# GENMO: A GENeralist Model for Human MOtion

Jiefeng Li Jinkun Cao Haotian Zhang Davis Rempe Jan Kautz Umar Iqbal Ye Yuan
NVIDIA
https://research.nvidia.com/labs/dair/genmo

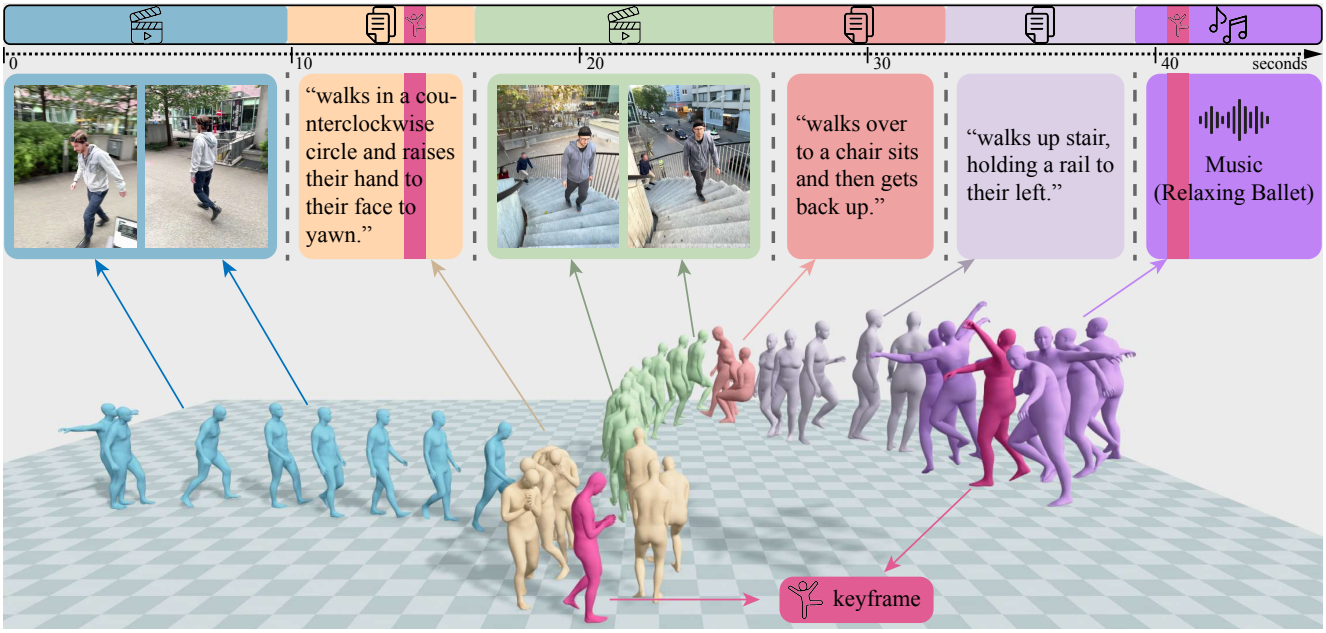


Figure 1. GENMO unifies human motion estimation and generation in a single framework and supports diverse conditioning signals including monocular videos, 2D keypoints, text descriptions, music, and 3D keyframes. GENMO can estimate accurate global human motion from videos with dynamic cameras and seamlessly handles arbitrary combinations and lengths of conditioning signals while generating smooth transitions between them. All of this is achieved in a single feedforward diffusion pass without complex post-processing.

## Abstract

*Human motion modeling traditionally separates motion generation and estimation into distinct tasks with specialized models. Motion generation models focus on creating diverse, realistic motions from inputs like text, audio, or keyframes, while motion estimation models aim to reconstruct accurate motion trajectories from observations like videos. Despite sharing underlying representations of temporal dynamics and kinematics, this separation limits knowledge transfer between tasks and requires maintaining separate models. We present GENMO, a unified Generalist Model for Human Motion that bridges motion estimation and generation in a single framework. Our key insight is to reformulate motion estimation as constrained motion generation, where the output motion must precisely satisfy observed conditioning signals. Leveraging the synergy between regression and diffusion, GENMO achieves accurate global motion estimation while enabling diverse motion generation. We also introduce an estimation-guided training objective that exploits in-the-wild videos with 2D annotations and text descriptions to enhance generative diversity. Furthermore, our novel architecture handles variable-length motions and mixed multimodal conditions (text, audio, video) at different time intervals, offering flexible control. This unified approach creates synergistic benefits: generative priors improve estimated motions under challenging conditions like occlusions, while diverse video data enhances generation capabilities. Extensive experiments demonstrate GENMO's effectiveness as a generalist framework that successfully handles multiple human motion tasks within a single model.*

## 1. Introduction

Human motion modeling is a longstanding topic in computer vision and graphics, with applications in gaming, animation, and 3D content creation. These creative applications typically require precise and intuitive user con-

trol. Consider a scenario where a user aims to generate motion sequences integrating multiple modalities: starting from a video clip, transitioning to follow textual descriptions, syncing with audio cues, and aligning with another video, all while providing fine-grained control via user-defined keyframes. Such sequences must precisely replicate observed human movements, reflect intended actions described by text or music, and adhere consistently to specified keyframes. While recent advances have made significant progress in individual tasks, achieving such precision and flexibility across multiple modalities remains challenging. Specifically, motion estimation from videos typically involves deterministic predictions focused on accuracy, whereas text/music-to-motion generation requires diversity to all possible motions. Consequently, these tasks are usually treated independently despite sharing common representations like temporal dynamics and kinematic structures. This separation limits cross-task knowledge transfer and requires maintaining distinct models.

Recent studies have revealed the synergistic relationship between motion estimation and generation tasks. Generative models [22, 49, 60] have provided robust priors for motion estimation, particularly in challenging scenarios such as world-space estimation [6, 31, 37, 69]. Conversely, leveraging large-scale video data for estimation has enhanced the realism of generative models by enriching their learned motion distributions [40]. This motivates developing a unified generalist model capable of handling both tasks concurrently across multiple modalities. However, developing such a framework presents significant challenges due to the contrasting objectives of these tasks: generation requires producing diverse and plausible outputs from abstract inputs like text or audio, while estimation demands precise motion reconstruction from concrete observations such as videos and keypoints. Creating a unified architecture that effectively balances diverse generation with accurate reconstruction while leveraging shared representations remains a complex challenge.

To address these issues, we propose GENMO, a Generalist Model for Human Motion that unifies estimation and generation within a single framework. We formulate motion estimation as constrained motion generation adhering to observed signals. This unification yields synergistic benefits: generative priors enhance plausibility in challenging estimation scenarios (e.g., occlusions), while diverse video data enrich generative diversity without requiring ground-truth 3D annotations.

GENMO is built upon a diffusion model framework incorporating a novel dual-mode training paradigm: (1) *estimation mode*, where we feed the GENMO diffusion denoiser with zero-initialized noise and the largest diffusion timestep, forcing the model to produce maximum likelihood estimation (MLE) of the motion based on the conditional signals; (2) *generation mode*, follows traditional diffusion training by sampling noisy motions and timesteps according to a predefined schedule, enabling the model to learn rich generative distributions from the conditioning signals. This dual-mode approach allows GENMO to excel at both precise estimation and diverse generation tasks. We further enhance the framework with an estimation-guided training objective that effectively leverages in-the-wild videos with 2D annotations, substantially expanding the model's generative capabilities. Furthermore, our architectural innovations enable the processing of variable-length motion sequences and seamlessly integrate arbitrary combinations of multi-modal conditioning signals at different time intervals, as demonstrated in Fig. 1. Notably, GENMO generates multi-conditioned motions in a single feedforward diffusion pass, without requiring complex post-processing steps.

Through extensive empirical evaluation, we demonstrate GENMO's capabilities across a comprehensive suite of tasks encompassing both global and local motion estimation, as well as diverse motion generation tasks including music-to-dance synthesis, text-to-motion generation, and motion-inbetweening. Our experimental results establish that GENMO achieves state-of-the-art performance across various tasks (global motion estimation, local motion estimation, and music-to-dance generation), validating its efficacy as a unified generalist framework for human motion modeling.

Our contributions are summarized as follows:

- We propose GENMO, the first generalist model unifying state-of-the-art global motion estimation with flexible human motion generation conditioned on videos, music, text, 2D keypoints, and 3D keyframes.
- Our architecture design supports seamless generation of variable-length motions conditioned on arbitrary numbers and combinations of multimodal inputs without complex post-processing.
- We propose a novel dual-mode training paradigm to explore the synergy between regression and diffusion, and introduce an estimation-guided training objective that enables effective training on in-the-wild videos.
- We demonstrate bidirectional benefits: generative priors improve estimation under challenging conditions like occlusions; conversely, diverse video data enhances generative expressiveness.

## 2. Related Work

### 2.1. Human Motion Generation

Human motion generation has progressed significantly in recent years [3, 7, 10–12, 19, 20, 22, 23, 46, 46, 49, 53, 60, 62, 67, 73, 74, 82] leveraging a variety of conditioning signals such as text [8, 13, 18, 27], actions [15], speech [1, 81], music [38, 54, 57, 59, 61, 62], and scenes/objects [21, 33,

66, 70, 78]. Recently, multimodal motion generation has also gained attention [4, 43, 75, 80] enabling multiple input modalities. However, most existing methods focus solely on generative tasks without supporting estimation. For instance, the method [75] supports video input but treats it as a generative task, resulting in motions that loosely imitate video content rather than precisely matching it. In contrast, our method jointly handles generation and estimation tasks, yielding more precise video-conditioned results.

For long-sequence motion generation, existing works mostly rely on ad-hoc post-processing techniques to stitch separately generated fixed-length motions [2, 45, 48, 76]. In contrast, our method introduces a novel diffusion-based architecture enabling seamless generation of arbitrary-length motions conditioned on multiple modalities without complex post-processing.

Existing datasets, such as AMASS [44], are limited in size and diversity. To address the scarcity of 3D data, Motion-X [40] and MotionBank [68] augment datasets using 2D videos and 3D pose estimation models [52, 72], but the resulting motions often contain artifacts. In contrast, our method directly leverages in-the-wild videos with 2D annotations without explicit 3D reconstruction, reducing reliance on noisy data and enhancing robustness and diversity.

### 2.2. Human Motion Estimation

Human pose estimation from images [28, 35, 50], videos [9, 14, 30], or even sparse marker data [34, 47, 71] has been studied extensively in the literature. Recent works focus primarily on estimating global human motion in world-space coordinates [31, 37, 52, 65, 69, 72]. This is an inherently ill-posed problem, hence these methods leverage generative priors and SLAM methods to constrain human and camera motions, respectively. However, these methods typically involve computationally expensive optimization or separate post-processing steps.

More recent approaches aim to estimate global human motion in a feed-forward manner [51, 52, 65, 77], offering faster solutions. Our method extends this direction by jointly modeling generation and estimation within a unified diffusion framework. This integration leverages shared representations and generative priors during training to produce more plausible estimations.

## 3. Generalist Model for Human Motion

GENMO unifies motion estimation and generation by formulating both tasks as conditional motion generation. Specifically, it synthesizes a human motion sequence $x$ of length $N$ based on a set of condition signals $\mathcal{C}$ and a set of corresponding condition masks $\mathcal{M}$, where $N$ can be arbitrarily large. The condition set $\mathcal{C}$ includes one or more of the following: video feature $c_{\text{video}} \in \mathbb{R}^{N\times d_{\text{video}}}$, camera motion $c_{\text{cam}} \in \mathbb{R}^{N\times d_{\text{cam}}}$, 2D skeleton $c_{\text{2d}} \in \mathbb{R}^{N\times d_{\text{2d}}}$, music clip $c_{\text{music}} \in \mathbb{R}^{N\times d_{\text{music}}}$, 2d bounding box $c_{\text{bbox}} \in \mathbb{R}^{N\times d_{\text{bbox}}}$, or natural language $c_{\text{text}} \in \mathbb{R}^{M\times d_{\text{text}}}$ that describes the motion where $M$ is the number of text tokens. The condition mask $\mathcal{M}$ consists of the mask $m_\star \in \mathbb{R}^{N\times d_\star}$ for each condition type $c_\star$ in $\mathcal{C}$. The mask matrix is of the same size as the condition feature and its element is one if the condition feature is available and zero otherwise.

**Joint Local and Global Motion Representation.** We now introduce the motion representation we use for $x$. Most text-to-motion generation methods adopt an egocentric motion representation that encodes human motion in a heading-free local coordinate system. However, for motion estimation, human motions are typically represented in the camera coordinate system to ensure better image feature alignment that facilitates learning. In this work, to obtain a unified generation and estimation model, we adopt a general human motion representation that encodes both the egocentric and camera-space human motions, along with the camera poses. Our approach leverages the gravity-view coordinate system [51], where the global trajectory of a person at frame $i$ includes the gravity-view orientation $\Gamma^i_{\text{gv}} \in \mathbb{R}^6$ and the local root velocities $v^i_{\text{root}} \in \mathbb{R}^3$. The local motion at the $i$-th frame is represented as the SMPL [42] parameters, which consists of joint angles $\theta^i \in \mathbb{R}^{24\times 6}$, shape parameters $\beta^i \in \mathbb{R}^{10}$, and root translation $t^i_{\text{root}} \in \mathbb{R}^3$. Camera pose information at frame $i$ is encoded through the camera-to-world transformation $\pi^i = \left(\Gamma^i_{\text{cv}}, t^i_{\text{cv}}\right)$, comprising the camera-view orientation $\Gamma^i_{\text{cv}} \in \mathbb{R}^6$ and camera translation $t^i_{\text{cv}} \in \mathbb{R}^3$. Additionally, we include contact labels $p^i \in \mathbb{R}^6$ for hands and feet (heels and toes). The complete motion sequence $x = \left\{x^i\right\}_{i=1}^N$ encompasses $N$ human poses, where each pose $x^i \in \mathbb{R}^D$ consists of global motion, local motion, and camera pose:

$$x^i = \left(\Gamma^i_{\text{gv}}, v^i_{\text{root}}, \theta^i, \beta^i, t^i_{\text{root}}, \pi^i, p^i\right). \tag{1}$$

### 3.1. Unified Estimation and Generation Design

In this section, we will present the architectural design of GENMO and elucidates how it unifies motion estimation and generation within a single model. The model architecture, illustrated in Figure 2, transforms a noisy motion sequence $x_t$ with the conditions $\mathcal{C}$ and condtion masks $\mathcal{M}$ into a clean motion sequence $x_0$ through a series of carefully designed components. The initial processing stage consists of an additive fusion block that converts $x_t$ into a sequence of per-frame motion tokens. This block utilizes dedicated multilayer perceptrons (MLPs) to process each condition type in $\mathcal{C}$ independently, combines their features through summation to create a unified condition representation, which is further fused with noisy motion $x_t$ to produce the motion token sequence. The resulting sequence is subsequently processed through $L$ GENMO modules, each comprising a RoPE-based Transformer block and our novel multi-text

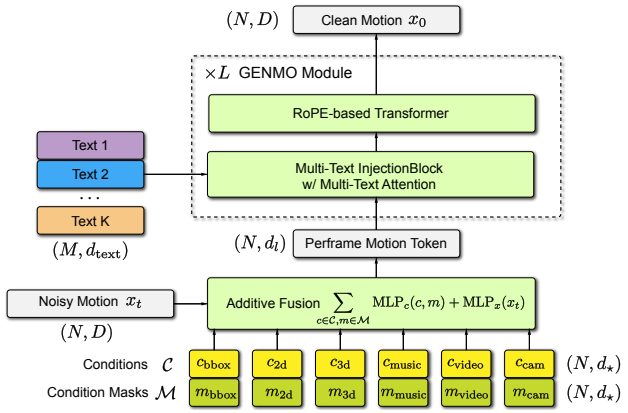


Figure 2. **GENMO Model Design** supports the generation of variable-length motion sequences in a single pass and enables seamless integration of multimodal conditioning signals, supporting both human motion generation and estimation.

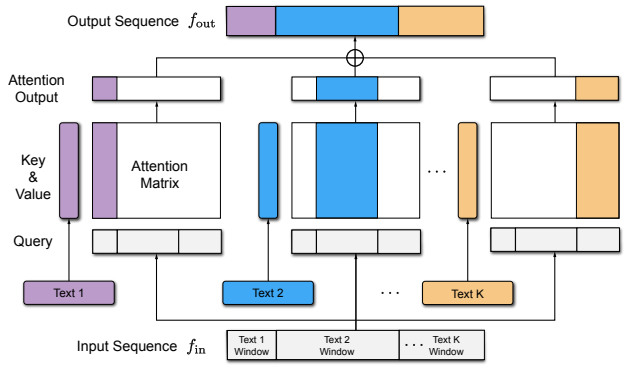


Figure 3. **Multi-text attention** enables flexible conditioning with multiple text inputs, each constrained to its specified time window.

injection block. Our architecture leverages Rotary Position Embedding (RoPE) [56], which computes attention based on relative temporal positions. This design choice enables processing of variable-length sequences and accommodates conditions lacking inherent temporal ordering, such as images and 2D skeletons.

However, text conditioning poses unique challenges. Unlike frame-aligned modalities such as video and music, text is not aligned with the motion frames. The conventional approach of concatenating text with the motion sequence is inadequate as inserting text at any positions can introduce temporal bias. To address this challenge, we propose a novel multi-text injection block that facilitates text-conditioned motion generation while accommodating multiple text inputs ($K$) with user-specified time windows. The multi-text injection block comprises a transformer block with our proposed multi-text attention mechanism at its core. As depicted in Figure 3, the multi-text attention mechanism processes K text embedding sequences $c^1_{\text{text}}, c^2_{\text{text}}, \ldots, c^K_{\text{text}}$ alongside the input motion feature sequence $f_{\text{in}}$ to generate the output feature sequence $f_{\text{out}}$:

$$f_{out} = \sum_{k=1}^{K} \text{MaskedMHA}\big(f_{in}, c^k_{\text{text}}, \Omega_k\big). \tag{2}$$

$$\Omega_k(i,j) = \begin{cases} 1 & \text{if } i \text{ is within time window of text } k \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

where MaskedMHA$(\cdot)$ represents a masked variant of the conventional multi-head attention mechanism. For each text input $k$, we employ a binary mask $\Omega_k$ that assumes a value of one when timestep $i$ lies within the designated time window of text $k$, and zero otherwise. Through the multiplication of attention weights with mask $\Omega_k$, we effectively restrict the influence of each text prompt to its corresponding time window. This multi-text attention mechanism enables highly flexible conditioning, allowing for any number of text inputs, each with a user-defined time window length. Although the mask introduces discontinuities at time window boundaries, GENMO successfully generates smooth motion sequences through the subsequent RoPE-based transformer block, which effectively captures and models temporal motion dynamics.

**Inference with Arbitrary Motion Length.** Our architecture employs relative positional embeddings rather than absolute embeddings for motion sequences, allowing GENMO to generate motions of arbitrary length in a single diffusion forward pass while naturally incorporating multiple text inputs across different time spans. During inference, we adopt a sliding window attention mechanism in the RoPE-based Transformer block, where each token attends only to tokens within a $W$-frame neighborhood. This design enables the generation of motion sequences longer than those seen during training while preserving computational efficiency and ensuring smooth, coherent motion transitions.

**Mixed Multimodal Conditions.** When conditioned on multiple modalities, our framework employs a principled approach for generation: text conditions, which lack frame-level alignment, are processed through our specialized multi-text attention mechanism, while frame-aligned modalities (e.g., video, music, 2D skeleton) are managed through a temporal masking strategy. As mentioned before, for each condition $c_\star$, we use a mask $m_\star$ of the same size to indicate whether the condition feature is (partially) present at each frame (one for present, zero otherwise). We also multiply the mask with the condition feature to nullify missing features. This simple yet effective approach enables seamless transitions between different conditioning modalities while maintaining temporal coherence.

### 3.2. Dual-Mode Training Paradigm

As a diffusion model, GENMO can theoretically be trained with the standard DDPM [24] objective:

$$\mathcal{L}_{\text{gen}} = \mathbb{E}_{t\sim[1,T],x_t\sim q(x_t|x_0)}\left[\left\|x_0 - \mathcal{G}(x_t, t, \mathcal{C}, \mathcal{M})\right\|^2\right], \quad (4)$$

where $t$ the sampled diffusion timestep, and $x_t$ is the noisy motion sampled from the forward diffusion process. Ideally, the model trained with this objective should be capable of generating motion sequences that satisfy the condition set $\mathcal{C}$ and mask $\mathcal{M}$, so it can be used as a motion estimation model when provided with video $c_{\text{video}}$ or 2D skeleton $c_{\text{2d}}$ conditions. However, we found that such a generative training objective is not enough to generate accurate motion sequences that are consistent with the input video. We observe a fundamental difference between motion estimation and text-to-motion generation tasks: motion estimation results exhibit substantially lower variability. To investigate this phenomenon, we trained separate diffusion models for text-to-motion generation and video-conditioned motion estimation, then visualized their predictions across all diffusion steps and different initial latent noises (Fig. 4). The results demonstrate that the video-conditioned model behaves more deterministically, in other words, the first-step prediction closely resembles predictions from subsequent steps with minimal variation. In contrast, the text-to-motion model exhibits significantly higher variance among steps. This observation has important implications for the estimation task: the accuracy of the first-step prediction becomes critical, as errors introduced early in the diffusion process are difficult to correct in later steps. Based on this insight, we propose a *dual-mode* training paradigm, which consists of (1) an *estimation mode* and (2) a *generation mode*. Intuitively, this dual-mode approach reinforces the quality of first-step predictions while maintaining the model's generative capabilities.

**Estimation Mode.** In the estimation mode, we formulate the problem as a regression task, employing maximum likelihood estimation to learn the conditional distribution $q(x|\mathcal{C}, \mathcal{M})$. This approach yields the following mean-square error (MSE) objective:

$$\mathcal{L}_{\text{est}} = \mathbb{E}_{z\sim\mathcal{N}(\mathbf{0},I)}\left[\left\|x_0 - \mathcal{G}(z, T, \mathcal{C}, \mathcal{M})\right\|^2\right]. \quad (5)$$

Rather than using noisy motion $x_t$, we utilize pure Gaussian noise $z \sim \mathcal{N}(\mathbf{0}, I)$ as input to the model, along with the corresponding maximum diffusion timestep $T$. This formulation ensures that the estimation mode aligns with the inherent variance characteristics of the diffusion model, thereby preventing conflicts between the generation and estimation modes.

To further enhance the quality of predicted motion sequences, we incorporate geometric regularization losses $\mathcal{L}_{\text{geo}}$ following established approaches in the literature [49, 60]. It involves decoding the predicted motion sequences into SMPL joints and vertices, followed by the application of constraints on world-space and camera-space vertices positions, world-space and camera-space joint positions, and joint contacts. In scenarios where only 2D annotations are available, we employ a 2D reprojection loss to effectively regularize the predicted motion sequences.

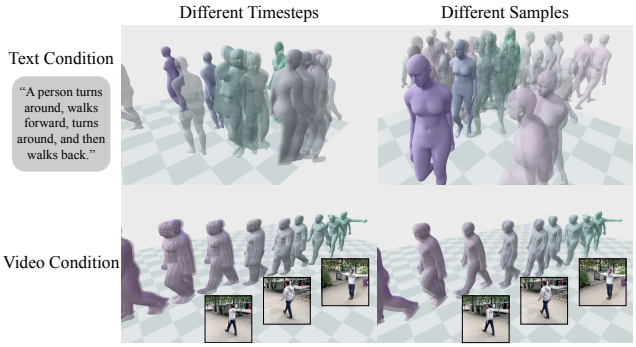


Figure 4. **Variance of video/text conditioned predictions.** Left: Intermediate predictions across 50 DDIM denoising steps. Right: Predictions with 10 different initial noises (including zero-noise). Motions are transparent except the first-step and zero-noise predictions. Video conditioning yields more deterministic outputs compared to text conditioning.

**Generation Mode.** For data with clean 3D annotations $x_0$, we can directly employ the standard diffusion objective in Eq. 4 to train the generation mode. In this section, we primarily focus on the more interesting scenario where only 2D annotations are available for the generation mode.

Unlike 3D annotations, 2D pose labels are more readily accessible through manual annotation or by applying robust 2D pose estimators on large-scale video datasets. 2D data also offers greater diversity compared to existing 3D motion capture data, which is constrained by the limited variety of subjects, motions, appearances, and environments.

Due to its inherent estimation capability, GENMO can naturally leverage 2D data for training the generation mode. Specifically, we propose an estimation-guided generation training strategy. First, we generate a pseudo-clean motion from the estimation mode using video or 2D skeleton as conditions: $\hat{x}_0 = \mathcal{G}(z, T, \mathcal{C})$. Subsequently, we sample a noisy motion sequence $\hat{x}_t$ through the forward diffusion process: $q(\hat{x}_t|\hat{x}_0)$. We then apply a 2D reprojection loss on the predicted clean motion using the 2D keypoint annotations $x_{\text{2d}}$:

$$\mathcal{L}_{\text{gen-2D}} = \mathbb{E}_{\hat{x}_t\sim q(\hat{x}_t|\hat{x}_0),t\sim[1,T]}\left[\left\|x_{\text{2d}} - \Pi(\mathcal{G}(\hat{x}_t, t, \mathcal{C}))\right\|^2\right], \quad (6)$$

where $\Pi$ represents the 2D projection function. For the generation mode, we also apply the aforementioned geometric losses $\mathcal{L}_{\text{geo}}$ to regularize the predicted motion sequences.

**Training Mode Selection.** We train the model on diverse datasets with various types of modalities. When training on datasets with strong conditioning signals that render the motion distribution more deterministic, such as video or 2D skeletons, we utilize both the estimation and generation modes to train GENMO. Conversely, when training on datasets with abstract conditions that result in more generative motion distributions, such as text and music, we exclusively employ the generation mode. This mode selection principle is applied to both 3D and 2D data. To efficiently batch the training process, we employ loss masks to selectively activate the relevant objectives for each sample, thereby ensuring that only the appropriate training modes contribute to the optimization.

# 4. Experiments

We evaluate the performance of GENMO on four different tasks including video-to-motion, music-to-dance, text-to-motion, and motion in-betweening. Note that for all experiments we use a single one-in-all checkpoint jointly trained for all tasks unless stated otherwise.

**Datasets.** GENMO is trained on a diverse collection of motion datasets spanning multiple tasks: (1) motion capture data from AMASS [44]; (2) motion estimation benchmarks including BEDLAM [5], Human3.6M [26], and 3DPW [63]; (3) music-to-dance data from AIST++ [38]; (4) text-to-motion data from HumanML3D [16]; (5) 2D keypoints and text descriptions from Motion-X [40]. Comprehensive details regarding the training procedure and implementation are provided in the supplementary material.

For evaluation, we use RICH [25], and EMDB [29] for global human motion estimation, 3DPW [63] for local human motion estimation, AIST++ [38] for music-to-dance generation, and HumanML3D [16] and Motion-X [40] for text-to-motion generation.

**Evaluation Metrics.** For music-to-dance generation, we follow the standard evaluation metrics [38, 61] and report the FID, Diversity, PFC, and BAS. For text-to-motion generation, we follow the standard evaluation metrics in previous works [16, 60] and report the R-Precision (Top 3), FID, Diversity, and MultiModal Dist. We also test the motion inbetweening performance by reporting the WA-MPJPE and PA-MPJPE for all the keyframes.

For motion estimation, we report MPJPE, PA-MPJPE, and PVE to evaluate the local motion. Acceleration error (Accel) is also reported to measure the smoothness of the motion. For global motion estimation, we report W-MPJPE$_{100}$ and WA-MPJPE$_{100}$. We also evaluate the error accumulation over long sequences by reporting RTE in %. Jitters and foot sliding (FS) during contacts are also reported. Details are provided in the supplementary material.

## 4.1. Evaluation of Motion Estimation

**Global Motion Estimation.** We compare GENMO with state-of-the-art (SOTA) methods for recovering global human motion from videos with dynamic cameras. To ensure fair comparison across methods that employ different SLAM techniques during inference, we also report results using ground-truth camera parameters provided by the datasets. As shown in Table 1, GENMO consistently outperforms specialized methods trained exclusively for human motion estimation. Notably, our approach achieves a W-MPJPE of 202.1 mm on the EMDB dataset, surpassing TRAM [65] (222.4 mm) despite both methods utilizing identical SLAM systems and backbone features for video encoding. This performance advantage stems from our unified motion generation and estimation framework, where the generative prior enhances the quality of reconstructed motions. GENMO also demonstrates superior performance on the RICH dataset compared to all existing methods. Extensive qualitative results are provided in the supplementary material.

**Local Motion Estimation.** We evaluate GENMO against SOTA methods for local 3D human motion estimation. Quantitative results in Table 2 demonstrate that GENMO surpasses existing approaches across most metrics. Additionally, we present results without training on 2D-only data, where the observed performance degradation highlights the effectiveness of our estimation-guided 2D training objective. Further evaluation on the challenging 3DPW-XOCC dataset [36] reveals that our generative prior enables GENMO to maintain robust performance even under severe occlusions and truncations. Comprehensive analyses and results on 3DPW-XOCC are provided in the supplementary material.

## 4.2. Evaluation of Motion Generation

**Comparison on Music-to-Dance.** We evaluate music-to-dance generation performance on the AIST++ dataset [38], with results presented in Table 3. GENMO is benchmarked against SOTA methods and a specialized variant of our model trained exclusively on AIST++ for music-to-dance generation. Notably, our generalist model, jointly trained across multiple estimation and generation tasks, demonstrates substantially enhanced motion diversity, physical plausibility, and motion-music correlation, as evidenced by superior Div$_k$, Div$_m$, PFC, and BAS metrics. While GENMO exhibits higher FID values compared to the specialized music-only variant, this performance differential is expected given that our generalist model was trained on considerably more heterogeneous motion data spanning multiple tasks and domains.

**Comparison on Text-to-Motion.** We evaluate the text-to-motion generation capabilities of GENMO on both Hu-

Table 1. **World-grounded human motion estimation.** We evaluate the global motion quality on the EMDB-2 [29] dataset and RICH [25]. Parenthesis denotes the number of joints used to compute WA-MPJPE$_{100}$, W-MPJPE$_{100}$ and Jitter.

| | EMDB (24) | | | | | RICH (24) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Models | WA-MPJPE$_{100}$ | W-MPJPE$_{100}$ | RTE | Jitter | Foot-Sliding | WA-MPJPE$_{100}$ | W-MPJPE$_{100}$ | RTE | Jitter | Foot-Sliding |
| GLAMR [72] | 280.8 | 726.6 | 11.4 | 46.3 | 20.7 | 129.4 | 236.2 | 3.8 | 49.7 | 18.1 |
| TRACE [58] | 529.0 | 1702.3 | 17.7 | 2987.6 | 370.7 | 238.1 | 925.4 | 610.4 | 1578.6 | 230.7 |
| SLAHMR [69] | 326.9 | 776.1 | 10.2 | 31.3 | 14.5 | 98.1 | 186.4 | 28.9 | 34.3 | 5.1 |
| COIN [37] | 152.8 | 407.3 | 3.5 | - | - | - | - | - | - | - |
| WHAM (w/ DPVO) [52] | 135.6 | 354.8 | 6.0 | 22.5 | 4.4 | 109.9 | 184.6 | 4.1 | 19.7 | 3.3 |
| WHAM (w/ GT extrinsics) [52] | 131.1 | 335.3 | 4.1 | 21.0 | 4.4 | 109.9 | 184.6 | 4.1 | 19.7 | 3.3 |
| GVHMR (w/ DPVO) [51] | 111.0 | 276.5 | 2.0 | 16.7 | **3.5** | 78.8 | 126.3 | 2.4 | **12.8** | **3.0** |
| GVHMR (w/ GT extrinsics) [51] | 109.1 | 274.9 | 1.9 | **16.5** | **3.5** | 78.8 | 126.3 | 2.4 | **12.8** | **3.0** |
| TRAM (w/ DROID-SLAM) [65] | 76.4 | 222.4 | 1.4 | - | - | - | - | - | - | - |
| Ours (w/ DROID-SLAM) | 74.3 | 202.1 | 1.2 | 17.8 | 8.8 | **75.3** | **118.6** | **1.9** | 15.0 | 6.7 |
| Ours (w/ GT extrinsics) | **69.5** | **185.9** | **0.9** | 17.7 | 8.6 | **75.3** | **118.6** | **1.9** | 15.0 | 6.7 |

Table 2. **Camera-space metrics.** We evaluate the camera-space motion quality on the 3DPW [63], RICH [25] and EMDB-1 [29] datasets. * denotes models trained with the 3DPW training set.

| | | 3DPW (14) | | | | RICH (24) | | | | EMDB (24) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Models | PA-MPJPE | MPJPE | PVE | Accel | PA-MPJPE | MPJPE | PVE | Accel | PA-MPJPE | MPJPE | PVE | Accel |
| per-frame | CLIFF* [39] | 43.0 | 69.0 | 81.2 | 22.5 | 56.6 | 102.6 | 115.0 | 22.4 | 68.1 | 103.3 | 128.0 | 24.5 |
| | HybrIK* [35] | 41.8 | 71.6 | 82.3 | – | 56.4 | 96.8 | 110.4 | – | 65.6 | 103.0 | 122.2 | – |
| | HMR2.0 [14] | 44.4 | 69.8 | 82.2 | 18.1 | 48.1 | 96.0 | 110.9 | 18.8 | 60.6 | 98.0 | 120.3 | 19.8 |
| | ReFit* [64] | 40.5 | 65.3 | 75.1 | 18.5 | 47.9 | 80.7 | 92.9 | 17.1 | 58.6 | 88.0 | 104.5 | 20.7 |
| temporal | VIBE* [30] | 51.9 | 82.9 | 98.4 | 18.5 | 68.4 | 120.5 | 140.2 | 21.8 | 81.4 | 125.9 | 146.8 | 26.6 |
| | TRACE* [58] | 50.9 | 79.1 | 95.4 | 28.6 | – | – | – | – | 70.9 | 109.9 | 127.4 | 25.5 |
| | SLAHMR [69] | 55.9 | – | – | – | 52.5 | – | – | 9.4 | 69.5 | 93.5 | 110.7 | 7.1 |
| | PACE [31] | – | – | – | – | 49.3 | – | – | 8.8 | – | – | – | – |
| | WHAM* [52] | 35.9 | 57.8 | 68.7 | 6.6 | 44.3 | 80.0 | 91.2 | 5.3 | 50.4 | 79.7 | 94.4 | 5.3 |
| | GVHMR* [51] | 36.2 | 55.6 | 67.2 | 5.0 | 39.5 | **66.0** | **74.4** | 4.1 | 42.7 | **72.6** | **84.2** | **3.6** |
| | TRAM* [65] | 35.6 | 59.3 | 69.6 | 4.9 | - | - | - | - | 45.7 | 74.4 | 86.6 | 4.9 |
| | Ours* (w/o 2D Training) | 35.2 | 55.4 | 67.0 | **4.8** | 40.6 | 66.4 | 75.4 | **4.0** | 44.3 | 76.0 | 88.9 | 4.3 |
| | Ours* | **34.6** | **53.9** | **65.8** | 5.2 | **39.1** | 66.8 | 75.4 | 4.1 | **42.5** | 73.0 | 84.8 | 3.8 |

Table 3. **Benchmark of Music-to-Dance Generation.** Motion quality is evaluated on the AIST++ [38] dataset.

| **Methods** | **FID$_k$** ↓ | **FID$_m$** ↓ | **Div$_k$** ↑ | **Div$_m$** ↑ | **PFC** ↓ | **BAS** ↑ |
|---|---|---|---|---|---|---|
| FACT [38] | 86.43 | 43.46 | 6.85 | 3.32 | 2.2543 | 0.1607 |
| Bailando [54] | 28.16 | **9.62** | 7.83 | 6.34 | 1.754 | 0.2332 |
| EDGE [61] | 42.16 | 22.12 | 3.96 | 4.61 | 1.5363 | 0.2334 |
| GCDance [41] | 36.81 | 20.10 | 6.44 | 6.47 | - | 0.2346 |
| Ours (music only) | **16.10** | 13.91 | 8.47 | 7.26 | 0.7340 | 0.2282 |
| Ours | 40.91 | 18.51 | **10.09** | **7.48** | **0.3702** | **0.2708** |

Table 4. **Benchmark of Text-to-Motion Generation** on the HumanML3D [16] dataset. R@3 denotes R-Precision (Top 3).

| **Methods** | Rep. | **R@3** ↑ | **FID** ↓ | **MM Dist** ↓ | **Diversity** → |
|---|---|---|---|---|---|
| Real | HumanML3D | 0.797 | 0.002 | 2.974 | 9.503 |
| T2M [17] | HumanML3D | 0.740 | 1.067 | 3.340 | 9.188 |
| MDM [60] | HumanML3D | 0.611 | 0.544 | 5.566 | 9.559 |
| M2DM [32] | HumanML3D | 0.763 | 0.352 | 3.134 | 9.926 |
| EMDM [79] | HumanML3D | 0.786 | 0.112 | 3.110 | 9.551 |
| MoMask [19] | HumanML3D | 0.781 | 0.204 | 2.779 | - |
| Ours (w/o 2D Training) | SMPL | 0.556 | 0.245 | 3.128 | 11.660 |
| Ours | SMPL | 0.632 | 0.216 | 3.466 | 11.342 |

manML3D (Table 4) and Motion-X (Table 5) datasets. Our method demonstrates superior performance compared to the baseline model MDM [60], exhibiting enhanced motion fidelity and improved text-prompt correspondence across both benchmarks. To assess the impact of 2D data training, we compare GENMO with its variant without training on Motion-X's 2D data. The results indicate that incorporating 2D training substantially enhances motion generation performance across both HumanML3D and Motion-X datasets. These findings substantiate the efficacy of leveraging 2D data within GENMO's framework for text-conditioned motion generation tasks.

**Discussion on HumanML3D Performance.** Although GENMO exhibits worse performance compared to SOTA methods like EMDM [79], this discrepancy can be stemmed from our representation choice: GENMO utilizes SMPL parameters to represent human motion for unified estimation and generation, whereas SOTA methods employ the HumanML3D representation — the same representation used by the encoders of the FID and R-Precision metrics. *This representational mismatch introduces an inherent disadvantage for GENMO*, as it necessitates bidirectional conversion of ground-truth motions from HumanML3D to SMPL during training and conversion of our generated motions to the HumanML3D format during evaluation. These conversion processes inevitably introduce distribution shifts through alterations in bone lengths, joint angles, and joint velocities, consequently affecting performance metrics and limiting the upper bound of GENMO's achievable performance on these HumanML3D-specific metrics.

**Experiments on Motion In-betweening.** We further evaluate the performance of conditional motion genera-

Table 5. **Benchmark of Text-to-Motion Generation.** Motion quality is evaluated on the Motion-X [40] dataset.

| **Methods** | **R@3** ↑ | **FID** ↓ | **MM Dist** ↓ | **Diversity** → |
|---|---|---|---|---|
| Real | 0.791 | 0.001 | 2.823 | 11.702 |
| MDM [60] | 0.313 | 2.389 | 6.745 | 8.720 |
| Ours (w/o 2D Training) | 0.401 | 0.515 | 5.210 | 12.124 |
| Ours | **0.472** | **0.207** | **4.801** | 11.719 |

Table 6. **Motion In-betweening Experiments.** The DDPM baseline is the proposed method without the estimation objective, only using the standard diffusion objective for training. "w/o Estimation." is the proposed method without training for the motion estimation task. "w/o 2D Training" is trained without $\mathcal{L}_{\text{gen-2D}}$. Results are reported using PA-MPJPE/WA-MPJPE.

| | HumanML3D | | Motion-X | |
|---|---|---|---|---|
| Models | 2-Keyframe | 5-Keyframe | 2-Keyframe | 5-Keyframe |
| Diffusion-only | 71.6/98.8 | 46.3/70.4 | 97.6/154.9 | 56.3/106.9 |
| w/o Estimation | 64.9/97.5 | 47.5/72.6 | 97.9/151.0 | 69.6/116.4 |
| w/o 2D Training | 56.4/**85.1** | **36.7**/59.5 | 68.3/136.8 | 44.6/98.6 |
| Ours | **53.5**/85.3 | 37.1/**58.5** | **58.8/122.7** | **40.5/89.5** |

Table 7. **Ablation studies on motion estimation.** The DDPM baseline is the proposed method without the estimation objective, only using the standard diffusion objective for training. The regression baseline is the proposed method without the generation objective.

| | RICH (24) | | EMDB (24) | |
|---|---|---|---|---|
| Models | WA-MPJPE$_{100}$ | W-MPJPE$_{100}$ | WA-MPJPE$_{100}$ | W-MPJPE$_{100}$ |
| Diffusion-only | 88.9 | 143.9 | 128.6 | 307.7 |
| Regression-only | 87.0 | 141.0 | 121.1 | 300.1 |
| Ours | **81.3** | **130.6** | **114.6** | **281.7** |

tion through the motion in-betweening task, following the methodology of prior diffusion-based approaches [60] by overwriting the noisy motion with the keyframe poses before each denoising step. Experiments are conducted on both HumanML3D and Motion-X test sets under two configurations with either 2 or 5 sampled keyframes. As shown in Table 6, GENMO achieves superior performance through its unified estimation and generation training compared to the diffusion-only baseline. Furthermore, the incorporation of additional 2D-only data and joint training with video-conditioned motion estimation substantially enhances motion in-betweening quality.

### 4.3. Ablation Study

**Impact of the Estimation Mode.** To assess the efficacy of our proposed estimation mode, we evaluate a variant of our method trained exclusively with the generation mode ("Diffusion-only"). Table 7 presents quantitative comparisons of global human motion estimation performance on the RICH and EMDB datasets, using direct model predictions without post-processing for static joints. The results demonstrate that omitting the estimation objective significantly degrades global motion estimation performance, confirming the estimation objective's crucial role in enhancing consistency between predicted motions and input videos. This finding is further corroborated by the motion in-betweening results in Table 6, which similarly indicate that the estimation objective improves in-betweening performance.

Table 8. **Effect of inference steps** on motion generation and estimation performance.

| | HumanML3D (Gen.) | EMDB (Est.) | |
|---|---|---|---|
| Models | FID | W-MPJPE$_{100}$ | MPJPE |
| Step=1 (Regression) | $0.260^{\pm.101}$ | 280.0 | 73.0 |
| Step=2 | $0.242^{\pm.083}$ | 276.8 | 72.5 |
| Step=5 | $0.231^{\pm.091}$ | **274.9** | **72.2** |
| Step=10 | $0.237^{\pm.126}$ | 275.8 | 72.3 |
| Step=50 | $\mathbf{0.216}^{\pm.119}$ | 278.7 | 72.7 |

**Impact of the Generation Mode.** We compare our unified model against a pure regression baseline (trained solely with the estimation mode, akin to SOTA human motion estimation methods) to evaluate the impact of the generation objective. Quantitative comparisons on the RICH and EMDB datasets (Table 7) reveal that our unified model consistently outperforms the regression baseline, suggesting that incorporating generative priors enhances motion quality in human motion estimation tasks.

**Different Inference Steps.** We evaluate the impact of denoising steps using the standard DDIM [55] inference pipeline. As shown in Table 8, motion estimation performance remains relatively stable across different step counts, while text-to-motion generation shows greater sensitivity. Notably, single-step denoising sufficiently produces video-consistent human motions, with optimal estimation performance achieved at 5 inference steps — a balance that effectively leverages generative priors without introducing excessive variance.

## 5. Conclusion

In this work, we introduced GENMO, a generalist framework for human motion modeling that bridges the gap between motion estimation and generation tasks. We showed that GENMO can effectively leverage shared representations to enable synergistic benefits: generative priors enhance motion estimation robustness under challenging conditions, while diverse video data enriches the generative capabilities. GENMO can produce variable-length motion generation in a single pass and supports flexible control using text, videos, music, 2D keypoints, and 3D keyframes. GENMO achieved state-of-the-art performance on both motion estimation and generation benchmarks, while also reducing reliance on 3D motion capture data. Extensive experiments demonstrated that GENMO is not only capable of handling multiple human motion tasks within a single framework but also achieves superior results compared to task-specific models.

## References


[1] Simon Alexanderson, Rajmund Nagy, Jonas Beskow, and Gustav Eje Henter. Listen, denoise, action! audio-driven motion synthesis with diffusion models. *ACM Transactions on Graphics (TOG)*, 2023. 2

[2] Nikos Athanasiou, Mathis Petrovich, Michael J. Black, and Gül Varol. TEACH: Temporal action composition for 3D humans. In *International Conference on 3D Vision (3DV)*, 2022. 3

[3] Emad Barsoum, John Kender, and Zicheng Liu. HP-GAN: Probabilistic 3D human motion prediction via GAN. In *CVPR Workshops*, 2018. 2

[4] Yuxuan Bian, Ailing Zeng, Xuan Ju, Xian Liu, Zhaoyang Zhang, Wei Liu, and Qiang Xu. Motioncraft: Crafting whole-body motion with plug-and-play multimodal controls. *arXiv preprint arXiv:2407.21136*, 2024. 3

[5] Michael J. Black, Priyanka Patel, Joachim Tesch, and Jinlong Yang. BEDLAM: A synthetic dataset of bodies exhibiting detailed lifelike animated motion. In *Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 8726–8737, 2023. 6

[6] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *ECCV*, 2016. 2

[7] Pablo Cervantes, Yusuke Sekikawa, Ikuro Sato, and Koichi Shinoda. Implicit neural representations for variable length human motion generation. In *ECCV*, 2022. 2

[8] Xin Chen, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, Jingyi Yu, and Gang Yu. Executing your commands via motion diffusion in latent space. In *CVPR*, 2023. 2

[9] Hongsuk Choi, Gyeongsik Moon, and Kyoung Mu Lee. Beyond static features for temporally consistent 3d human pose and shape from a video. In *CVPR*, 2021. 3

[10] B. Chopin, N. Otberdout, M. Daoudi, and A. Bartolo. Human motion prediction using manifold-aware wasserstein gan. In *FG*, 2021. 2

[11] Setareh Cohan, Guy Tevet, Daniele Reda, Xue Bin Peng, and Michiel van de Panne. Flexible motion in-betweening with diffusion models. *SIGGRAPH*, 2024.

[12] 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*, 2024. 2

[13] Anindita Ghosh, Noshaba Cheema, Cennet Oguz, Christian Theobalt, and Philipp Slusallek. Synthesis of compositional animations from textual descriptions. In *ICCV*, 2021. 2

[14] Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. Humans in 4d: Reconstructing and tracking humans with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14783–14794, 2023. 3, 7

[15] 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 *ACM International Conference on Multimedia (ACMMM)*, 2020. 2

[16] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5152–5161, 2022. 6, 7

[17] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5152–5161, 2022. 7

[18] 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*, 2022. 2

[19] Chuan Guo, Yuxuan Mu, Muhammad Gohar Javed, Sen Wang, and Li Cheng. Momask: Generative masked modeling of 3d human motions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1900–1910, 2024. 2, 7

[20] Ikhsanul Habibie, Daniel Holden, Jonathan Schwarz, Joe Yearsley, and Taku Komura. A recurrent variational autoencoder for human motion synthesis. In *BMVC*, 2017. 2

[21] Mohamed Hassan, Duygu Ceylan, Ruben Villegas, Jun Saito, Jimei Yang, Yi Zhou, and Michael J Black. Stochastic scene-aware motion prediction. In *ICCV*, 2021. 2

[22] Chengan He, Jun Saito, James Zachary, Holly Rushmeier, and Yi Zhou. Nemf: Neural motion fields for kinematic animation. In *NeurIPS*, 2022. 2

[23] Gustav Eje Henter, Simon Alexanderson, and Jonas Beskow. MoGlow: Probabilistic and controllable motion synthesis using normalising flows. *ACM Transactions on Graphics (TOG)*, 2020. 2

[24] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 5

[25] Chun-Hao P. Huang, Hongwei Yi, Markus Höschle, Matvey Safroshkin, Tsvetelina Alexiadis, Senya Polikovsky, Daniel Scharstein, and Michael J Black. Capturing and inferring dense full-body human-scene contact. In *CVPR*, 2022. 6, 7

[26] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments. *TPAMI*, 36(7):1325–1339, 2014. 6

[27] Peng Jin, Yang Wu, Yanbo Fan, Zhongqian Sun, Yang Wei, and Li Yuan. Act as you wish: Fine-grained control of motion diffusion model with hierarchical semantic graphs. In *NeurIPs*, 2023. 2

[28] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *CVPR*, 2018. 3

[29] Manuel Kaufmann, Jie Song, Chen Guo, Kaiyue Shen, Tianjian Jiang, Chengcheng Tang, Juan José Zárate, and Otmar Hilliges. EMDB: The Electromagnetic Database of Global 3D Human Pose and Shape in the Wild. In *ICCV*, 2023. 6, 7

[30] Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. VIBE: Video inference for human body pose and shape estimation. In *CVPR*, 2020. 3, 7

[31] Muhammed Kocabas, Ye Yuan, Pavlo Molchanov, Yunrong Guo, Michael J. Black, Otmar Hilliges, Jan Kautz, and Umar Iqbal. PACE: Human and motion estimation from in-the-wild videos. In *3DV*, 2024. 2, 3, 7

[32] Hanyang Kong, Kehong Gong, Dongze Lian, Michael Bi Mi, and Xinchao Wang. Priority-centric human motion generation in discrete latent space. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14806–14816, 2023. 7

[33] Nilesh Kulkarni, Davis Rempe, Kyle Genova, Abhijit Kundu, Justin Johnson, David Fouhey, and Leonidas Guibas. NIFTY: Neural object interaction fields for guided human motion synthesis. *arXiv:2307.07511*, 2023. 2

[34] Jiye Lee and Hanbyul Joo. Mocap everyone everywhere: Lightweight motion capture with smartwatches and a head-mounted camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 3

[35] Jiefeng Li, Chao Xu, Zhicun Chen, Siyuan Bian, Lixin Yang, and Cewu Lu. Hybrik: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation. In *CVPR*, 2021. 3, 7

[36] Jiefeng Li, Siyuan Bian, Qi Liu, Jiasheng Tang, Fan Wang, and Cewu Lu. Niki: Neural inverse kinematics with invertible neural networks for 3d human pose and shape estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12933–12942, 2023. 6

[37] Jiefeng Li, Ye Yuan, Davis Rempe, Haotian Zhang, Pavlo Molchanov, Cewu Lu, Jan Kautz, and Umar Iqbal. Coin: Control-inpainting diffusion prior for human and camera motion estimation. In *ECCV*, pages 426–446. Springer, 2024. 2, 3, 7

[38] Ruilong Li, Shan Yang, David A Ross, and Angjoo Kanazawa. Ai choreographer: Music conditioned 3d dance generation with aist++. In *ICCV*, 2021. 2, 6, 7

[39] Zhihao Li, Jianzhuang Liu, Zhensong Zhang, Songcen Xu, and Youliang Yan. Cliff: Carrying location information in full frames into human pose and shape estimation. In *ECCV*, 2022. 7

[40] Jing Lin, Ailing Zeng, Shunlin Lu, Yuanhao Cai, Ruimao Zhang, Haoqian Wang, and Lei Zhang. Motion-x: A large-scale 3d expressive whole-body human motion dataset. In *NeurIPS*, 2023. 2, 3, 6, 8

[41] Xinran Liu, Xu Dong, Diptesh Kanojia, Wenwu Wang, and Zhenhua Feng. Gcdance: Genre-controlled 3d full body dance generation driven by music. *arXiv preprint arXiv:2502.18309*, 2025. 7

[42] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: a skinned multi-person linear model. *ACM Transactions on Graphics (TOG)*, 34(6):1–16, 2015. 3

[43] Mingshuang Luo, Ruibing Hou, Zhuo Li, Hong Chang, Zimo Liu, Yaowei Wang, and Shiguang Shan. $M^3$gpt: An advanced multimodal, multitask framework for motion comprehension and generation. In *NeurIPs*, 2024. 3

[44] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes. In *ICCV*, 2019. 3, 6

[45] Mathis Petrovich, Or Litany, Umar Iqbal, Michael J. Black, Gül Varol, Xue Bin Peng, and Davis Rempe. Multi-track timeline control for text-driven 3d human motion generation. In *CVPR Workshop on Human Motion Generation*, 2024. 3

[46] Ekkasit Pinyoanuntapong, Pu Wang, Minwoo Lee, and Chen Chen. Mmm: Generative masked motion model. In *CVPR*, 2024. 2

[47] Jose Luis Ponton, Haoran Yun, Andreas Aristidou, Carlos Andujar, and Nuria Pelechano. Sparseposer: Real-time full-body motion reconstruction from sparse data. *ACM Transactions on Graphics*, 2023. 3

[48] Yijun Qian, Jack Urbanek, Alexander G. Hauptmann, and Jungdam Won. Breaking the limits of text-conditioned 3D motion synthesis with elaborative descriptions. In *ICCV*, 2023. 3

[49] Davis Rempe, Tolga Birdal, Aaron Hertzmann, Jimei Yang, Srinath Sridhar, and Leonidas J. Guibas. Humor: 3d human motion model for robust pose estimation. In *ICCV*, 2021. 2, 5

[50] István Sárándi and Gerard Pons-Moll. Neural localizer fields for continuous 3d human pose and shape estimation. 2024. 3

[51] Zehong Shen, Huaijin Pi, Yan Xia, Zhi Cen, Sida Peng, Zechen Hu, Hujun Bao, Ruizhen Hu, and Xiaowei Zhou. World-grounded human motion recovery via gravity-view coordinates. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–11, 2024. 3, 7

[52] Soyong Shin, Juyong Kim, Eni Halilaj, and Michael J Black. Wham: Reconstructing world-grounded humans with accurate 3d motion. In *CVPR*, 2024. 3, 7

[53] Ayumi Shiobara and Makoto Murakami. Human motion generation using wasserstein GAN. In *International Conference on Digital Signal Processing (ICDSP)*, 2021. 2

[54] Li Siyao, Weijiang Yu, Tianpei Gu, Chunze Lin, Quan Wang, Chen Qian, Chen Change Loy, and Ziwei Liu. Bailando: 3d dance generation by actor-critic gpt with choreographic memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11050–11059, 2022. 2, 7

[55] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2020. 8

[56] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. 4

[57] Jiangxin Sun, Chunyu Wang, Huang Hu, Hanjiang Lai, Zhi Jin, and Jian-Fang Hu. You never stop dancing: Non-freezing dance generation via bank-constrained manifold projection. In *NeurIPS*, 2022. 2

[58] 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 *CVPR*, 2023. 7

[59] Taoran Tang, Jia Jia, and Hanyang Mao. Dance with melody: An LSTM-autoencoder approach to music-oriented dance synthesis. In *ACM International Conference on Multimedia (ACMMM)*, 2018. 2

[60] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H Bermano. Human motion diffusion model. In *ICLR*, 2023. 2, 5, 6, 7, 8

[61] Jonathan Tseng, Rodrigo Castellon, and Karen Liu. Edge: Editable dance generation from music. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 448–458, 2023. 2, 6, 7

[62] Guillermo Valle-Pérez, Gustav Eje Henter, Jonas Beskow, André Holzapfel, Pierre-Yves Oudeyer, and Simon Alexanderson. Transflower: probabilistic autoregressive dance generation with multimodal attention. *ACM Transactions on Graphics (TOG)*, 2021. 2

[63] Timo von Marcard, Roberto Henschel, Michael Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *ECCV*, 2018. 6, 7

[64] Yufu Wang and Kostas Daniilidis. Refit: Recurrent fitting network for 3d human recovery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14644–14654, 2023. 7

[65] Yufu Wang, Ziyun Wang, Lingjie Liu, and Kostas Daniilidis. Tram: Global trajectory and motion of 3d humans from in-the-wild videos. In *European Conference on Computer Vision*, pages 467–487. Springer, 2024. 3, 6, 7

[66] Zan Wang, Yixin Chen, Tengyu Liu, Yixin Zhu, Wei Liang, and Siyuan Huang. HUMANISE: Language-conditioned human motion generation in 3d scenes. In *Neural Information Processing Systems (NeurIPS)*, 2022. 3

[67] Liang Xu, Ziyang Song, Dongliang Wang, Jing Su, Zhicheng Fang, Chenjing Ding, Weihao Gan, Yichao Yan, Xin Jin, Xiaokang Yang, Wenjun Zeng, and Wei Wu. ActFormer: A gan-based transformer towards general action-conditioned 3d human motion generation. In *ICCV*, 2023. 2

[68] Liang Xu, Shaoyang Hua, Zili Lin, Yifan Liu, Feipeng Ma, Yichao Yan, Xin Jin, Xiaokang Yang, and Wenjun Zeng. Motionbank: A large-scale video motion benchmark with disentangled rule-based annotations, 2024. 3

[69] Vickie Ye, Georgios Pavlakos, Jitendra Malik, and Angjoo Kanazawa. Decoupling human and camera motion from videos in the wild. In *CVPR*, 2023. 2, 3, 7

[70] Hongwei Yi, Justus Thies, Michael J Black, Xue Bin Peng, and Davis Rempe. Generating human interaction motions in scenes with text control. In *European Conference on Computer Vision*, pages 246–263. Springer, 2024. 3

[71] Xinyu Yi, Yuxiao Zhou, and Feng Xu. Transpose: Real-time 3d human translation and pose estimation with six inertial sensors. *ACM Transactions on Graphics*, 2021. 3

[72] Ye Yuan, Umar Iqbal, Pavlo Molchanov, Kris Kitani, and Jan Kautz. Glamr: Global occlusion-aware human mesh recovery with dynamic cameras. In *CVPR*, 2022. 3, 7

[73] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *arXiv preprint arXiv:2208.15001*, 2022. 2

[74] Mingyuan Zhang, Xinying Guo, Liang Pan, Zhongang Cai, Fangzhou Hong, Huirong Li, Lei Yang, and Ziwei Liu. Remodiffuse: Retrieval-augmented motion diffusion model. In *arXiv preprint arXiv:2304.01116*, 2023. 2

[75] Mingyuan Zhang, Daisheng Jin, Chenyang Gu, Fangzhou Hong, Zhongang Cai, Jingfang Huang, Chongzhi Zhang, Xinying Guo, Lei Yang, Ying He, et al. Large motion model for unified multi-modal motion generation. In *ECCV*, 2024. 3

[76] Qinsheng Zhang, Jiaming Song, Xun Huang, Yongxin Chen, and Ming yu Liu. Diffcollage: Parallel generation of large content with diffusion models. In *CVPR*, 2023. 3

[77] Siwei Zhang, Bharat Lal Bhatnagar, Yuanlu Xu, Alexander Winkler, Petr Kadlecek, Siyu Tang, and Federica Bogo. Rohm: Robust human motion reconstruction via diffusion. In *CVPR*, 2024. 3

[78] Xiaohan Zhang, Bharat Lal Bhatnagar, Sebastian Starke, Ilya A. Petrov, Vladimir Guzov, Helisa Dhamo, Eduardo Pérez Pellitero, and Gerard Pons-Moll. Force: Dataset and method for intuitive physics guided human-object interaction. In *International Conference on 3D Vision (3DV)*, 2025. 3

[79] Wenyang Zhou, Zhiyang Dou, Zeyu Cao, Zhouyingcheng Liao, Jingbo Wang, Wenjia Wang, Yuan Liu, Taku Komura, Wenping Wang, and Lingjie Liu. Emdm: Efficient motion diffusion model for fast and high-quality motion generation. In *European Conference on Computer Vision*, pages 18–38. Springer, 2024. 7

[80] Zixiang Zhou and Baoyuan Wang. Ude: A unified driving engine for human motion generation. In *CVPR*, 2023. 3

[81] Lingting Zhu, Xian Liu, Xuanyu Liu, Rui Qian, Ziwei Liu, and Lequan Yu. Taming diffusion models for audio-driven co-speech gesture generation. In *CVPR*, 2023. 2

[82] Wentao Zhu, Xiaoxuan Ma, Dongwoo Ro, Hai Ci, Jinlu Zhang, Jiaxin Shi, Feng Gao, Qi Tian, and Yizhou Wang. Human motion generation: A survey. *TPAMI*, 2023. 2