Odoriko: A Shape-Aware Multimodal Diffusion Framework for Human Motion

Dongseok Shim[1,†], Julian Tanke[2], Kengo Uchida[2], Christian Simon[1], Koichi Saito[2], Takashi Shibuya[2], Shusuke Takahashi[1], and Yuki Mitsufuji[1,2]

[1]Sony Group Corporation [2]Sony AI

Abstract. Human motion generation has been widely studied across diverse input modalities — text, music, and video — and recent efforts have unified these into single multimodal frameworks. However, while morphological factors such as gender and body shape are known to produce distinct kinematic signatures, no existing unified framework incorporates this into generation, treating all subjects as morphologically equivalent. We present Odoriko, the first unified multimodal motion generation framework that reflects subject bio-morphological information directly in synthesized motion output. Rather than averaging over subject variation, Odoriko generates motion that is consistent with who is moving, not just what they are asked to do — across text, music, and video conditions within a single model. When explicit morphological information is unavailable, Odoriko additionally recovers subject morphology alongside motion, unifying estimation and generation in one framework. Extensive experiments across text-to-motion, music-to-dance, and video-to-motion benchmarks demonstrate that Odoriko matches or exceeds prior specialized models on standard metrics, while enabling morphology-consistent generation that no existing unified framework supports. Project page: https://dsshim0125.github.io/odoriko.github.io/

Keywords: Human Motion · Multimodal Model · Shape-aware Generation

1 Introduction

Human motion synthesis and reconstruction are central to animation, VR, robotics, gaming, and embodied AI [19,26,39,49,59]. Recent advances in generative modeling have enabled realistic motion generation from diverse inputs such as text, music, video, and sparse keypoints, lowering the barrier for content creation and enabling richer human–computer interaction.

Despite this rapid progress, current motion generation systems remain fundamentally limited in two critical aspects. First, most models are designed as single-modality specialists, excelling at one specific task—such as text-to-motion [9,14,41,51,55,61], music-to-dance [24,28,29,31,47], or video-based pose reconstruction [45,46,50,58]—while failing to generalize across modalities. This

† Project lead and corresponding author. Email: dongseok.shim@sony.com.

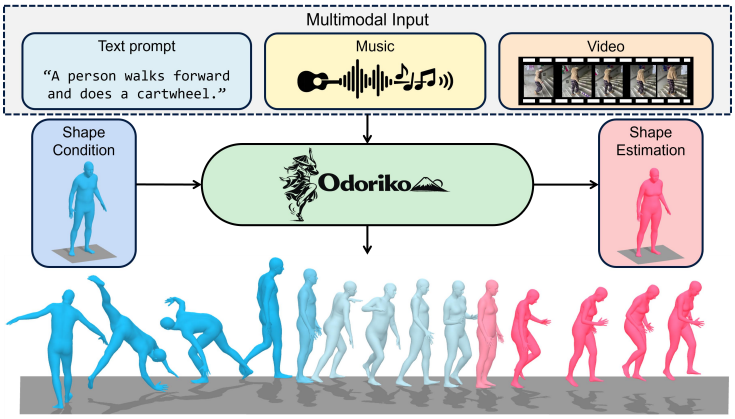

Fig. 1: We present Odoriko, a shape-aware multimodal human motion framework that unifies motion generation and 3D pose estimation within a single model. Our approach explicitly models subject-specific body shape, enabling both shape-conditioned motion synthesis and shape-aware pose estimation under diverse multimodal inputs.

fragmented design requires separate architectures and training pipelines for each modality, preventing unified control across tasks. Second, and more critically, existing approaches largely overlook the subject-specific nature of human motion. Human motion is not only a function of action intent or external conditions, but of bio-morphological characteristics of the subject such as limb proportions, mass distribution, and gender-specific structure—which systematically influence kinematics and dynamics [2,8,21,53]. For example, the same textual instruction or musical rhythm should yield systematically different motions when performed by subjects with different body shapes.

Recent works have begun incorporating body shape into motion synthesis [1, 7,32,57], highlighting the importance of morphology-aware modeling. However, most approaches focus on a single modality and rely on handcrafted or static shape representations, which limit their ability to capture rich subject-specific variation. Moreover, as they are typically designed around one input modality, they do not naturally extend to heterogeneous conditions such as music or video.

Yet, extending shape-aware modeling across modalities is inherently non-trivial, since the influence of body morphology on motion depends on context—for instance, affecting rhythmic dynamics in music-driven dance differently from semantic articulation in text-based actions. Modeling these modality-dependent shape–motion interactions within a unified framework therefore remains relatively underexplored, motivating the need for a general multimodal, shape-aware motion system.

In this work, we present **Odoriko**, a unified framework for shape-aware multimodal human motion generation and reconstruction. To the best of our knowledge, Odoriko is the first model to explicitly incorporate subject-specific shape factors either through conditioning or estimation across multiple input modalities, and to consistently reflect them in the generated motion. The framework supports text-to-motion synthesis, music-to-dance generation, and local human

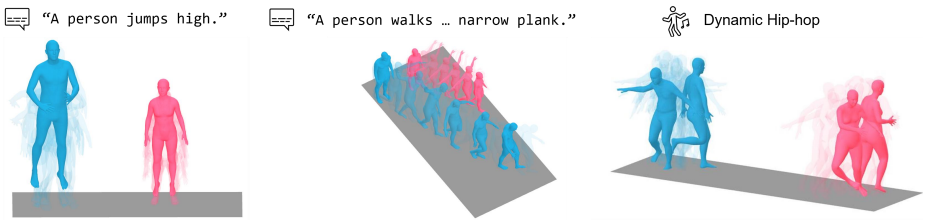

Fig. 2: Odoriko visualizations showing shape-dependent motion variations under identical text or music conditions.

motion estimation from videos and 2D keypoints within a single shared architecture, enabling controllable and anatomically coherent motion modeling across tasks and modalities.

We represent subject shape using gender and SMPL [35] $\boldsymbol{\beta}$ parameters, providing a compact yet expressive encoding of body morphology. To integrate shape with multimodal conditions, we propose *Shape-Aware Motion Transformer (SAMT)* within a diffusion-based framework, injecting shape cues at multiple network levels to modulate motion dynamics according to subject morphology. Odoriko further leverages pretrained modality-specific encoders, decoupling feature extraction from motion synthesis and enabling the generator to focus on learning shape-conditioned motion distributions.

We evaluate Odoriko on standard text-to-motion, music-to-dance, and video-to-motion (local human pose estimation) benchmarks, and introduce new protocols to quantify shape fidelity in generated motions. Results show that Odoriko matches or surpasses state-of-the-art performance in motion quality and diversity while more faithfully capturing subject-specific shape characteristics.

In short, our contribution can be summarized as follows.

- We introduce **Odoriko**, the first unified framework for shape-aware multimodal human motion generation and reconstruction, capable of handling diverse input modalities within a single model.
- We propose the Shape-Aware Motion Transformer (SAMT) that effectively injects subject shape cues jointly with multimodal conditioning signals.
- Extensive experiments across multiple benchmarks, including text-to-motion, music-to-dance, and local human motion estimation, demonstrate that Odoriko achieves competitive or superior performance while generalizing across input modalities.
- We introduce shape-aware motion generation benchmarks that evaluate subject–motion consistency, demonstrating Odoriko's superiority in preserving shape-dependent motion patterns and highlighting the importance of shape awareness in human motion generation and estimation.

2 Related Work

Human Motion Generation and Estimation. Text-driven motion synthesis has advanced rapidly through deep generative models. Diffusion-based meth-

ods [51, 62] progressively denoise motion representations to produce semantically aligned sequences, while latent variants [5, 9] improve efficiency by operating in compact latent spaces. Discrete approaches leveraging VQ-VAE pretraining [14, 41] and masked autoregressive modeling [38] further improve quality and long-range temporal coherence. For music-conditioned dance generation, methods have evolved from autoregressive Transformers [31] and GAN-based models [47] to diffusion-based frameworks with strong musical alignment [29, 30, 54]. In contrast, video-based motion estimation recovers 3D pose trajectories from observations rather than synthesizing them; recent feed-forward frameworks [45, 46, 58] achieve high accuracy without generative pipelines. Unlike these single-task methods, our work unifies all three paradigms, text-to-motion, music-to-dance, and video-to-motion estimation, within a single model.

Multimodal Motion Synthesis. Building generalist motion models over heterogeneous input modalities is an emerging direction. M^3GPT [36] integrates motion and language via instruction tuning on large language models. MCM [34] and MotionCraft [3] address optimization conflicts in mixed-modality training via ControlNet-based conditioning and coarse-to-fine training strategies, respectively. Most closely related to our work, GENMO [27] unifies motion estimation and generation under a single diffusion framework, supporting diverse conditioning signals including text, music, and video. MotionLab [17] further unifies generation and editing tasks under a Motion-Condition-Motion paradigm using rectified flows. However, none of these methods incorporate subject-specific body shape as a conditioning signal, nor do they enforce biomechanically consistent motion across diverse body morphologies—both of which are central to our approach.

Shape-Aware Motion Modeling. Biological morphology—mass distribution, limb proportions, and joint torque limits—fundamentally determines how individuals move [2, 8, 21, 53]. Ignoring shape leads to physically inconsistent outputs such as floating, ground penetration, or biomechanically implausible postures. Several works integrate shape parameters into pose estimation and reconstruction with physically grounded constraints [1, 7, 13], and HUMOS [52] introduces a shape-conditioned generative model using cycle consistency and differentiable physics losses. More recently, ShapeMove [32] proposes a text-driven shape-aware generation framework combining a dedicated FSQ-VAE with a pretrained language model, and AttrMoGen [57] introduces attribute-controlled motion generation by disentangling action semantics from body attributes via a structural causal model. Despite these advances, shape-aware motion modeling in *multimodal* generation settings—where driving signals such as music or text carry no identity information—remains largely unaddressed. Our work is the first to tackle this challenge in a unified multi-task framework spanning text, audio, and video conditioning.

3 Method

In this section, we present **Odoriko**, a unified and shape-aware framework for multimodal human motion modeling. Our approach enables the generation of

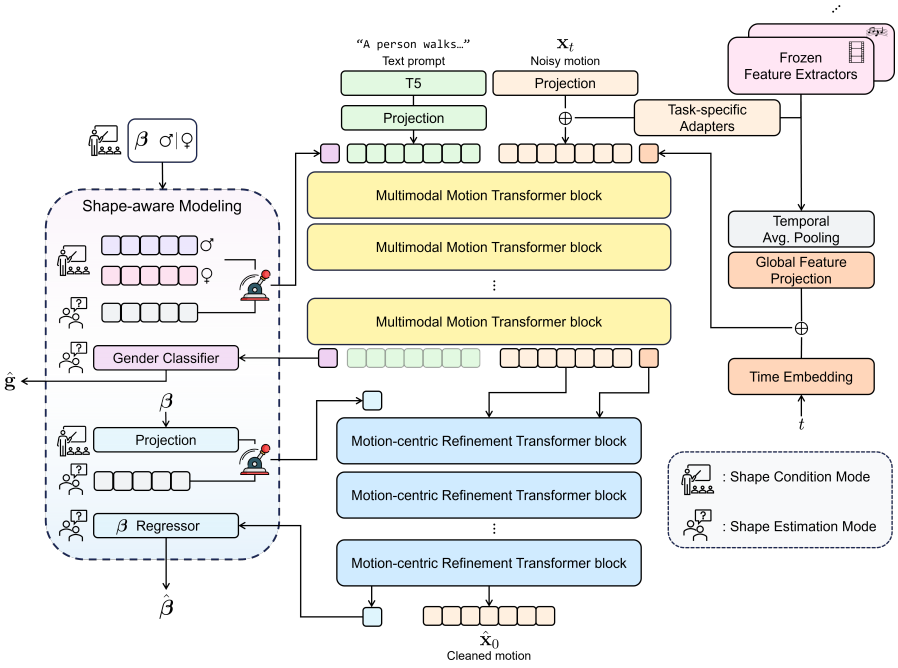

Fig. 3: Odoriko Overview. Temporally aligned multimodal inputs are fused with noisy motion tokens, while their global features condition the denoising timestep t. A shape token constructed from gender $\mathbf{g}$ and body shape $\boldsymbol{\beta}$ is integrated via joint attention, enabling explicit morphology-aware generation. The model outputs denoised motion $\hat{\mathbf{x}}_0$, and additionally predicts shape parameters in shape estimation mode. For clarity, the CLIP text embedding path is omitted.

diverse and realistic motion sequences conditioned on heterogeneous inputs, including text, music, video, and 2D keypoints.

Beyond multimodal controllability, our framework explicitly models subject-specific characteristics by incorporating intrinsic biomechanical and kinematic properties of the motion subject. This design allows the generated motions to faithfully reflect variations in body shape, gender, and physical structure, leading to improved physical plausibility and subject consistency across different generation scenarios.

3.1 Motion Representation

Human motion can be represented in various forms, including parametric body models such as SMPL [35], dataset-specific formats such as HumanML3D [15], and discrete or latent motion codes adopted in recent generative frameworks [14, 38]. The choice of representation plays a critical role in model stability, physical plausibility, and learning efficiency, particularly for diffusion-based generative models.

We directly utilize the articulated pose parameters and transform the global transformation into a rotation and translation-invariant canonical frame [20,40]. Each motion sequence $\mathbf{x}$ is represented as:

$$\mathbf{x} = \{r_z,\ \dot{r}_x,\ \dot{r}_y,\ \dot{\alpha},\ \boldsymbol{\theta}\}, \tag{1}$$

where r_z denotes the root (pelvis) height from the ground, assuming the xy-plane corresponds to the ground plane and the z-axis points upward. The terms $\dot{r}_x$ and $\dot{r}_y$ represent the root linear velocities in the ground plane, and $\dot{\alpha}$ denotes the angular velocity around the vertical (z) axis. The variable $\boldsymbol{\theta}$ corresponds to local joint rotations derived from SMPL body pose parameters. To ensure continuity and avoid singularities inherent in Euler angles or quaternions, joint rotations are encoded using a continuous 6D rotation representation [66].

To explicitly model camera-centric orientation when required (e.g., in human pose estimation settings), we introduce an auxiliary parameter $\boldsymbol{\theta}_r^l \in \mathbb{R}^6$ that represents the camera-aligned root orientation in a 6D rotation formulation. This auxiliary parameter is used exclusively for camera-dependent tasks such as pose estimation, and is omitted in generation scenarios (e.g., text-to-motion) where camera alignment is not required.

3.2 Data-Domain Diffusion Framework

We adopt a data-domain diffusion formulation [18] with direct $\mathbf{x}_0$ prediction, following prior human motion synthesis works [27–29,44,51,63]. Human motion lies on a highly structured, physically constrained manifold characterized by strong inter-joint correlations and temporal coherence. Modeling diffusion directly in the pose space preserves these geometric and kinematic dependencies, facilitating anatomically consistent and dynamically plausible motion reconstruction.

Accordingly, we employ the standard $\mathbf{x}_0$-estimation objective in the data domain:

$$\hat{\mathbf{x}}_0 = f_{\text{Odoriko}}(\mathbf{x}_t, t, \mathbf{c}, \mathbf{s}), \tag{2}$$

where $\mathbf{x}_t$ denotes the noisy motion representation at timestep t, $\mathbf{c}$ represents multimodal conditioning signals (e.g., text, music, video, or 2D keypoints), and $\mathbf{s}$ encodes subject-specific shape factors, further described in Sec. 3.3.

3.3 Model Architecture

In this section, we present the architecture of **Odoriko**, illustrated in Fig. 3. Odoriko is built upon the proposed *Shape-Aware Motion Transformer* (SAMT), a diffusion-based architecture that explicitly incorporates subject-specific body shape information to enable morphology-aware motion modeling from multimodal inputs. The detailed design of SAMT is described below.

Multimodal Conditioning. Odoriko supports heterogeneous inputs, including text, music, video, and 2D keypoints. To decouple representation learning from motion synthesis and ensure stable optimization, we adopt modality-specific

frozen pre-trained feature extractors for all modalities except 2D keypoints. This design allows the diffusion backbone to focus on motion dynamics while leveraging semantically rich representations from large-scale foundation models.

For text encoding, we employ T5-Base [43] to extract token-level contextual embeddings that capture fine-grained linguistic structure, and CLIP [42] to obtain sentence-level global representations that encode holistic semantic meaning.

For music, audio waveforms are processed by the Jukebox encoder [10], followed by the EDGE transformer [54], capturing rhythmic structure and high-level musical semantics. Music inputs are temporally resampled to align with the target motion sequence.

For video and 2D keypoints, video features are extracted using the frozen TRAM encoder [58], producing spatially compressed linearized visual embeddings. For 2D keypoints, we extract 18-joint OpenPose-style detections [4] using DWPose [60] and project them into the motion embedding space via lightweight MLP adapters.

Except for text, all modality features are temporally aligned to the motion sequence through resampling or linear interpolation.

Multimodal Motion Transformer. The diffusion backbone of **Odoriko** builds upon the Multimodal Diffusion Transformer (MM-DiT) [11], which provides a unified self-attention framework for modeling heterogeneous conditioning signals.

Following MM-DiT, text tokens from T5 are concatenated with motion tokens along the temporal axis, forming a unified token sequence processed via full self-attention. For temporally aligned modalities (music, video, and 2D keypoints), modality-specific features are projected into the motion embedding space through lightweight MLP adapters and added to the corresponding motion tokens. This additive fusion preserves temporal synchronization while maintaining parameter efficiency and stable diffusion training. Also, to handle variable-length motion sequences, we zero-pad shorter sequences and apply masked self-attention to ignore padded positions. For video-to-motion tasks that do not require textual input, text tokens are similarly masked out, so attention is restricted to valid modality tokens only.

Global Sequence-Level Conditioning. Beyond token-level cues, we incorporate a sequence-level global signal to promote long-range coherence. For all modalities except text, a global feature is obtained by temporally pooling frame-wise features from the respective frozen encoder, followed by a linear projection; for text, we directly use the CLIP sentence-level embedding, which already provides a holistic global representation. The resulting vector is added to the denoising timestep embedding and prepended to the motion token sequence as a global conditioning token, anchoring semantic consistency across the full denoising trajectory.

Efficient Hybrid Transformer Design. To balance multimodal modeling capacity and computational efficiency, we adopt a staged hybrid transformer architecture inspired by FLUX [25] and MMAudio [6]. The network is divided into two functional halves with distinct roles.

The first half consists of Multimodal Motion Transformer blocks, where motion tokens attend jointly to text tokens, global conditioning tokens, and a subject-level shape token introduced later. These layers establish cross-modal semantic grounding while preserving motion structure.

The second half comprises Motion-Centric Refinement Transformer blocks. After semantic alignment is established, text tokens are removed to reduce attention complexity, and motion tokens are further refined using contextual information from global conditioning tokens and the shape token.

Shape-Aware Modeling. A central contribution of Odoriko is the explicit integration of human body shape into multimodal motion modeling. We represent subject morphology using SMPL parameters $\mathbf{s} = \{\mathbf{g}, \boldsymbol{\beta}\}$, where biological gender $\mathbf{g} \in \{\text{male}, \text{female}\}$ selects a template and blendshape basis, and body shape coefficients $\boldsymbol{\beta} \in \mathbb{R}^{10}$ describe continuous deformations within that basis.

Importantly, SMPL exhibits a hierarchical dependency: gender determines the underlying template and basis, while $\boldsymbol{\beta}$ parameterizes variations within that template. We therefore align shape conditioning with this structure through staged injection.

Multimodal stage (gender conditioning). In the Multimodal Motion blocks, we introduce a gender token derived from a learnable embedding lookup. Since early layers perform cross-modal grounding, conditioning them on $\mathbf{g}$ provides a stable template-level identity prior during semantic alignment.

Refinement stage ($\boldsymbol{\beta}$ conditioning). In the Motion-Centric Refinement blocks, we inject a $\boldsymbol{\beta}$ token obtained by projecting the raw $\boldsymbol{\beta} \in \mathbb{R}^{10}$ through an MLP and linear layer. Because later layers refine detailed kinematics, conditioning them on $\boldsymbol{\beta}$ enables modulation of fine-grained biomechanical characteristics.

Operational Modes. Odoriko operates under a unified formulation with two modes that differ only in how the shape token is instantiated. In `shape conditioning`, a subject-level shape token is constructed from provided gender $\mathbf{g}$ and SMPL shape coefficients $\boldsymbol{\beta}$, and injected into the designated transformer stages to enforce morphology-consistent motion synthesis. In `shape estimation`, the explicit shape token is replaced by a learnable estimation token; the corresponding transformer output is decoded to predict $\hat{\mathbf{g}}$ and $\hat{\boldsymbol{\beta}}$ via a lightweight classifier and regressor, enabling joint motion reconstruction and implicit morphology inference.

By default, Odoriko operates in `shape conditioning` for generation (text-to-motion, music-to-dance) and in `shape estimation` for reconstruction/estimation (video-to-motion). When ground-truth shape annotations are unavailable (e.g., FineDance [30]), we replace explicit shape inputs with a learnable placeholder token; thus, Odoriko can be applied to such datasets even for generation, in a shape-agnostic manner, without modifying the architecture.

3.4 Loss Function and Inference

Following Sec. 3.1, Odoriko adopts direct $\mathbf{x}_0$ prediction as the diffusion parameterization. Given a noised sample $\mathbf{x}_t \sim q(\mathbf{x}_t \mid \mathbf{x}_0)$ at timestep $t \sim \mathcal{U}(0, T)$,

the network predicts the clean motion $\hat{\mathbf{x}}_0$. We optimize the standard DDPM objective [18]:

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{t,\mathbf{x}_t}\left[\|\mathbf{x}_0 - \hat{\mathbf{x}}_0\|^2\right]. \tag{3}$$

This loss directly supervises reconstruction of the clean motion signal and yields stable training with high sample fidelity.

Shape-Aware Auxiliary Supervision. When operating in `shape estimation`, Odoriko additionally predicts biological gender $\hat{\mathbf{g}}$ and SMPL body shape coefficients $\hat{\boldsymbol{\beta}}$ from the global motion representation. We apply auxiliary supervision:

$$\mathcal{L}_\beta = \left\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right\|^2, \qquad \mathcal{L}_g = \text{CE}(\mathbf{g}, \hat{\mathbf{g}}), \tag{4}$$

where $\text{CE}(\cdot)$ denotes cross-entropy for gender classification. The overall objective is

$$\mathcal{L}_{total} = \mathcal{L}_{\text{diff}} + \lambda_\beta \mathcal{L}_\beta + \lambda_g \mathcal{L}_g. \tag{5}$$

We set the auxiliary weights by mode: in `shape conditioning`, shape inputs are provided and we disable auxiliary prediction losses ($\lambda_\beta = \lambda_g = 0$); in `shape estimation`, we enable auxiliary supervision and set $\lambda_\beta = \lambda_g = 0.1$ to balance motion reconstruction and shape supervision.

Inference. In `shape conditioning`, user-specified $\mathbf{g}$ and $\boldsymbol{\beta}$ are encoded into the shape token and injected into the diffusion backbone to guide reverse denoising. In `shape estimation`, the network predicts $\hat{\mathbf{g}}$ and $\hat{\boldsymbol{\beta}}$ alongside motion reconstruction at each denoising step. To obtain stable final estimates, we aggregate the timestep-wise predictions by averaging them over the entire denoising trajectory of T steps for robust shape estimation:

$$\hat{\boldsymbol{\beta}} = \frac{1}{T}\sum_{t=1}^{T} \hat{\boldsymbol{\beta}}^{(t)}, \qquad \hat{\mathbf{g}} = \frac{1}{T}\sum_{t=1}^{T} \hat{\mathbf{g}}^{(t)}.$$

For efficient and accurate sampling, we adopt the UniPC sampler [65], a higher-order predictor–corrector solver, instead of conventional DDPM or DDIM [48] sampling schemes.

4 Experiment

4.1 Experiment Setup

Datasets. *Text-to-Motion.* We use the HumanML3D dataset [15], constructed from AMASS [37] and HumanAct12 [16]. Since our model conditions on body shape parameters ($\boldsymbol{\beta}$), we bypass the retargeting step applied in the original dataset construction and train directly on the AMASS subset, which retains ground-truth SMPL gender and shape annotations; HumanAct12 lacks such annotations and is excluded.

Table 1: Comparison with state-of-the-art methods on the HumanML3D benchmark. Odoriko is compared primarily against generalist models that handle multiple tasks and modalities. Task-specific text-to-motion methods are shown in gray for reference. Best and second-best results are highlighted.

Category	Method	#param	R-Precision ↑			FID ↓	MMDist ↓	Diversity →
			Top-1	Top-2	Top-3			
N/A	Real	-	$0.511^{\pm.003}$	$0.703^{\pm.003}$	$0.797^{\pm.002}$	$0.002^{\pm N/A}$	$2.974^{\pm.008}$	$9.503^{\pm.065}$
Text-only	MotionDiffuse [62]	87M	$0.491^{\pm.001}$	$0.681^{\pm.001}$	$0.782^{\pm.001}$	$0.630^{\pm.001}$	$3.113^{\pm.001}$	$9.410^{\pm.049}$
	MDM [51]	18M	$0.320^{\pm.005}$	$0.498^{\pm.004}$	$0.611^{\pm.007}$	$0.544^{\pm.044}$	$5.566^{\pm.027}$	$9.559^{\pm.086}$
	MLD [5]	26M	$0.481^{\pm.003}$	$0.673^{\pm.003}$	$0.772^{\pm.002}$	$0.473^{\pm.013}$	$3.196^{\pm.010}$	$9.724^{\pm.082}$
	T2M-GPT [61]	248M	$0.492^{\pm.003}$	$0.679^{\pm.002}$	$0.775^{\pm.002}$	$0.141^{\pm.005}$	$3.121^{\pm.009}$	$9.722^{\pm.081}$
	MotionLCM [9]	40M	$0.502^{\pm.003}$	$0.698^{\pm.002}$	$0.798^{\pm.002}$	$0.304^{\pm.012}$	$3.012^{\pm.007}$	$9.607^{\pm.066}$
	MMM [41]	254M	$0.504^{\pm.003}$	$0.696^{\pm.003}$	$0.794^{\pm.002}$	$0.080^{\pm.003}$	$2.998^{\pm.007}$	$9.411^{\pm.058}$
	MoMask [14]	45M	$0.521^{\pm.002}$	$0.713^{\pm.002}$	$0.807^{\pm.002}$	$0.045^{\pm.002}$	$2.958^{\pm.008}$	-
	MoLA [55]	49M	$0.516^{\pm.006}$	$0.712^{\pm.005}$	$0.805^{\pm.004}$	$0.115^{\pm.004}$	$3.008^{\pm.016}$	$9.885^{\pm.152}$
Multimodal	TM2D [12]	72M	$0.319^{\pm N/A}$	-	-	$1.021^{\pm N/A}$	$4.098^{\pm N/A}$	$9.513^{\pm N/A}$
	LMM [64]	90M	$0.496^{\pm.002}$	$0.685^{\pm.002}$	$0.785^{\pm.002}$	$0.415^{\pm.002}$	$3.087^{\pm.012}$	$9.176^{\pm.074}$
	MCM [34]	-	$0.502^{\pm.002}$	$0.692^{\pm.004}$	$0.788^{\pm.006}$	$0.053^{\pm.007}$	$3.037^{\pm.003}$	$9.585^{\pm.082}$
	MotionCraft [3]	77M	$0.501^{\pm.003}$	$0.697^{\pm.003}$	$0.796^{\pm.002}$	$0.173^{\pm.002}$	$3.025^{\pm.008}$	$9.543^{\pm.098}$
	GENMO [27]	504M	-	-	$0.632^{\pm N/A}$	$0.216^{\pm N/A}$	$3.466^{\pm N/A}$	$11.342^{\pm N/A}$
	Odoriko (Ours)	44M	$0.512^{\pm.002}$	$0.709^{\pm.003}$	$0.805^{\pm.002}$	$0.103^{\pm.004}$	$2.930^{\pm.010}$	$9.672^{\pm.070}$

Music-to-Dance. We primarily train on FineDance [30] to leverage long, high-quality dance sequences, and additionally use AIST++ [31] to provide explicit shape supervision via dancers' gender annotations.

Video-to-Motion. We train on EMDB [23], 3DPW [56], and Human3.6M [22], excluding EMDB Sequence 1 as it is reserved for evaluation. 2D keypoints are extracted with DWPose [60] for both training and inference.

Evaluation Metrics. Following standard protocols, we report Top-3 R-Precision, FID, MMDist, and Diversity [15, 51] for text-to-motion; FID, Diversity, and BAS [29–31] for music-to-dance; and PA-MPJPE, MPJPE, PVE, and Acceleration error [50, 58] for video-to-motion.

4.2 Comparison on Task-Specific Benchmarks

Text-to-Motion. We evaluate Odoriko on the HumanML3D benchmark and compare it with recent state-of-the-art methods. For fair comparison, we convert our generated motion representation into SMPL parameters and subsequently into the HumanML3D format using the official conversion pipeline. During evaluation, Odoriko is conditioned on the text prompts, gender, and $\boldsymbol{\beta}$ parameters provided in the HumanML3D test split. For samples originating from HumanAct12, which do not include shape annotations, we follow the dataset convention and set the gender to *male* and $\boldsymbol{\beta}$ to a zero vector, corresponding to a canonical body shape.

In Table 1, Odoriko achieves leading performance among multimodal motion generators (e.g., MotionCraft [3], GENMO [27]) and remains competitive with recent task-specific text-to-motion specialists [9, 14, 41, 55] despite jointly modeling multiple modalities.

Notably, Odoriko outperforms our primary competitor, GENMO, while using substantially fewer parameters, achieving efficiency comparable to single-

Table 2: Comparison of motion quality and diversity on the FineDance benchmark. Lower is better for FID; higher is better for BAS and Diversity. Task-specific music-to-dance methods are shown in gray for reference. Best and second-best results are highlighted. *Results computed using official checkpoints.

Category	Method	Motion Qual.		Motion Div.		BAS ↑
		FID_k ↓	FID_g ↓	Div_k ↑	Div_g ↑	
N/A	Real	-	-	9.73	7.44	0.2120
M2D-only	FACT [31]	113.38	97.05	3.36	6.37	0.1831
	MNet [24]	104.71	90.31	3.12	6.14	0.1864
	Bailando [47]	82.81	28.17	7.74	6.25	0.2029
	EDGE [54]	94.34	50.38	8.13	6.45	0.2116
	LODGE [29]	45.56	34.29	6.75	5.64	0.2397
	LODGE++ [28]	40.77	30.79	5.53	5.01	0.2423
Multimodal	MotionCraft* [3]	53.61	46.14	10.74	7.72	0.2143
	VerMo [33]	42.89	-	14.61	-	0.2162
	M^3GPT [36]	42.66	-	8.24	-	0.2261
	Odoriko (Ours)	37.73	25.36	7.33	6.49	0.2496

modality models such as MoMask [14] and MoLA [55]. Moreover, Odoriko attains the best results on most metrics while operating under a representation mismatch with the HumanML3D evaluation space—a discrepancy reported to degrade evaluation scores [27]—whereas prior works such as MCM [34] operate directly in HumanML3D's native representation.

We attribute this improvement in part to explicit shape conditioning. Most prior text-to-motion methods condition solely on text and thus absorb inter-subject body-shape variation into the learned motion prior [9,14,15,41,51,55,62]. Because AMASS comprises motion capture from subjects with diverse body morphologies, shape-dependent kinematics can become entangled with semantic motion patterns in such implicit representations. By explicitly conditioning on gender and SMPL shape coefficients $\boldsymbol{\beta}$, Odoriko encourages a clearer separation between morphology and semantic intent, allowing the model to learn motion content without compensating for shape-induced variation.

Music-to-Dance. We evaluate Odoriko on the FineDance benchmark. Unlike text-to-motion, music-to-dance generation prioritizes rhythmic precision and dynamic (temporal) coherence over semantic alignment. Since FineDance does not provide ground-truth body shape annotations, we replace the shape token with a learnable embedding, without explicitly conditioning on dancer-specific gender or $\boldsymbol{\beta}$ parameters.

As shown in Table 2, Odoriko achieves the best performance among multimodal baselines in dance quality metrics (FID_k and FID_g), and further surpasses recent dedicated music-to-dance methods in BAS. These results indicate that our shape-aware architecture does not hinder rhythmic fidelity; instead, it enhances motion realism while maintaining strong music–motion alignment.

Video-to-Motion. We evaluate Odoriko on EMDB and 3DPW for camera-centric 3D human pose estimation in Table 3. On EMDB, Odoriko surpasses the multimodal baseline GENMO, as well as achieving competitive performance compared to recent specialist methods [45, 46, 50, 58]. On 3DPW, our model attains reasonable accuracy, while retaining the flexibility of multimodal conditioning within a unified diffusion framework.

Table 3: Camera-space human motion estimation on EMDB-1 and 3DPW test split. Best and second-best results are highlighted.

Category	Method	EMDB (24)				3DPW (14)			
		PA-MPJPE ↓	MPJPE ↓	PVE ↓	Accel ↓	PA-MPJPE ↓	MPJPE ↓	PVE ↓	Accel ↓
V2M-only	TRACE [50]	70.9	109.9	127.4	25.5	50.9	79.1	95.4	28.6
	WHAM [46]	50.4	79.7	94.4	5.3	35.9	57.8	68.7	6.6
	GVHMR [45]	42.7	72.6	84.2	3.6	36.2	55.6	67.2	5.0
	TRAM [58]	45.7	74.4	86.6	4.9	35.6	59.3	69.6	4.9
Multimodal	GENMO [27]	42.5	73.0	84.8	3.8	34.6	53.9	65.8	4.8
	Ours (w/o θ_r^l)	41.1	285.9	345.9	7.2	41.1	255.8	287.4	8.6
	Ours (w/o 2D kpts)	46.6	71.1	83.7	9.4	52.8	100.2	114.4	16.3
	Ours	41.1	70.2	82.9	7.5	41.1	69.9	82.0	8.3

When we exclude the camera-centric root orientation parameter θ_r^l and instead directly calculate global orientation using ground-truth camera trajectories, performance drops significantly. This suggests that explicitly modeling camera-centric local orientation—rather than relying on global orientation signals—facilitates more stable and accurate pose reconstruction.

Similarly, removing 2D keypoint inputs leads to marginal degradation on EMDB but substantially larger drops on 3DPW. We attribute this discrepancy to dataset characteristics: EMDB primarily contains single-person videos, whereas 3DPW frequently includes multiple subjects. In multi-person scenes, 2D keypoints act as spatial cues that help the model disambiguate the target subject, resulting in more pronounced performance gains on 3DPW.

4.3 Shape-Aware Motion Generation Evaluation

We further evaluate whether generated motions faithfully reflect the underlying body shape and gender attributes. Unlike conventional motion quality metrics (e.g., FID, diversity), this evaluation explicitly measures the degree to which shape-dependent motion patterns are preserved in the synthesized sequences.

Text-to-Motion.

Shape Attribute Predictors. We train three Transformer-based predictors on the HumanML3D training set and freeze them during evaluation: a gender classifier that infers biological gender from root-relative joint trajectories; a continuous $\boldsymbol{\beta}$ regressor via direct regression; and a discretized $\boldsymbol{\beta}$ classifier that partitions each dimension into 10 bins for categorical prediction.

Evaluation Metrics. Using the above predictors, we define four shape-reflection metrics:

- **Gender Accuracy**: classification accuracy of predicted gender,
- **$\boldsymbol{\beta}$-bin Accuracy**: classification accuracy over discretized shape bins,
- **$\boldsymbol{\beta}$ MAE**: mean absolute error between predicted and ground-truth $\boldsymbol{\beta}$,
- **$\boldsymbol{\beta}$ PVE**: per-vertex error induced by predicted $\boldsymbol{\beta}$.

These metrics collectively measure whether the generated motion encodes shape-dependent dynamics consistent with the intended subject.

Baseline Evaluation Protocol. For text-to-motion methods [14,51,55,61] trained on HumanML3D, explicit shape conditioning is generally not supported. Nevertheless, since HumanML3D retargets all motions to a single canonical body

Table 4: Shape attribute evaluation on HumanML3D test set. Classification metrics (Gender Acc. and $\boldsymbol{\beta}$-bin Acc.) favor higher values; regression metrics ($\boldsymbol{\beta}$ PVE and MAE) favor lower. SA: shape-awareness during generation.

Method	SA	Gender Acc. ↑	$\boldsymbol{\beta}$-bin (10) ↑	$\boldsymbol{\beta}$ PVE ↓	$\boldsymbol{\beta}$ MAE ↓
Real	-	96.9	93.0	1.88	0.102
MDM [51]	✗	80.1	41.9	8.91	0.526
T2M-GPT [61]	✗	68.2	41.5	7.71	0.566
MoMask [14]	✗	78.6	26.3	9.06	0.614
MoLA [55]	✗	65.6	37.8	8.26	0.592
ShapeMove [32]	✓	74.5	52.0	8.67	0.574
Ours	✓	92.0	71.8	4.42	0.279

Table 5: Gender preservation on AIST++ test split. D: Gender tendency direction, M<F (female > male) or M>F (male > female); Pres.: Preservation of biomechanical tendency.

Metric	Real			Generated			Pres.
	M	F	D	M	F	D	
HEA →	0.043	0.063	M<F	0.112	0.122	M<F	✓
HSA →	0.085	0.156	M<F	0.247	0.251	M<F	✓
MCD →	0.487	0.885	M<F	0.545	0.709	M<F	✓
Gender Acc. ↑	88.5			61.5			-

shape, we use that canonical gender and $\boldsymbol{\beta}$ parameters as reference labels for evaluation. For ShapeMove [32], which conditions on shape information extracted from text prompts, we follow their prompt construction strategy for fair comparison. Specifically, we generate prompts corresponding to the ground-truth gender and $\boldsymbol{\beta}$ of each test sample using SHAPY [7], consistent with their training protocol.

Results. As shown in Table 4, Odoriko significantly outperforms both non-shape-aware and existing shape-aware baselines across all shape-reflection metrics. Combined with the motion quality improvements reported in Table 1, these results demonstrate that Odoriko not only generates high-quality motions but also faithfully preserves subject-specific shape characteristics in the synthesized outputs.

Music-to-Dance. Although music-to-dance generation is primarily driven by tempo and genre, dance motions may still exhibit latent kinematic traits associated with dancer morphology. To assess whether such traits are preserved, we evaluate generated motions on the AIST++ test split, which provides paired music, dance, and gender annotations.

We adopt two complementary evaluation strategies. First, following text-to-motion protocols, we train a Transformer-based gender classifier on joint coordinates from the AIST++ training set and evaluate it on both real and generated dances. Second, we compute three biomechanics-inspired metrics from joint coordinates—hip sagittal ellipse amplitude (HEA), hip sagittal semi-major axis (HSA), and mediolateral center-of-mass displacement range (MCD)—based on prior biological motion studies [21, 53].

As shown in Table 5, generated dances exhibit gender-consistent patterns in these biomechanical metrics while achieving gender classification accuracy approaching that of real data. These results suggest that Odoriko retains morphology-related kinematic tendencies, indicating shape-aware generation even under music-driven conditions.

Qualitative results. In Fig. 2, we show that, under identical text or music conditions, Odoriko generates distinct motions when the input shape parameters change. For example, different body morphologies lead to variations in jump

Table 6: Ablation on shape-awareness across motion estimation (left) and motion generation (right).

Method	EMDB (24)			
	PA-MPJPE ↓	MPJPE ↓	PVE ↓	Accel ↓
GT shape	41.1	70.0	82.8	7.5
neutral gender	42.2	76.3	89.3	7.6
zero-$\boldsymbol{\beta}$	41.4	72.7	86.0	7.6
Estimated	41.1	70.2	82.9	7.5

Shape	HumanML3D	
	R3↑	FID↓
GT	0.805	0.103
Cano.	0.768	0.592

Table 7: FID across samplers and steps.

Sampler	Step	HumanML3D	FineDance
		FID↓	FID_k↓
UniPC	1	1.804	303.42
	5	0.906	38.65
	10	0.200	37.96
	25	0.103	37.73
	50	0.113	37.85
DDIM	25	0.107	37.74

height and balance during walking on a narrow plank. Similarly, in music-to-dance generation, the model produces stylistic variations under identical audio when conditioned on different shape inputs. These results indicate that Odoriko produces motions that consistently reflect the provided bio-morphological attributes.

4.4 Ablation Study

Effectiveness of Shape-Awareness. To verify that Odoriko truly leverages subject-specific morphology—rather than generating shape-agnostic yet plausible motions—we perform ablations on both generation and estimation tasks.

Text-to-Motion. We assess the role of shape conditioning by replacing the ground-truth gender and $\boldsymbol{\beta}$ with a canonical body (*male*, $\boldsymbol{\beta} = \mathbf{0}$) while keeping the same text prompts. As shown in Table 6, this substitution degrades FID and R-Precision (R3), indicating that removing subject-specific morphology harms both motion realism and text–motion alignment. This confirms that explicit shape conditioning is essential: Odoriko meaningfully encodes gender and body shape into the synthesized motion.

Video-to-Motion. For motion estimation, we replace the predicted gender and $\boldsymbol{\beta}$ with a neutral gender template and zero $\boldsymbol{\beta}$ at evaluation time. As reported in Table 6, this leads to consistent degradation across pose metrics, demonstrating the importance of accurate shape estimation. In contrast, substituting predicted shape with ground-truth parameters yields only marginal gains over our predictions, suggesting that Odoriko estimates morphology reliably in practice.

Inference Steps. As shown in Table 7, performance improves with more denoising steps and saturates around 25, beyond which no consistent gain is observed. Among samplers, UniPC [65] outperforms DDIM [48] at the same step budget.

5 Conclusion

In this work, we introduced Odoriko, a unified shape-aware multimodal diffusion framework for human motion modeling. Unlike prior approaches that assume a canonical body shape, Odoriko explicitly incorporates biological gender and SMPL shape parameters $\boldsymbol{\beta}$ as conditioning signals within a multimodal architecture. We proposed a Shape-Aware Motion Transformer that integrates shape tokens with heterogeneous modalities (text, music, video, and 2D keypoints) via hybrid token-wise and global conditioning, enabling joint reasoning over motion

dynamics, multimodal context, and subject-specific morphology. Odoriko further unifies motion generation and camera-centric human pose estimation within a single diffusion backbone. Extensive experiments demonstrate competitive performance across tasks, with consistent improvements in shape consistency and controllability.

6 Acknowledgments

We sincerely thank Yingruo Fan for the constructive internal review and valuable feedback that helped improve this work.

References

1. Árbol, B.R., Casas, D.: Bodyshapegpt: Smpl body shape manipulation with llms. In: European Conference on Computer Vision. pp. 388–396. Springer (2024) 2, 4
2. Barclay, C.D., Cutting, J.E., Kozlowski, L.T.: Temporal and spatial factors in gait perception that influence gender recognition. Perception & psychophysics **23**(2), 145–152 (1978) 2, 4
3. Bian, Y., Zeng, A., Ju, X., Liu, X., Zhang, Z., Liu, W., Xu, Q.: Motioncraft: Crafting whole-body motion with plug-and-play multimodal controls. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 1880–1888 (2025) 4, 10, 11
4. Cao, Z., Hidalgo, G., Simon, T., Wei, S.E., Sheikh, Y.: Openpose: Realtime multi-person 2d pose estimation using part affinity fields. IEEE transactions on pattern analysis and machine intelligence **43**(1), 172–186 (2019) 7
5. Chen, X., Jiang, B., Liu, W., Huang, Z., Fu, B., Chen, T., Yu, G.: Executing your commands via motion diffusion in latent space. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 18000–18010 (2023) 4, 10
6. Cheng, H.K., Ishii, M., Hayakawa, A., Shibuya, T., Schwing, A., Mitsufuji, Y.: Mmaudio: Taming multimodal joint training for high-quality video-to-audio synthesis. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 28901–28911 (2025) 7
7. Choutas, V., Müller, L., Huang, C.H.P., Tang, S., Tzionas, D., Black, M.J.: Accurate 3d body shape regression using metric and semantic attributes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2718–2728 (2022) 2, 4, 13
8. Cutting, J.E.: Generation of synthetic male and female walkers through manipulation of a biomechanical invariant. Perception **7**(4), 393–405 (1978) 2, 4
9. Dai, W., Chen, L.H., Wang, J., Liu, J., Dai, B., Tang, Y.: Motionlcm: Real-time controllable motion generation via latent consistency model. In: European Conference on Computer Vision (2024) 1, 4, 10, 11
10. Dhariwal, P., Jun, H., Payne, C., Kim, J.W., Radford, A., Sutskever, I.: Jukebox: A generative model for music. arXiv preprint arXiv:2005.00341 (2020) 7
11. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al.: Scaling rectified flow transformers for high-resolution image synthesis. In: Forty-first international conference on machine learning (2024) 7

12. Gong, K., Lian, D., Chang, H., Guo, C., Jiang, Z., Zuo, X., Mi, M.B., Wang, X.: Tm2d: Bimodality driven 3d dance generation via music-text integration. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9942–9952 (2023) 10
13. Gralnik, O., Gafni, G., Shamir, A.: Semantify: Simplifying the control of 3d morphable models using clip. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14554–14564 (2023) 4
14. Guo, C., Mu, Y., Javed, M.G., Wang, S., Cheng, L.: Momask: Generative masked modeling of 3d human motions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1900–1910 (2024) 1, 4, 5, 10, 11, 12, 13
15. Guo, C., Zou, S., Zuo, X., Wang, S., Ji, W., Li, X., Cheng, L.: Generating diverse and natural 3d human motions from text. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5152–5161 (2022) 5, 9, 10, 11
16. Guo, C., Zuo, X., Wang, S., Zou, S., Sun, Q., Deng, A., Gong, M., Cheng, L.: Action2motion: Conditioned generation of 3d human motions. In: Proceedings of the 28th ACM international conference on multimedia. pp. 2021–2029 (2020) 9
17. Guo, Z., Hu, Z., Soh, D.W., Zhao, N.: Motionlab: Unified human motion generation and editing via the motion-condition-motion paradigm. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13869–13879 (2025) 4
18. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020) 6, 9
19. Holden, D., Komura, T., Saito, J.: Phase-functioned neural networks for character control. ACM Transactions on Graphics (TOG) **36**(4), 1–13 (2017) 1
20. Holden, D., Saito, J., Komura, T.: A deep learning framework for character motion synthesis and editing. ACM Transactions on Graphics (2016) 6
21. Hsue, B.J., Su, F.C.: Effects of age and gender on dynamic stability during stair descent. Archives of physical medicine and rehabilitation **95**(10), 1860–1869 (2014) 2, 4, 13
22. Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. IEEE transactions on pattern analysis and machine intelligence **36**(7), 1325–1339 (2013) 10
23. Kaufmann, M., Song, J., Guo, C., Shen, K., Jiang, T., Tang, C., Zárate, J.J., Hilliges, O.: Emdb: The electromagnetic database of global 3d human pose and shape in the wild. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14632–14643 (2023) 10
24. Kim, J., Oh, H., Kim, S., Tong, H., Lee, S.: A brand new dance partner: Music-conditioned pluralistic dancing controlled by multiple dance genres. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3490–3500 (2022) 1, 11
25. Labs, B.F., Batifol, S., Blattmann, A., Boesel, F., Consul, S., Diagne, C., Dockhorn, T., English, J., English, Z., Esser, P., Kulal, S., Lacey, K., Levi, Y., Li, C., Lorenz, D., Müller, J., Podell, D., Rombach, R., Saini, H., Sauer, A., Smith, L.: Flux.1 kontext: Flow matching for in-context image generation and editing in latent space (2025), https://arxiv.org/abs/2506.15742 7
26. Li, J., Wu, J., Liu, C.K.: Object motion guided human motion synthesis. ACM Transactions on Graphics (TOG) **42**(6), 1–11 (2023) 1

27. Li, J., Cao, J., Zhang, H., Rempe, D., Kautz, J., Iqbal, U., Yuan, Y.: Genmo: A generalist model for human motion. In: International Conference on Computer Vision (2025) 4, 6, 10, 11, 12
28. Li, R., Zhang, H., Zhang, Y., Zhang, Y., Zhang, Y., Guo, J., Zhang, Y., Li, X., Liu, Y.: Lodge++: High-quality and long dance generation with vivid choreography patterns. arXiv preprint arXiv:2410.20389 (2024) 1, 6, 11
29. Li, R., Zhang, Y., Zhang, Y., Zhang, H., Guo, J., Zhang, Y., Liu, Y., Li, X.: Lodge: A coarse to fine diffusion network for long dance generation guided by the characteristic dance primitives. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1524–1534 (2024) 1, 4, 6, 10, 11
30. Li, R., Zhao, J., Zhang, Y., Su, M., Ren, Z., Zhang, H., Tang, Y., Li, X.: Finedance: A fine-grained choreography dataset for 3d full body dance generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10234–10243 (2023) 4, 8, 10
31. Li, R., Yang, S., Ross, D.A., Kanazawa, A.: Ai choreographer: Music conditioned 3d dance generation with aist++. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 13401–13412 (2021) 1, 4, 10, 11
32. Liao, T.H., Zhou, Y., Shen, Y., Huang, C.H.P., Mitra, S., Huang, J.B., Bhattacharya, U.: Shape my moves: Text-driven shape-aware synthesis of human motions. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 1917–1928 (2025) 2, 4, 13
33. Ling, Z., Han, B., Li, S., Cheng, J., Shen, H., Zou, C.: Versatilemotion: A unified framework for motion synthesis and comprehension. arXiv preprint arXiv:2411.17335 (2024) 11
34. Ling, Z., Han, B., Wong, Y., Lin, H., Kankanhalli, M., Geng, W.: Mcm: Multi-condition motion synthesis framework. In: Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24. pp. 1083–1091 (2024) 4, 10, 11
35. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: Smpl: a skinned multi-person linear model. ACM Transactions on Graphics (TOG) **34**(6), 1–16 (2015) 3, 5
36. Luo, M., Hou, R., Li, Z., Chang, H., Liu, Z., Wang, Y., Shan, S.: M^3 gpt: An advanced multimodal, multitask framework for motion comprehension and generation. Advances in Neural Information Processing Systems **37**, 28051–28077 (2024) 4, 11
37. Mahmood, N., Ghorbani, N., Troje, N.F., Pons-Moll, G., Black, M.J.: Amass: Archive of motion capture as surface shapes. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 5442–5451 (2019) 9
38. Meng, Z., Xie, Y., Peng, X., Han, Z., Jiang, H.: Rethinking diffusion for text-driven human motion generation: Redundant representations, evaluation, and masked autoregression. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 27859–27871 (2025) 4, 5
39. Parent, R.: Computer animation: algorithms and techniques. Newnes (2012) 1
40. Petrovich, M., Litany, O., Iqbal, U., Black, M.J., Varol, G., Bin Peng, X., Rempe, D.: Multi-track timeline control for text-driven 3d human motion generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1911–1921 (2024) 6
41. Pinyoanuntapong, E., Wang, P., Lee, M., Chen, C.: Mmm: Generative masked motion model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1546–1555 (2024) 1, 4, 10, 11

42. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021) 7
43. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of machine learning research **21**(140), 1–67 (2020) 7
44. Ren, Z., Huang, S., Li, X.: Realistic human motion generation with cross-diffusion models. In: European Conference on Computer Vision. pp. 345–362. Springer (2024) 6
45. Shen, Z., Pi, H., Xia, Y., Cen, Z., Peng, S., Hu, Z., Bao, H., Hu, R., Zhou, X.: World-grounded human motion recovery via gravity-view coordinates. In: SIGGRAPH Asia 2024 Conference Papers. pp. 1–11 (2024) 1, 4, 11, 12
46. Shin, S., Kim, J., Halilaj, E., Black, M.J.: Wham: Reconstructing world-grounded humans with accurate 3d motion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2070–2080 (2024) 1, 4, 11, 12
47. Siyao, L., Yu, W., Gu, T., Lin, C., Wang, Q., Qian, C., Loy, C.C., Liu, Z.: Bailando: 3d dance generation by actor-critic gpt with choreographic memory. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11050–11059 (2022) 1, 4, 11
48. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: International Conference on Learning Representations (2021) 9, 14
49. Starke, S., Zhao, Y., Zinno, F., Komura, T.: Neural animation layering for synthesizing martial arts movements. ACM Transactions on Graphics (TOG) **40**(4), 1–16 (2021) 1
50. Sun, Y., Bao, Q., Liu, W., Mei, T., Black, M.J.: 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. pp. 8856–8866 (2023) 1, 10, 11, 12
51. Tevet, G., Raab, S., Gordon, B., Shafir, Y., Cohen-Or, D., Bermano, A.H.: Human motion diffusion model. In: The Eleventh International Conference on Learning Representations (2023) 1, 4, 6, 10, 11, 12, 13
52. Tripathi, S., Taheri, O., Lassner, C., Black, M., Holden, D., Stoll, C.: Humos: Human motion model conditioned on body shape. In: European Conference on Computer Vision. pp. 133–152. Springer (2024) 4
53. Troje, N.F.: Decomposing biological motion: A framework for analysis and synthesis of human gait patterns. Journal of vision **2**(5), 2–2 (2002) 2, 4, 13
54. Tseng, J., Castellon, R., Liu, K.: Edge: Editable dance generation from music. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 448–458 (2023) 4, 7, 11
55. Uchida, K., Shibuya, T., Takida, Y., Murata, N., Tanke, J., Takahashi, S., Mitsufuji, Y.: Mola: Motion generation and editing with latent diffusion enhanced by adversarial training. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 2910–2919 (2025) 1, 10, 11, 12, 13
56. Von Marcard, T., Henschel, R., Black, M.J., Rosenhahn, B., Pons-Moll, G.: Recovering accurate 3d human pose in the wild using imus and a moving camera. In: Proceedings of the European conference on computer vision (ECCV). pp. 601–617 (2018) 10
57. Wang, X., Xu, K., Li, F., Sheng, C., Yu, J., Mu, Y.: Generating attribute-aware human motions from textual prompt. In: Proceedings of the AAAI Conference on Artificial Intelligence (2026) 2, 4

58. Wang, Y., Wang, Z., Liu, L., Daniilidis, K.: Tram: Global trajectory and motion of 3d humans from in-the-wild videos. In: European Conference on Computer Vision. pp. 467–487. Springer (2024) 1, 4, 7, 10, 11, 12
59. Yamane, K., Revfi, M., Asfour, T.: Synthesizing object receiving motions of humanoid robots with human motion database. In: 2013 IEEE International Conference on Robotics and Automation. pp. 1629–1636. IEEE (2013) 1
60. Yang, Z., Zeng, A., Yuan, C., Li, Y.: Effective whole-body pose estimation with two-stages distillation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4210–4220 (2023) 7, 10
61. Zhang, J., Zhang, Y., Cun, X., Zhang, Y., Zhao, H., Lu, H., Shen, X., Shan, Y.: Generating human motion from textual descriptions with discrete representations. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14730–14740 (2023) 1, 10, 12, 13
62. Zhang, M., Cai, Z., Pan, L., Hong, F., Guo, X., Yang, L., Liu, Z.: Motiondiffuse: Text-driven human motion generation with diffusion model. IEEE transactions on pattern analysis and machine intelligence **46**(6), 4115–4128 (2024) 4, 10, 11
63. Zhang, M., Guo, X., Pan, L., Cai, Z., Hong, F., Li, H., Yang, L., Liu, Z.: Remodiffuse: Retrieval-augmented motion diffusion model. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 364–373 (2023) 6
64. Zhang, M., Jin, D., Gu, C., Hong, F., Cai, Z., Huang, J., Zhang, C., Guo, X., Yang, L., He, Y., et al.: Large motion model for unified multi-modal motion generation. In: European Conference on Computer Vision. pp. 397–421. Springer (2024) 10
65. Zhao, W., Bai, L., Rao, Y., Zhou, J., Lu, J.: Unipc: A unified predictor-corrector framework for fast sampling of diffusion models. Advances in Neural Information Processing Systems **36**, 49842–49869 (2023) 9, 14
66. Zhou, Y., Barnes, C., Lu, J., Yang, J., Li, H.: On the continuity of rotation representations in neural networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5745–5753 (2019) 6