LaMP: Learning Robust Latent Motion Prior for Optimization-Based Human Motion Generation

Xiaozhong Lyu Korrawe Karunratanakul Kaifeng Zhao Siyu Tang
ETH Zürich
{xiaozhong.lyu, korrawe.karunratanakul, kaifeng.zhao, siyu.tang}@inf.ethz.ch

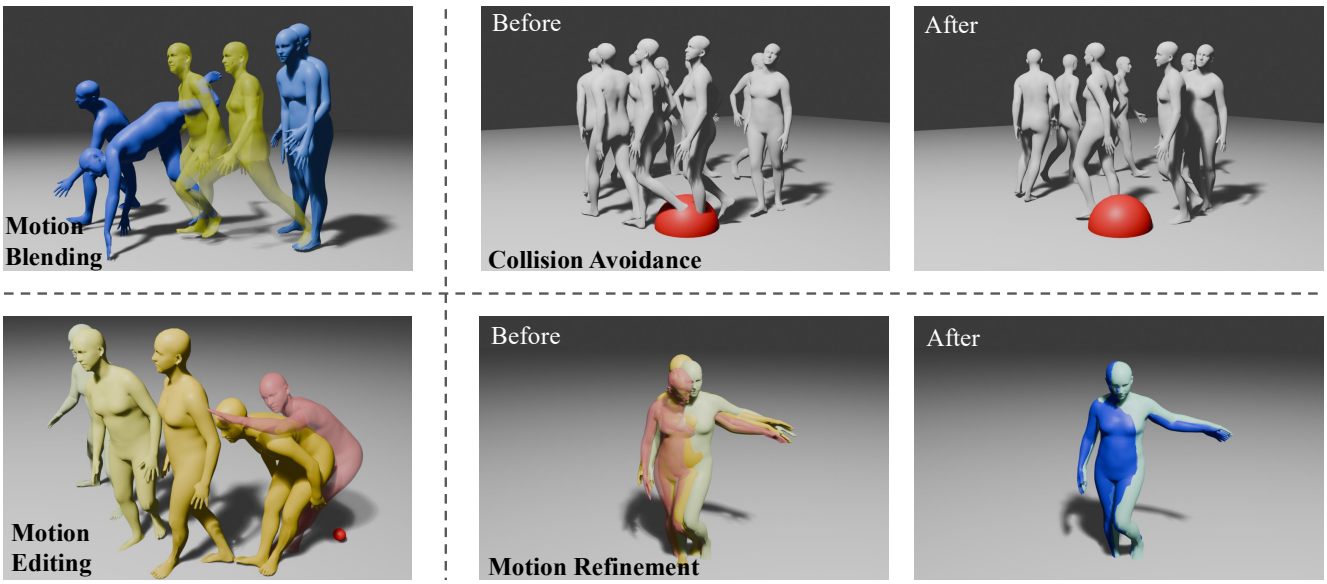

Figure 1. Our proposed **Latent Motion Prior (LaMP)** framework provides a robust, universal latent motion prior suitable for both text-to-motion generation and optimization-based downstream tasks. **Motion Blending**: LaMP generates a smooth transition (yellow) between two distinct input motions (blue), such as standing and performing a cartwheel. **Motion Editing**: The model adapts a motion sequence to reach a specific target position (indicated by the red sphere). **Collision Avoidance**: Given an input motion that intersects with an obstacle (left), LaMP optimizes the trajectory to avoid the red obstacle while preserving the original motion semantics (right). **Motion Refinement**: LaMP effectively removes artifacts from a noisy input (left) to produce a physically plausible, refined motion (right).

Abstract

We present Latent Motion Prior (LaMP) , a novel framework for learning a generalizable human motion prior that enables efficient optimization for a wide range of motion-related tasks, including text-to-motion generation, motion editing, motion blending, motion refinement, and environment-aware collision avoidance. LaMP employs a body part-based encoder to learn a disentangled latent representation of human motion, together with a masked training strategy that encourages the model to capture the most informative structural and dynamic aspects of the motion. As a result, LaMP produces a robust and expressive latent space that serves as a latent motion prior across diverse downstream tasks. We evaluate the learned representation on a wide range of optimization-based downstream tasks. Experimental results show that the current families of text-to-motion models are generally not suitable to serve as a motion prior, while LaMP consistently outperforms the state-of-the-art methods across all optimization tasks. The code is available at: https://github.com/lvsean/LaMP.

1. Introduction

Human movement is purposeful and reactive: people plan whole-body actions to achieve intentions while adapting to environmental feedback. Modeling and generating such behavior is a central goal across robotics, graphics, biome-

chanics, and virtual environments, where the key challenge is to produce realistic, controllable motion that respects physics and high-level intent. Recent data-driven approaches [2, 3, 10, 11, 41, 49] condition on text, action labels, video, or 2D/3D poses, with diffusion- and transformer-based models now leading the field and enabling coherent, human-like motion for immersive VR/AR applications.

The diversity of these motion-centric applications highlights the need for a **universal human motion prior** capable of capturing the underlying patterns across disparate tasks [19, 20, 36, 38, 39, 46]. Such human motion prior should encapsulate the inherent expectations of human movement, including kinematic constraints, realistic joint trajectories, and the fundamental dynamics of everyday actions. By filtering out anatomically or physically implausible motions, this prior serves as a critical regularizer for downstream tasks like 2D/3D pose estimation [36, 46], conditioned generation [20], and action understanding [51]. Ideally, an effective prior is data-efficient and transferable across various downstream tasks without requiring specialized models or retraining. The development of such a generalizable prior is particularly vital given the scarcity of human-motion data [12, 26, 42, 50], especially datasets for task-specific supervision, which are even more limited [34, 35, 40].

Recent lines of task-specialized motion generation work, especially for text-to-motion, have achieved notable improvements in performance but often at the cost of flexibility and generalization. Discrete tokenization methods [10, 14, 31, 32] utilize transformers, demonstrating strong performance in text-to-motion. However, by quantizing motion into discrete tokens, they discard the continuous nature of motion data, making them unsuitable for optimization-based tasks such as editing. On the other hand, diffusion-based approaches [2, 21, 41, 43, 49] can serve as powerful motion priors, but prevailing designs often optimize for a higher text-to-motion score at the expense of its applicability to other tasks by collapsing to overly low-dimensional latent space that discards kinematic detail [15]. Furthermore, we observe that adversarial objectives can inadvertently narrow the latent space, further hindering generalization to unseen tasks.

To address these challenges, we propose a novel framework LaMP to learn a robust latent motion prior. Our key insight is that the quality of a diffusion-based motion prior heavily depends on the structure of the latent space. The proposed method utilizes **a part-based encoder** to preserve local kinematic details and introduce **a spatial-temporal masking training strategy** during training. This strategy forces the model to distill the most informative features from redundant raw motion data. This enables the model to learn a compact yet expressive latent space.

Unlike previous works [2, 10, 32] that rely on either high-dimensional or overly compressed representations, our model provides a balanced latent space that enables an effective motion prior which supports various downstream tasks, including motion generation, editing, refinement, blending, and physically-aware tasks like collision avoidance. The main contributions of this paper are:

1. We introduce a novel Masked Motion Autoencoder to learn a low-dimensional and semantically meaningful representation of human motion. We developed a diffusion-based generative model over this latent space, enabling flexible and robust human motion synthesis.
2. We tested and showed that popular text-to-motion pipelines are generally not suitable as a motion prior for the downstream tasks.
3. We evaluate our method on benchmark human motion datasets and demonstrate that it produces realistic, diverse, and controllable motion sequences across diverse scenarios, including motion generation, editing, and refinement, achieving state-of-the-art performance with improved control and generality.

2. Related Work

2.1. Human motion representation

Human motion representation is fundamental in synthesizing human movement. Early approaches relied on motion capture datasets such as CMU Mocap [16], where motion was typically encoded as joint trajectories or projected into latent spaces using PCA. Traditional representations [7, 41] were largely human-designed, leveraging velocities and other kinematic or dynamic features. However, these representations are often sparse and redundant, forcing models to waste capacity on high-dimensional noise rather than learning the underlying motion distribution from limited data.

To overcome these limitations, sequence-to-sequence models [4, 25] have been employed to learn latent embeddings for motion prediction generation. Variational Autoencoders (VAEs) [2, 43] further advanced this direction by learning compact latent spaces for motion, though VAEs often struggle to preserve fine-grained motion details. Recent autoregressive models [10, 14, 32] have shifted attention to discrete vector quantization, which trades off the inherent continuity of natural human motion for learning a more regularized and structured latent space.

2.2. Human Motion Generation

Recent advancements in human motion generation have explored various modalities to drive the synthesis of realistic movements. Those works generate human motion both conditionally and unconditionally [29, 37, 45, 47, 48]. The condition signals include inputs from different modalities

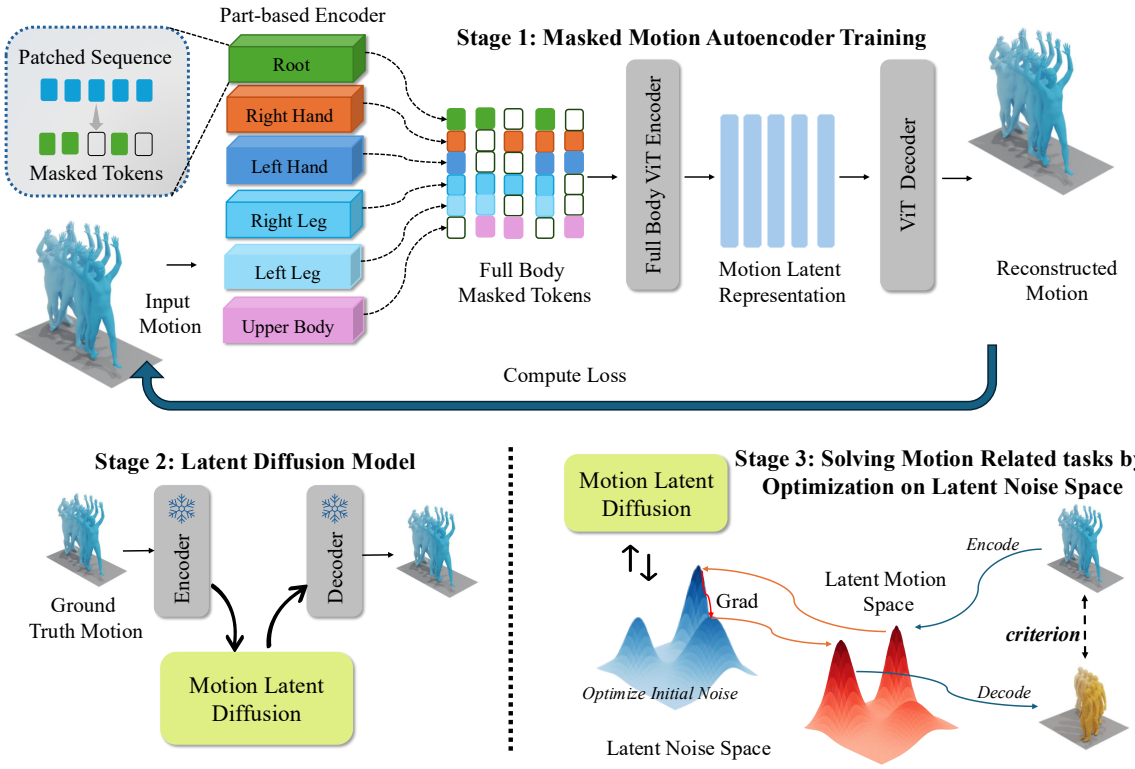

Figure 2. **Approach Overview of LaMP. Stage (a):** We first train a Masked Autoencoder to learn a robust latent space for human motion. Each training sequence is encoded using six part-based encoders and one full-body encoder to combine them. During training, a portion of the motion sequence is randomly masked. The model then reconstructs it from the remaining visible parts. **Stage (b):** We use the frozen MME to encode motion data into the latent space. Then, we train a latent diffusion model following the standard text-to-motion setting. **Stage (c):** Finally, we demonstrate that the resulting model can be used as a general motion prior for downstream tasks, including motion refinement, motion editing, motion blending, and collision avoidance, where the desired motion can be found by optimizing the latent diffusion noise.

such as text [1, 2, 7, 9, 32], audio [17, 18, 22], and actions [3, 11, 30]. Text-driven approaches leverage natural language descriptions to generate corresponding human motions, enabling intuitive control over the motion.

VQ and transformer-based methods [10, 21, 27, 31–33] remain competitive for text-to-motion generation, while lacking the ability to do fine-grained, differentiable control. Diffusion-based models offer a continuous space but also exhibit limitations. MDM [41] uses high-dimensional representations, leading to sparse and unstable latent noise spaces. MLD [2], on the other hand, compresses motions into low-dimensional vectors losing the details of motions. More recent EMDM [49] incorporating adversarial losses (e.g., GANs) narrows the optimization space, reducing the generalizability and robustness of the model.

2.3. Human Motion Prior

Recent advances in text-to-motion generation have shown impressive progress, but their generalization ability to other tasks remains limited. In computer vision, human motion priors are often employed as statistical or generative models that capture the inherent regularities of human movement. Emerging research demonstrates that learning more expressive motion priors can significantly benefit a wide range of downstream tasks [19, 20, 24, 36, 38, 39, 44, 46].

DNO [15] demonstrates that the learned diffusion space can serve as an effective prior for optimization-based motion tasks, although this optimization is confined to a joint representation space. The challenge of optimization, starting from the latent space of diffusion models [2], remains unresolved. Our model demonstrates that optimization over latent space can be achieved within the learned representation of the motion prior.

3. Method

Figure 2 shows the overall approach of LaMP, which is composed of three stages. The first stage learns a flexible and robust latent representation from raw human motion using a *Masked Motion Autoencoder*. The model architec-

ture is introduced in Sec. 3.1. In the second stage, using the frozen encoder, we learn a *Latent Diffusion Model* that generates motion tokens conditioned on text prompts, as described in Sec. 3.2. Finally, we demonstrate that the learned diffusion process can serve as a general motion prior for downstream tasks via optimization in the diffusion noise space in Sec. 3.3.

3.1. Masked Motion Autoencoder

Our goal is to learn a sequence of latent motion tokens in continuous space from partially observed raw human motion input and reconstruct the original sequence. We introduce a Masked Motion Autoencoder that follows the training paradigm of [13], with body part-based encoders that observe only a subset of spatio-temporal patches and a lightweight decoder that reconstructs the masked motion.

Notation. Let a motion clip be $\mathbf{X} \in \mathbb{R}^{L\times J}$ with L frames and a J-dimensional representation for N joints (e.g. 3D positions and rotation parameters). We partition the body skeleton into $P = 6$ disjoint parts $\mathcal{P} = \{\text{Head}, \text{Left Arm}, \text{Right Arm}, \text{Left Leg}, \text{Right Leg}, \text{Root}\}$, with part joint index sets $\mathcal{S}_p \subset \{1,\dots,J\}$ and $J_p = |\mathcal{S}_p|$ so that $\sum_{p=1}^{P} J_p = J$.

3.1.1 Part-based Encoder

Patchify the Motion We segment the sequence along time into non-overlapping windows of length τ frames, yielding $L' = \lfloor L/\tau \rfloor$ temporal patches. For each part p and time index $l \in \{1,\dots,L'\}$ we extract a spatio-temporal patch $\mathbf{X}_l^{(p)} \in \mathbb{R}^{\tau\times J_p}$, which we vectorize and linearly project to a fixed d-dimensional token:

$$\mathbf{h}_l^{(p)} = \mathbf{X}_l^{(p)}\mathbf{W}_{\text{emb}}^{(p)} + \mathbf{b}_{\text{emb}}^{(p)}, \tilde{\mathbf{h}}_l^{(p)} = \mathbf{h}_l^{(p)} + \mathbf{p}_l^{\text{time}} + \mathbf{p}_p^{\text{part}},$$

where $\mathbf{W}_{\text{emb}}^{(p)} \in \mathbb{R}^{(\tau J_p)\times d}$, $\mathbf{h}_l^{(p)} \in \mathbb{R}^d$. This yields $N = L'\cdot P$ tokens. We add a temporal positional encoding $\mathbf{p}_l^{\text{time}} \in \mathbb{R}^d$ and a learned part-based type embedding $\mathbf{p}_p^{\text{part}} \in \mathbb{R}^d$:

Random Masking Following the training paradigm from [13], we randomly sample a mask $\mathcal{M}_p \subset \{1,\dots,L'\}, p \in \{1,\dots,P\}$ for each body part by masking a fraction $\rho \in (0,1)$ of tokens uniformly. The visible set is $\mathcal{V}_p = \{1,\dots,L'\} \setminus \mathcal{M}_p$ with $|\mathcal{V}_p| = (1-\rho)L'$. Only the visible tokens are fed into the corresponding part-wise Transformer encoders, while masked tokens will be masked out as zero, which mitigates bias towards specific joints or keyframes and forces the model to learn strong motion priors across parts and time.

Full Body Transformer Encoder The model then concatenates the encoded tokens from the part-wise encoders to form a unified token sequence for the full body representation.

$$\mathbf{u}_l = \text{concat}_{p=1}^{P}\tilde{\mathbf{h}}_l^{(p)} \in \mathbb{R}^D, \quad D = P \times d,$$

Therefore, the input sequence $\mathbf{u}$ will have a shape $L' \times D$. The masked token sequence is fed to a stack of Transformer blocks (ViT-style) with pre-norm, multi-head self-attention, and MLP. Let $\mathbf{z} \in \mathbb{R}^{L'\times D}$ denote the final output from the transformer as the latent representation for the input motion $\mathbf{X}$, enabling cross-part interaction and global motion understanding. Collectively, $\mathbf{z} = \mathcal{E}(\mathbf{u}) \in \mathbb{R}^{L'\times D}$.

3.1.2 Part-based Decoder

The decoder receives learned mask tokens and reconstructs the motion from the latent representation $\mathbf{z}$. We add the same positional encoding as in the encoder and a shared Transformer decoder predicts latent representations for all tokens. A part-specific linear head projects each decoded token back to the vectorized patch space:

$$\hat{\mathbf{X}}_l = \mathcal{D}(\mathbf{z})_l \in \mathbb{R}^{\tau\times J},$$

Finally, we *unpatchify* by reshaping $\hat{\mathbf{X}}_l$ to $\mathbb{R}^{\tau\times J}$ and stitching all parts and time windows to obtain the reconstruction (we drop trailing frames if L is not divisible by τ).

3.1.3 Variational Regularization

To encourage a structured and robust latent space, we augment the encoder with a variational branch that predicts token-wise mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\sigma}$ that parameterize a diagonal Gaussian posterior for each visible token:

$$q(\mathbf{z} \mid \mathbf{X}) = \mathcal{N}\big(\boldsymbol{\mu}, \text{diag}(\boldsymbol{\sigma}^2)\big),$$

During training we apply the reparameterization trick for visible tokens and replace $\mathbf{z}$ by $\hat{\mathbf{z}}$ in the decoder input.

3.1.4 Training Objective

The overall objective is a weighted sum of three terms:

$$\mathcal{L} = \alpha\,\mathcal{L}_{\text{rec}} + \gamma\,\mathcal{L}_{\text{joint}} + \beta\,\mathcal{L}_{\text{KL}},$$

Reconstruction Loss We impose a reconstruction loss on the recovered motion representation:

$$\mathcal{L}_{\text{rec}} = \frac{1}{T'J}\sum_{t=1}^{T'}\sum_{j=1}^{J}\big\|\hat{\mathbf{X}}_t^{(j)} - \mathbf{X}_t^{(j)}\big\|_2^2,$$

Joint-space consistency After unpatchifying, we enforce accuracy in the original joint space using a forward kinematics function Π that maps predicted parameters to 3D joints.

$$\mathcal{L}_{\text{joint}} = \frac{1}{T'\tau} \sum_{s=1}^{T'\tau} \frac{1}{N} \sum_{n=1}^{N} \left\| \Pi(\hat{\mathbf{X}}_{s,n,:}) - \Pi(\mathbf{X}_{s,n,:}) \right\|_2^2,$$

KL divergence We regularize the token-wise posterior towards the standard normal prior:

$$\mathcal{L}_{\text{KL}} = D_{\text{KL}}\big(\mathcal{N}(\boldsymbol{\mu}, \text{diag}(\boldsymbol{\sigma}^2)) \parallel \mathcal{N}(\mathbf{0}, \mathbf{I})\big),$$

Modeling at the part level preserves fine-grained motion details while the shared Transformer enables cross-part information. The masking training strategy prevents shortcut copying and encourages the encoder learning a more robust motion prior. The variational prediction provides a calibrated latent space that supports sampling and improves robustness.

3.2. Latent Diffusion on Motion Tokens

After training the masked autoencoder, we freeze the encoder and follow [28] to train a latent text-to-motion diffusion model.

Denoiser For each text/motion training data pair, we learn a latent diffusion model conditioned on the text to generate latent token sequence with full observation of the motion $\mathbf{z}_0 \in \mathbb{R}^{L' \times D}$, where $D = P \cdot d$. A transformer denoiser $\boldsymbol{\epsilon}_\theta$ predicts noise conditioned on the given timestep and optional conditioning signal c (text input)

$$\widehat{\boldsymbol{\epsilon}} = \boldsymbol{\epsilon}_\theta(\mathbf{z}_t,\, t,\, c),$$

We train with the simple objective using classifier-free guidance by randomly dropping c with probability $p_\emptyset$.

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{t, \mathbf{z}_0, \boldsymbol{\epsilon}} \|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{z}_t, t, c)\|_2^2,$$

Generation At inference time, we sample $\mathbf{z}_{T_s} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, iterate reverse steps to $\mathbf{z}_0$, then decode to motion via the frozen decoder pipeline: (i) split $\mathbf{z}_0$ to tokens (t, p), (ii) run the decoder heads and unpatchify to get $\hat{\mathbf{X}}$. Guided sampling uses $\boldsymbol{\epsilon}_{\text{guided}} = (1{+}w)\, \boldsymbol{\epsilon}_\theta(\cdot, c) - w\, \boldsymbol{\epsilon}_\theta(\cdot, \emptyset)$ with scale w.

3.3. Optimization on Diffusion Noise

The learned diffusion model introduced in Sec. 3.2 is able to generate motion sequences conditioned on text prompts. Beyond text-to-motion generation, the learned diffusion process itself serves as a powerful **motion prior**. Instead of treating the denoiser as a black-box generator, we can explicitly optimize the **initial noise** to satisfy downstream objectives.

Noise optimization formulation Given an initial Gaussian noise $\boldsymbol{\epsilon}_0 \sim \mathcal{N}(0, I)$, the diffusion process generates a motion sequence via the reverse dynamics of the denoiser $\boldsymbol{\epsilon}_\theta$. Instead of fixing $\boldsymbol{\epsilon}_0$, we optimize it with respect to a downstream task objective $\mathcal{L}_{\text{task}}(\hat{\mathbf{X}})$, where $\hat{\mathbf{X}}$ is the reconstructed motion. The optimization is performed by back-propagating through the full denoising chain with a learning rate η:

$$\boldsymbol{\epsilon}_0 \;\leftarrow\; \boldsymbol{\epsilon}_0 - \eta \nabla_{\boldsymbol{\epsilon}_0} \mathcal{L}_{\text{task}}(\hat{\mathbf{X}}),$$

Applications We adopt this optimization strategy for downstream motion-related tasks introduced in [15] by defining task-specific objectives $\mathcal{L}_{\text{task}}$:

- **Motion Refinement** Given a noisy motion $\mathbf{X}^*$, motion refinement projects $\mathbf{X}^*$ to the closest clean motion in the prior by minimizing the distance between noisy motion and the generated motion:

$$\mathcal{L}_{\text{refine}} = \frac{1}{LT} \sum_{t=1}^{T} \sum_{j=1}^{J} \left\| \hat{\mathbf{X}}_{t,j} - \mathbf{X}^*_{t,j} \right\|_2^2,$$

- **Obstacle Collision Avoidance** Considering an obstacle $\mathcal{O}$, let $sdf(\mathbf{p}, \mathcal{O})$ denote the signed distance field from a joint position $\mathbf{p}$ to the obstacle surface. Obstacle collision avoidance task penalizes the collision with the object:

$$\mathcal{L}_{\text{coll}} = \frac{1}{T} \sum_{t=1}^{T} \Big[\min\big(0,\, sdf(\hat{\mathbf{X}}_t, \mathcal{O})\big) \Big]^2,$$

- **Motion Editing** Given a set key-frames $\mathcal{K} = \{(t_k, \mathbf{X}^*_{t_k})\}$, motion editing task edits the input motion at the supervised frames:

$$\mathcal{L}_{\text{edit}} = \frac{1}{|\mathcal{K}|} \sum_{(t_k, \mathbf{X}^*_{t_k}) \in \mathcal{K}} \left\| \hat{\mathbf{X}}_{t_k} - \mathbf{X}^*_{t_k} \right\|_2^2,$$

- **Motion Blending** Motion blending task is to interpolate smooth transition between two motion clips $\mathbf{X}^A$ and $\mathbf{X}^B$ while preserving the start of motion $\mathbf{X}^A$ and the end of motion $\mathbf{X}^B$:

$$\mathcal{L}_{\text{blend}} = \left\| \hat{\mathbf{X}}_{start} - \mathbf{X}^A_{\text{start}} \right\|_2^2 + \left\| \hat{\mathbf{X}}_{end} - \mathbf{X}^B_{\text{end}} \right\|_2^2,$$

Each of these objectives is differentiable, enabling end-to-end optimization of the initial noise $\boldsymbol{\epsilon}_0$. The optimization procedure effectively adapts the generated motion to satisfy external constraints while preserving the motion semantic through the strong motion prior imposed by the diffusion model.

4. Experiments

We first evaluate our approach on the standard text-to-motion task in Sec. 4.1. Then, we assess the learned model

Table 1. **Performance of Text-to-Motion Generation** on the HumanML3D [6] dataset. R@3 denotes R-Precision (Top 3).

Methods	Optimization Support	FID ↑	R@3 ↓	MM Dist ↓	Diversity →
Real		0.002	0.797	2.974	9.503
Momask [10]	✗	0.045	0.807	2.958	-
MMM [32]	✗	0.089	0.804	2.926	9.577
BAMM [31]	✗	0.055	0.808	2.936	9.636
T2M [7]	✓	1.067	0.740	3.340	9.188
MDM [41]	✓	0.544	0.611	5.566	9.559
MLD [2]	✓	0.473	0.773	3.196	9.724
EMDM [49]	✓	0.112	0.786	3.110	9.551
LaMP(Res)	✓	0.469	0.700	3.609	9.317
LaMP(ViT)	✓	0.292	0.752	3.293	9.465

as a motion prior across various downstream applications in Sec. 4.2, demonstrating its effectiveness as a prior in optimization. For each task, we compare against state-of-the-art baselines in terms of motion quality, controllability, accuracy, and diversity. Additionally, we conduct ablation studies to show the impact of each design decision in Sec. 4.3. We also train a ResNet-based variant of LaMP; in what follows, we refer to it as LaMP(Res), and we use LaMP(ViT) for the model trained with a ViT architecture.

4.1. Text-to-Motion Generation

We conduct experiments in text-to-motion generation to show that the LaMP feature improves motion generation by learning a better motion representation. Our study utilizes HumanML3D [6] for text-to-motion generation experiment.

Dataset HumanML3D [6] encompasses 14,616 unique human motion sequences from AMASS [23] and HumanAct12 [5] datasets with 44,970 individual text annotations. For uniformity, all motion sequences HumanML3D are padded to a length of 256 frames for training.

Metrics (1) *Fréchet Inception Distance (FID)* measures the distributional gap between real and generated motions by computing the Fréchet distance between their feature embeddings extracted with a pretrained model [8]. (2) *R-Precision@3*. For each generated sequence, we compare its embedding to 32 candidate texts (1 ground-truth + 31 distractors) using Euclidean distance and report whether the correct text appears in the top-3. (3) *Diversity*. Assesses coverage by averaging pairwise embedding distances over randomly paired generations across the test prompts. (4)*MM-Dist* measures intra-prompt diversity by averaging pairwise embedding distances among multiple samples generated from the same input text.

Results As shown in Tab. 1, our method achieves comparable performance with state-of-the-art diffusion models. The VQVAE-based methods [10, 31, 32] obtain the best scores in terms of FID, R-Precision, and Diversity, but they lack the ability to optimize for the target objectives. In the following experiments, we will demonstrate that although some diffusion-based baselines [49] achieve higher scores on text-to-motion metrics, their latent distributions are less robust for optimization, leading to failure in motion-related downstream tasks.

4.2. Optimization Based Motion Tasks

4.2.1 Motion Refinement

Table 2. Performance Metrics for Motion Refinement

Action	FID ↓ (gen)	Dist ↓ (gen/gt)	FID ↓ (gen/gt)	Foot ↓ skating ratio
MLD [2]	4.86	0.04	4.55	3.44
MDM-DNO [15]	2.06	0.02	2.65	**0.12**
EMDM [49]	22.89	0.07	23.19	0.68
LaMP(Res)	<u>1.01</u>	**0.01**	<u>1.43</u>	**0.12**
LaMP(ViT)	**0.72**	**0.01**	**0.91**	**0.12**

Given a noisy input motion, motion refinement aims to refine the noisy motion to be more realistic by projecting the input motion to the motion prior. The optimization-based approach addresses this by initializing the latent with a random $x_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, setting the target objective to the noisy input, and optimizing $\mathcal{L}_{\text{refine}}$.

In this experiment, we compare the learned latent motion prior of LaMP representation with baselines across MDM [49], MLD [2], and EMDM [49]. The original DNO [15] is built on the MDM diffusion model. Therefore, we labeled them in the same column. We sample 1,024 motions from the HumanML3D test dataset and add three types of noise—Gaussian, Perlin, and Sinusoidal—to each body joint as the noisy joint input and see the refined motion from each model.

Metrics We evaluate each optimized motion from two perspectives: (1) Realism with respect to the training data assessed via the *Foot skating ratio* and the *FID (gen) score*. The **Foot skating ratio** measures the incoherence between the human motion by calculating the proportion of frames in which a foot skates for more than a certain distance (2.5 cm) which contact with the ground (foot height $\leq$ 5 cm). **FID (gen)** computes the distance between the refined motions feature distribution and the ground truth motions feature distribution. (2) Fidelity to the original noisy input assessed by measuring the MSE distance of latent representations between the refined and the original motions (reported as **dist (gen/gt)** in the table), as well as the **FID (gen/gt)**, which is computed as the difference between the distribution of the refined motions and the original motions.

Results As shown in Tab. 2, LaMP achieves the best performance on the test data for both *foot-skating ratio* and *FID (gen)*. These results show that input noisy motions are

refined by our model into motions which are more close to the ground truth distribution. Moreover, LaMP (ViT) attains the lowest *dist(gen/gt)* and *FID(gen/gt)* scores, demonstrating that the refined motions maintain high fidelity to the original inputs.

4.2.2 Obstacle Collision Avoidance

We further evaluated our method in obstacle collision avoidance. In this task, the input motion must navigate around obstacles while preserving the original semantics. For the experiment, we randomly sampled 1,024 motion sequences from the test dataset and placed eight spherical obstacles with a radius of 30 cm around the initial position of each motion. The objective function is to minimize $\mathcal{L}_{\text{coll}}$ such that collisions with obstacles are avoided. Since this task needs to obtain the initial noise of the input motion, while EMDM [49] doesn't provide a solution to obtain the initial noise from the given motion, we do not include it as a baseline here.

We assess performance using four metrics: **FID(gen)**, **dist(gen/gt)**, **FID(gen/gt)**, and **collision loss**. The *collision loss* is defined as the mean distance between the character's feet and the obstacles in frames where collisions occur, normalized by the sequence length. Other metrics are mentioned before.

Table 3. Performance Metrics for SDF Collision Avoidance

Model	FID ↓ (gen)	Dist ↓ (gen/gt)	FID ↓ (gen/gt)	Collision ↓ loss(m)
MLD [2]	2.19	0.02	1.07	2.28
MDM-DNO [15]	1.47	**0.00**	0.55	**0.00**
LaMP(Res)	**1.24**	0.01	0.52	**0.00**
LaMP(ViT)	1.36	**0.00**	**0.35**	**0.00**

Results As shown in Tab. 3, LaMP with ViT architecture achieves the lowest *collision loss*, demonstrating the capability to guide avoiding obstacle collisions based on environmental feedback. At the same time, it maintains a low *FID (gen)*, indicating that the optimized motions remain realistic. It also reaches the lowest *dist (gen/gt)* and *FID (gen/gt)*, further confirming that doing optimization on the learned motion prior closely preserves the characteristics of the original sequences. We can also find that MDM [49] provides worse but reasonable results. However, MLD [2] model fails to avoid the collision since its compressed latent representation loses the details of the motion.

4.2.3 Motion Editing

We then conduct experiments on motion editing task. For evaluation, we sampled 1,024 motions from the HumanML3D test set. Given an input motion, we randomly selected a target pelvis position within a range of -3 m to $+3$ m relative to the initial position and edited the motion to align with the target. To further analyze semantic preservation in edited motions, we followed prior work [15] and examined four representative motion types—*jumping*, *long jumping*, *walking with raised hands*, and *crawling*—to provide a detailed case study of editing performance.

Table 4. Performance Metrics for Motion Editing

Model	FID ↓ (gen)	Dist ↓ (gen/gt)	Foot ↓ skating ratio	Objective ↓ Error(m)
MLD [2]	13.07	0.08	12.55	0.46
MDM-DNO [15]	7.12	0.03	**0.10**	0.01
LaMP(Res)	5.58	0.03	0.11	**0.00**
LaMP(ViT)	**1.35**	**0.00**	**0.10**	0.01

Table 5. Performance Metrics for Motion Editing

Action	Content ↑ Preserve	Objective ↓ Error (m)	Foot ↓ skating ratio	Jitter ↓
"jumping"				
Input	1.00	1.59	0.00	0.23
MLD [2]	0.95	0.16	0.26	0.75
MDM-DNO [15]	0.85	**0.00**	**0.01**	0.67
LaMP(Res)	0.87	**0.00**	0.04	**0.45**
LaMP(ViT)	**0.98**	**0.00**	0.02	0.65
"doing a long jump"				
Input	1.00	1.99	0.00	0.69
MLD	0.90	**0.00**	0.14	**0.78**
MDM-DNO	0.59	**0.00**	**0.01**	1.20
LaMP(Res)	0.78	**0.00**	0.04	1.45
LaMP(ViT)	**0.96**	**0.00**	0.04	1.05
"walking with raised hand"				
Input	1.00	1.95	0.00	0.21
MLD	0.76	0.24	0.13	0.76
MDM-DNO	**0.98**	**0.00**	**0.04**	**0.41**
LaMP(Res)	0.90	**0.00**	**0.04**	0.44
LaMP(ViT)	**0.98**	**0.00**	**0.04**	0.72
"crawling"				
Input	1.00	1.67	0.01	0.45
MLD	0.97	0.20	0.02	1.00
MDM-DNO	0.93	**0.00**	**0.03**	**0.67**
LaMP(Res)	0.96	**0.00**	0.09	1.89
LaMP(ViT)	**0.97**	**0.00**	0.05	1.26

We evaluate motion editing with four metrics: **FID (gen)**, **dist (gen/gt)**, **foot-skating ratio**, and **objective error**. The objective error is the Euclidean distance (in meters) between the edited motion's pelvis (root) position and the target position at the specified frame.

For the four selected motion types, we additionally report the *semantic preservation ratio*—the proportion of frames whose semantic label matches that of the target motion. We therefore omit *FID (gen)* in this setting since computing the distribution distance between the test set and a single class of motion is unreliable. Instead, we report the *Jittering Ratio*, defined as the mean per-joint change in acceleration (jerk) over time, reported in $10^2\,\mathrm{m/s^3}$.

Results As shown in Tab. 4, our method achieves the best performance. In contrast, the MLD [2] exhibits the highest objective error, as its latent representation lacks fine-grained control, leading to difficulty in reaching target positions.

From Tab. 5, our method also achieves the highest *semantic preservation ratio* across all four motion types, demonstrating that semantic meaning is preserved during optimization in the latent motion prior. Although the MDM [49]-based editing method produces lower jittering ratios, this improvement comes at the cost of semantic fidelity. For example, in the case of *long jumping*, the MDM latent space fails to preserve the semantic structure of repeated forward jumps, instead producing a smoother but semantically inconsistent motion.

4.2.4 Motion Blending

In this experiment, we evaluate the performance of our method in producing a smooth transition between the two input motions. We randomly sample 1,024 pairs of motions from the HumanML3D test dataset, concatenate each pair to form a single motion sequence, and mask the intermediate frames. The objective is to use the diffusion model to generate a new motion that produces a smooth transition between the two input motions. We evaluate blending performance using **FID(gen)**, **dist(gen/gt)**, **Diversity**, and **Objective Error**. The metrics is the same as the previous sections.

Table 6. Performance Metrics for Motion Blending

Model	FID ↓	Dist ↓ (gen/gt)	Diversity ↑	Objective ↓ Error(m)
MLD	9.04	0.06	**11.69**	3.00
MDM-DNO	3.65	**0.02**	6.65	<u>0.60</u>
EMDM	25.04	0.06	2.51	**0.16**
LaMP(Res)	**2.24**	**0.02**	<u>6.99</u>	0.89
LaMP(ViT)	<u>2.77</u>	**0.02**	6.72	0.87

Results As shown in Tab. 6, LaMP based methods achieve the lowest *FID(gen)*, *dist(gen/gt)*, demonstrating strong fidelity to the original motion, and a high *Diversity* score, showing that the model generates various transitions. In contrast, MDM [49] and EMDM [49] methods achieve lower *Objective Error*; however, their latent spaces fail to produce plausible motions in the unseen portions of the sequence, resulting in higher *FID(gen)* scores.

4.3. Ablation Studies

We conduct ablation studies to analyze the impact of different design choices in our method. The two most critical components are the *part-based encoder–decoder* and the *masking-based training strategy*. As shown in Tab. 7, we systematically remove either the part-wise encoder–decoder, the masking strategy, or both, and evaluate the performance on the motion refinement task. The results demonstrate that the model combining both components consistently achieves the lowest scores in *FID(gen)*, *dist(gen/gt)*, and *FID(gen/gt)*, while also yielding the second-lowest *foot-skating ratio*. These results indicate that the part-wise encoder/decoder learns disentangled motion representations for different body parts, while the masking strategy encourages to learn more informative features from partially observed motion. Together, these components lead to a more effective latent representation of the motion space, which ultimately serves as a robust motion prior for improving performance on the motion-related task.

Table 7. Ablation for Motion Refinement

Part-wise Encoder	Masking Training	FID ↓ (gen)	Dist ↓ (gen/gt)	FID ↓ (gen/gt)	Foot ↓ skating ratio
✗	✗	1.12	<u>0.02</u>	1.52	<u>0.12</u>
✗	✓	<u>0.78</u>	<u>0.02</u>	1.00	**0.11**
✓	✗	1.10	<u>0.02</u>	1.47	<u>0.12</u>
✓	✓	**0.72**	**0.01**	**0.91**	<u>0.12</u>

5. Conclusion and Limitations

Conclusion We present LaMP, a masked motion autoencoder paired with latent diffusion, as a new approach on learning transferable motion priors. The core design, the part-based encoder, and the masking training strategy yield a compact yet expressive latent space that preserves the continuity and semantics of human motion while enabling effective optimization in various downstream applications. In summary, our results indicate that a carefully structured latent space provides a promising foundation for motion priors that scale beyond task-specific supervision. We believe LaMP is a step toward a versatile, data-efficient motion prior that can support a broad range of applications.

Limitations Using the existing text–to–motion datasets, we observe that the data is biased toward common actions, single-person scenes, and curated capture conditions. Consequently, rare skills, complex multi-agent interactions, and contact-rich human–object behaviors are underrepresented in the learned motion prior. Future extensions include modeling human–human and human–object interactions, egocentric motion prediction, tighter scene/physics coupling, and continual or domain-adaptive learning to mitigate dataset bias. We leave these directions to future work.

Acknowledgements We sincerely acknowledge the anonymous reviewers for their insightful feedback. Kaifeng Zhao was supported by the SDSC PhD fellowship. We sincerely thank Siwei Zhang and Yan Wu for the fruitful discussions.

References

[1] Chaitanya Ahuja and Louis-Philippe Morency. Language2pose: Natural language grounded pose forecasting. In *2019 International conference on 3D vision (3DV)*, pages 719–728. IEEE, 2019. 3

[2] Xin Chen, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, and Gang Yu. Executing your commands via motion diffusion in latent space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18000–18010, 2023. 2, 3, 6, 7, 8

[3] Yinglin Duan, Tianyang Shi, Zhengxia Zou, Yenan Lin, Zhehui Qian, Bohan Zhang, and Yi Yuan. Single-shot motion completion with transformer. *arXiv preprint arXiv:2103.00776*, 2021. 2, 3

[4] Katerina Fragkiadaki, Sergey Levine, Panna Felsen, and Jitendra Malik. Recurrent network models for human dynamics. In *Proceedings of the IEEE international conference on computer vision*, pages 4346–4354, 2015. 2

[5] 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 *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2021–2029, 2020. 6

[6] 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] 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. 2, 3, 6

[8] 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. 6

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

[10] 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, 3, 6

[11] Félix G Harvey, Mike Yurick, Derek Nowrouzezahrai, and Christopher Pal. Robust motion in-betweening. *ACM Transactions on Graphics (TOG)*, 39(4):60–1, 2020. 2, 3

[12] Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J Black. Resolving 3d human pose ambiguities with 3d scene constraints. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2282–2292, 2019. 2

[13] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 4

[14] Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. Motiongpt: Human motion as a foreign language. *Advances in Neural Information Processing Systems*, 36:20067–20079, 2023. 2

[15] Korrawe Karunratanakul, Konpat Preechakul, Emre Aksan, Thabo Beeler, Supasorn Suwajanakorn, and Siyu Tang. Optimizing diffusion noise can serve as universal motion priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1334–1345, 2024. 2, 3, 5, 6, 7

[16] F. De la Torre, J. Hodgins, J. Montano, S. Valcarcel, R. Forcada, and J. Macey. Guide to the carnegie mellon university multimodal activity (cmu-mmac) database. In *Tech. report CMU-RI-TR-08-22, Robotics Institute, Carnegie Mellon University*, 2009. 2

[17] Hsin-Ying Lee, Xiaodong Yang, Ming-Yu Liu, Ting-Chun Wang, Yu-Ding Lu, Ming-Hsuan Yang, and Jan Kautz. Dancing to music. *Advances in neural information processing systems*, 32, 2019. 3

[18] Buyu Li, Yongchi Zhao, Shi Zhelun, and Lu Sheng. Danceformer: Music conditioned 3d dance generation with parametric motion transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1272–1279, 2022. 3

[19] Jiaman Li, Ruben Villegas, Duygu Ceylan, Jimei Yang, Zhengfei Kuang, Hao Li, and Yajie Zhao. Task-generic hierarchical human motion prior using vaes. In *2021 International Conference on 3D Vision (3DV)*, pages 771–781. IEEE, 2021. 2, 3

[20] 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 *European Conference on Computer Vision*, pages 426–446. Springer, 2024. 2, 3

[21] Jiefeng Li, Jinkun Cao, Haotian Zhang, Davis Rempe, Jan Kautz, Umar Iqbal, and Ye Yuan. Genmo: A generalist model for human motion. *arXiv preprint arXiv:2505.01425*, 2025. 2, 3

[22] Ruilong Li, Shan Yang, David A Ross, and Angjoo Kanazawa. Ai choreographer: Music conditioned 3d dance generation with aist++. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 13401–13412, 2021. 3

[23] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. Amass: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5442–5451, 2019. 6

[24] Mathieu Marsot, Stefanie Wuhrer, Jean-Sébastien Franco, and Stephane Durocher. A structured latent space for human body motion generation. In *2022 International Conference on 3D Vision (3DV)*, pages 557–566. IEEE, 2022. 3

[25] Julieta Martinez, Michael J Black, and Javier Romero. On human motion prediction using recurrent neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2891–2900, 2017. 2

[26] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3d human pose estimation in the wild using improved cnn supervision. In *3D Vision (3DV), 2017 Fifth International Conference on*. IEEE, 2017. 2

[27] Zichong Meng, Yiming Xie, Xiaogang Peng, Zeyu Han, and Huaizu Jiang. Rethinking diffusion for text-driven human motion generation. *arXiv preprint arXiv:2411.16575*, 2024. 3

[28] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, pages 8162–8171. PMLR, 2021. 5

[29] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10975–10985, 2019. 2

[30] Mathis Petrovich, Michael J Black, and Gül Varol. Action-conditioned 3d human motion synthesis with transformer vae. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10985–10995, 2021. 3

[31] Ekkasit Pinyoanuntapong, Muhammad Usama Saleem, Pu Wang, Minwoo Lee, Srijan Das, and Chen Chen. Bamm: Bidirectional autoregressive motion model. In *European Conference on Computer Vision*, pages 172–190. Springer, 2024. 2, 3, 6

[32] Ekkasit Pinyoanuntapong, Pu Wang, Minwoo Lee, and Chen Chen. Mmm: Generative masked motion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1546–1555, 2024. 2, 3, 6

[33] Ekkasit Pinyoanuntapong, Muhammad Saleem, Korrawe Karunratanakul, Pu Wang, Hongfei Xue, Chen Chen, Chuan Guo, Junli Cao, Jian Ren, and Sergey Tulyakov. Maskcontrol: Spatio-temporal control for masked motion synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9955–9965, 2025. 3

[34] Matthias Plappert, Christian Mandery, and Tamim Asfour. The kit motion-language dataset. *Big data*, 4(4):236–252, 2016. 2

[35] Abhinanda R Punnakkal, Arjun Chandrasekaran, Nikos Athanasiou, Alejandra Quiros-Ramirez, and Michael J Black. Babel: Bodies, action and behavior with english labels. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 722–731, 2021. 2

[36] Davis Rempe, Tolga Birdal, Aaron Hertzmann, Jimei Yang, Srinath Sridhar, and Leonidas J Guibas. Humor: 3d human motion model for robust pose estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11488–11499, 2021. 2, 3

[37] Alla Safonova and Jessica K. Hodgins. Construction and optimal search of interpolated motion graphs. *ACM Trans. Graph.*, 26(3):106–es, 2007. 2

[38] Yonatan Shafir, Guy Tevet, Roy Kapon, and Amit H Bermano. Human motion diffusion as a generative prior. *arXiv preprint arXiv:2303.01418*, 2023. 2, 3

[39] Mingyi Shi, Sebastian Starke, Yuting Ye, Taku Komura, and Jungdam Won. Phasemp: Robust 3d pose estimation via phase-conditioned human motion prior. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14725–14737, 2023. 2, 3

[40] Omid Taheri, Nima Ghorbani, Michael J Black, and Dimitrios Tzionas. Grab: A dataset of whole-body human grasping of objects. In *European conference on computer vision*, pages 581–600. Springer, 2020. 2

[41] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H Bermano. Human motion diffusion model. *arXiv preprint arXiv:2209.14916*, 2022. 2, 3, 6

[42] Matthew Trumble, Andrew Gilbert, Charles Malleson, Adrian Hilton, and John Collomosse. Total capture: 3d human pose estimation fusing video and inertial sensors. In *Proceedings of 28th British Machine Vision Conference*, pages 1–13, 2017. 2

[43] Kengo Uchida, Takashi Shibuya, Yuhta Takida, Naoki Murata, Julian Tanke, Shusuke Takahashi, and Yuki Mitsufuji. Mola: Motion generation and editing with latent diffusion enhanced by adversarial training. *arXiv preprint arXiv:2406.01867*, 2024. 2

[44] Jiachen Xu, Min Wang, Jingyu Gong, Wentao Liu, Chen Qian, Yuan Xie, and Lizhuang Ma. Exploring versatile prior for human motion via motion frequency guidance. In *2021 International Conference on 3D Vision (3DV)*, pages 606–616. IEEE, 2021. 3

[45] Sijie Yan, Zhizhong Li, Yuanjun Xiong, Huahan Yan, and Dahua Lin. Convolutional sequence generation for skeleton-based action synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4394–4402, 2019. 2

[46] Siwei Zhang, Bharat Lal Bhatnagar, Yuanlu Xu, Alexander Winkler, Petr Kadlecek, Siyu Tang, and Federica Bogo. Rohm: Robust human motion reconstruction via diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14606–14617, 2024. 2, 3

[47] Yan Zhang, Michael J Black, and Siyu Tang. Perpetual motion: Generating unbounded human motion. *arXiv preprint arXiv:2007.13886*, 2020. 2

[48] Rui Zhao, Hui Su, and Qiang Ji. Bayesian adversarial human motion synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6225–6234, 2020. 2

[49] 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. 2, 3, 6, 7, 8

[50] Yue Zhu, Nermin Samet, and David Picard. H3wb: Human3. 6m 3d wholebody dataset and benchmark. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 20166–20177, 2023. 2

[51] Tao Zhuo, Zhiyong Cheng, Peng Zhang, Yongkang Wong, and Mohan Kankanhalli. Explainable video action reasoning

via prior knowledge and state transitions. In *Proceedings of the 27th acm international conference on multimedia*, pages 521–529, 2019. 2