{\pm 1.33}$	$21.84^{\pm 0.11}$	$18.80^{\pm 0.25}$	$8.76^{\pm 0.15}$	$111.37^{\pm 0.74}$	$37.87^{\pm 0.23}$	$50.35^{\pm 0.33}$	$37.52^{\pm 0.27}$
0.75	$4.52^{\pm 0.03}$	$13.84^{\pm 0.32}$	$510.14^{\pm 1.41}$	$21.54^{\pm 0.12}$	$18.28^{\pm 0.27}$	$6.99^{\pm 0.16}$	$152.87^{\pm 1.01}$	$34.84^{\pm 0.28}$	$44.22^{\pm 0.39}$	$32.80^{\pm 0.29}$
MAR										
0.00	$6.55^{\pm 0.10}$	$30.19^{\pm 0.34}$	$251.94^{\pm 1.46}$	$20.16^{\pm 0.13}$	$25.48^{\pm 0.29}$	$28.25^{\pm 0.43}$	$108.28^{\pm 1.83}$	$52.17^{\pm 0.32}$	$67.31^{\pm 0.49}$	$55.48^{\pm 0.52}$
0.50	$4.90^{\pm 0.06}$	$24.28^{\pm 0.31}$	$281.39^{\pm 0.83}$	$21.90^{\pm 0.17}$	$26.43^{\pm 0.40}$	$17.48^{\pm 0.26}$	$100.66^{\pm 0.92}$	$55.43^{\pm 0.27}$	$69.76^{\pm 0.50}$	$47.87^{\pm 0.44}$
1.00	$4.62^{\pm 0.04}$	$22.85^{\pm 0.26}$	$308.10^{\pm 1.03}$	$22.46^{\pm 0.14}$	$26.43^{\pm 0.32}$	$15.36^{\pm 0.16}$	$134.96^{\pm 0.98}$	$52.74^{\pm 0.28}$	$61.26^{\pm 0.30}$	$41.28^{\pm 0.38}$
1.50	$4.75^{\pm 0.03}$	$23.19^{\pm 0.36}$	$330.03^{\pm 0.97}$	$22.63^{\pm 0.12}$	$26.44^{\pm 0.35}$	$14.31^{\pm 0.24}$	$177.61^{\pm 0.81}$	$48.47^{\pm 0.22}$	$55.72^{\pm 0.21}$	$34.24^{\pm 0.33}$

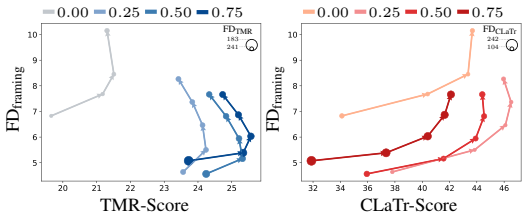
Qualitative results. Figures 17 and 18 present qualitative results with Aux sampling on the pure subset for DiT and MAR, respectively. In these examples, the human motion is accurately aligned with the prompts: in DiT, the person jumps as specified, and in MAR, the person leans forward. In the DiT example, the camera follows the person closely both vertically and laterally, maintaining proper on-screen framing, while in both cases the camera performs smooth pull-out motions. These examples further demonstrate that Aux generates human and camera behavior that faithfully follows the input prompts, producing precise motion and well-framed sequences across architectures.

E.2.4 Ablation study on pure dataset



(a) $FD_{framing}$-TMR-Score (b) $FD_{framing}$-CLaTr-Score

Figure 19: **Auxiliary guidance ablation in DiT.** Trade-off between framing quality and modality-text alignment for textual guidance ranges from 4 to 12. The optimal region is at the bottom-right (low framing error, high alignment).



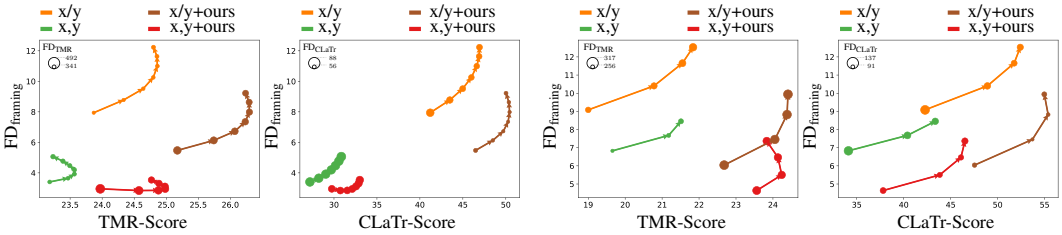
(a) $FD_{framing}$-TMR-Score (b) $FD_{framing}$-CLaTr-Score

Figure 20: **Auxiliary guidance ablation in MAR.** Trade-off between framing quality and modality-text alignment for textual guidance ranges from 1 to 5. The optimal region is at the bottom-right (low framing error, high alignment).

To assess controllability and validate the effectiveness of Aux sampling, we ablate the auxiliary guidance weight w_z (Equation (8)) on both DiT and MAR; results are shown in Table 9. We see that a moderate guidance weight improves framing and text–modality alignment. On DiT, increasing w_z from 0.00 to 0.25 reduces $FD_{framing}$ $6.78 \rightarrow 5.03$ and Out-rate $36.25 \rightarrow 24.92$; on MAR, w_z=0.50 lowers them $6.55 \rightarrow 4.90$ and $30.19 \rightarrow 24.28$.
Pushing w_z further continues to aid framing but degrades fidelity: FD_{TMR} and FD_{CLaTr} rise (DiT $424.81 \rightarrow 460.87$, MAR $281.39 \rightarrow 308.10$).
At high weights (w_z=0.75 for DiT, 1.50 for MAR), the trend becomes unstable, with FD_{TMR} spiking to 510.60 and FD_{CLaTr} to 177.61.

We then illustrate Figures 19 and 20 for the trade-off between framing quality ($FD_{framing}$, lower is better) and text–modality alignment (TMR, CLaTr; higher is better) as the Aux guidance weight w_z varies. The optimum lies near the bottom-right of each plot. Across both architectures, we see: (1) introducing guidance yields a large gain: w_z:0 $\rightarrow$ 0.25 (DiT) and 0.50 (MAR) shift points toward the bottom-right; (2) further increases, 0.50 (DiT), 1.0 (MAR), continue to improve framing but begin to reduce fidelity, reflected by larger markers (higher Fréchet distances); and (3) at very high weights, 0.75 (DiT), 1.50 (MAR), performance degrades on both axes.

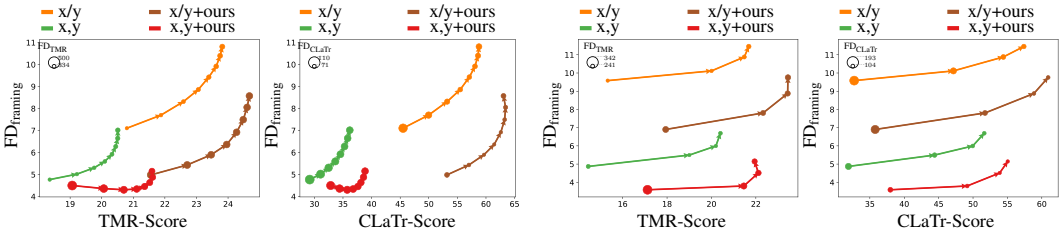
E.2.5 Ablation study on modality independence



(a) DiT TMR (b) DiT CLaTr (a) MAR TMR (b) MAR CLaTr

Figure 21: **Independent modality ablation in DiT on mixed subset.** Trade-off between framing quality and modality-text alignment for textual guidance ranges from 4 to 12. The optimal region is at the bottom-right (low framing error, high alignment).

Figure 22: **Independent modality ablation in MAR on mixed subset.** Trade-off between framing quality and modality-text alignment for textual guidance ranges from 4 to 12. The optimal region is at the bottom-right (low framing error, high alignment).



(a) DiT TMR (b) DiT CLaTr (a) MAR TMR (b) MAR CLaTr

Figure 23: **Independent modality ablation in DiT on pure subset.** Trade-off between framing quality and modality-text alignment for textual guidance ranges from 4 to 12. The optimal region is at the bottom-right (low framing error, high alignment).

Figure 24: **Independent modality ablation in MAR on pure subset.** Trade-off between framing quality and modality-text alignment for textual guidance ranges from 4 to 12. The optimal region is at the bottom-right (low framing error, high alignment).

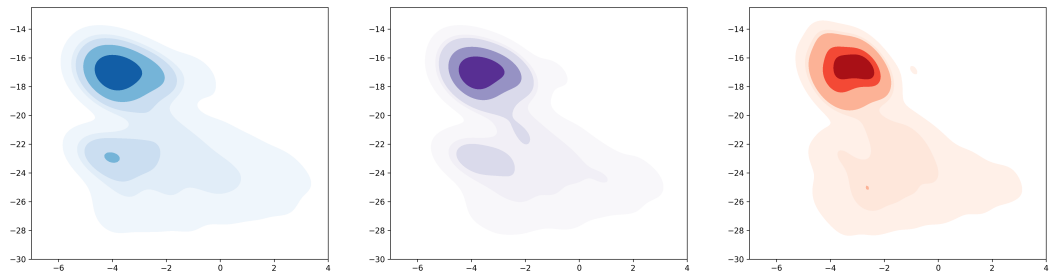
In this section, we analyze the influence of independent modality generation $(\mathbf{x}|\mathbf{y})$ versus dual-modality generation $(\mathbf{x}, \mathbf{y})$. We illustrate the trade-off between framing quality and modality-text alignment for both DiT and MAR architectures in Figure 21 and Figure 22 for the *mixed* subset, and in Figure 23 and Figure 24 for the *pure* subset.

Across all settings, the same phenomena are consistently observed:

- **The dual-modality generation setup $(\mathbf{x}|\mathbf{y})$ tends to improve inter-modality alignment at the cost of lower modality-wise performance.** This is visible in the figures when comparing green vs. orange or red vs. brown curves: the dual-modality setting appears further to the left (worse modality-wise metrics) but lower on the vertical axis (better inter-modality alignment, framing).
- **In both independent and dual-modality cases, our auxiliary sampling (Aux) consistently enhances overall performance.** Comparing green vs. red and orange vs. brown curves, Aux shifts the points toward the bottom-right, closer to the optimal balance between framing quality and modality-text alignment.

E.2.6 Qualitative visualization of auxiliary sampling influence

Figure 25 illustrates the evolution of UMAP projection density for a subset of generated samples as we vary the auxiliary sampling weight w_z. We analyze three settings: $w_z = 0.00$ in Figure 25a (no auxiliary sampling, corresponding to the $(\mathbf{x}, \mathbf{y})$ baseline); $w_z = 0.25$ in Figure 25b (the optimal



(a) $w_z = 0.00$ (b) $w_z = 0.25$ (c) $w_z = 1.00$

Figure 25: Visualization of UMAP projection density of $2,000$ generated samples for different values of auxiliary sampling weight w_z.

value used for numerical comparisons); and $w_z = 1.00$ in Figure 25c (an extreme value chosen to exaggerate the behavior).

Overall, we observe that auxiliary sampling shifts the probability mass towards a single mode. While Figure 25a exhibits bimodality with peaks near (-4,-22.5) and (-4,-16), increasing the weight to $w_z = 1.00$ concentrates the density into a single mode around (-4,-16).

We posit that this behavior mirrors the distribution tilting effect observed in guidance-based generation (Karras et al., 2024), where stronger guidance pushes samples toward high-density regions of the conditional distribution, improving prompt adherence at the cost of reduced diversity. In our case, auxiliary sampling similarly strengthens cross-modal consistency (i.e. enhancing $\text{FD}_{\text{framing}}$ in **??**) while diminishing sample diversity (i.e. decreasing the modality coverage metrics Table 8).

Input prompt:

```
Task:
  Observe the full motion sequence of the highlighted person carefully.
  Summarize their motion in one concise, accurate, and complete sentence
  in English.
Instructions:
  - Clearly state the primary global actions (e.g., walk, run, jump,
  sit, squat).
  - Specify relevant details of the motion:
  - Directions (e.g., forward, backward, left, right, clockwise).
  - Body parts involved explicitly (e.g., left arm, right foot, hips).
  - Movement trajectory patterns if present (e.g., circular, zig-zag,
  straight line).
  - Movement speed or intensity (e.g., slowly, quickly, rapidly).
  - Order and sequencing of actions clearly (what happens first,
  next, last).
  - Always describe global movement explicitly, even when emphasizing
  local actions (e.g., "A person stands still and turns their head"
  instead of just "turns head").
  - Refer only to "a person," avoiding mention of highlighting or boxes.
  Avoid:
  - Inferring emotional states or intentions.
  - Vague or overly generalized descriptions ("A person moves,"
  "A person acts").
  - Excessive spatial precision or measurement details ("moves
  exactly 5 cm forward").
  Good Examples:
  - "A person rises from sitting, walks quickly in a clockwise circle,
  then sits down again."
  - "A person jumps repeatedly side-to-side while holding their
  left wrist with the right hand, then runs rapidly backward."
  - "A person crouches, then climbs upward onto an object
  positioned to their right."
  Bad Examples (DO NOT PRODUCE):
  - Over-general or no main motion: "A person looks."
  - Too detailed spatially: "A person moves exactly
  10 centimeters left."
  - Explicit mention of highlighting: "The character in the red box jumps."
  Your response must strictly follow the format of the good examples provided.
```

Response:

```
A person points forward with their right hand while walking slightly to the left,
then stops and turns their head to the right.
```

Figure 26: **Example of human motion caption instruction.**