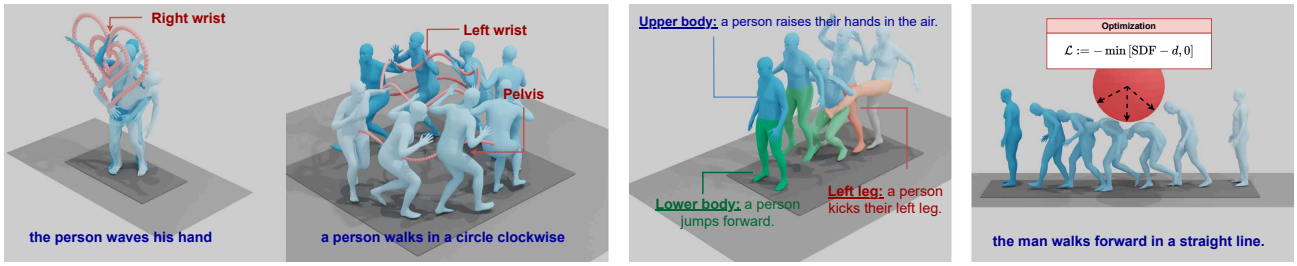
MaskControl: Spatio-Temporal Control for Masked Motion Synthesis

Ekkasit Pinyoanuntapong[1], Muhammad Usama Saleem[1], Korrawe Karunratanakul[2],
Pu Wang[1], Hongfei Xue[1], Chen Chen[3], Chuan Guo[4], Junli Cao[4], Jian Ren[4], Sergey Tulyakov[4]
[1]University of North Carolina at Charlotte [2]ETH Zürich [3]University of Central Florida [4]Snap Inc.
epinyoan@charlotte.edu



(a) Any-Joint-Any-Frame Control (b) Body-Part Timeline Control (c) Zero-shot Objective Control

Figure 1. MaskControl enables a wide range of applications in text-to-motion generation with high quality and precision. (a) Any-Joint Any-Frame Control: Controls specific joints at designated frames. (b) Body-Part Timeline Control: Generates motion from multiple text prompts, each corresponding to different body parts. (c) Arbitrary Objective Control: Supports any control function at inference time.

Abstract

Recent advances in motion diffusion models have enabled spatially controllable text-to-motion generation. However, these models struggle to achieve high-precision control while maintaining high-quality motion generation. To address these challenges, we propose MaskControl, the first approach to introduce controllability to the generative masked motion model. Our approach introduces two key innovations. First, Logits Regularizer implicitly perturbs logits at training time to align the distribution of motion tokens with the controlled joint positions, while regularizing the categorical token prediction to ensure high-fidelity generation. Second, Logit Optimization explicitly optimizes the predicted logits during inference time, directly reshaping the token distribution that forces the generated motion to accurately align with the controlled joint positions. Moreover, we introduce Differentiable Expectation Sampling (DES) to combat the non-differential distribution sampling process encountered by logits regularizer and optimization. Extensive experiments demonstrate that MaskControl outperforms state-of-the-art methods, achieving superior motion quality (FID decreases by 77%) and higher control precision (average error 0.91 vs. 1.08). Additionally, MaskControl enables diverse applications, including any-joint-any-frame control, body-part timeline control, and zero-shot objective control. Video visualization can be found at https://www.ekkasit.com/ControlMM-page/

1. Introduction

Text-driven human motion generation has recently gained significant attention due to the semantic richness and intuitive nature of natural language descriptions. This approach has broad applications in animation, film, virtual/augmented reality (VR/AR), and robotics. While text descriptions offer a wealth of semantic guidance for motion generation, they often fall short in providing precise joint control over specific human joints, such as the pelvis and hands. As a result, achieving natural interaction with the environment and fluid navigation through 3D space remains a challenge.

To tackle this challenge, a few controllable motion generation models have been developed recently to synthesize realistic human movements that align with both text prompts and joint control signals [42, 43, 51, 55]. However, existing solutions face significant difficulties in generating high-fidelity motion with precise and flexible joint control. In particular, current models struggle to support both sparse and dense joint control signals simultaneously. For instance, some models excel at generating natural human movements that traverse sparse waypoints [21, 42], while

others are more effective at synthesizing motions that follow detailed trajectories specifying human positions at each time point [51]. Recent attempts to support both sparse and dense joint inputs encounter issues with control precision; the generated motion often is not aligned well with the control conditions [55]. Besides unsatisfied joint flexibility and accuracy, the quality of motion generation in controllable models remains suboptimal, as evidenced by much worse FID scores compared to models that rely solely on text inputs. Moreover, most current methods utilize motion-space diffusion models, applying diffusion processes directly to raw motion sequences. While this design facilitates the incorporation of joint control signals, the redundancy in raw data introduces computational overhead, resulting in slower motion generation speeds.

Current controllable motion generation methods mostly rely on diffusion model [7, 21, 55, 55]. However, these methods exhibit lower control precision and less generation quality. The current SOTA method, TLControl, [51] achieves high-precision control through test-time optimization with a simple feedforward transformer, but this comes at the cost of lower generation quality, as shown in Tab. 1. Moreover, these methods cannot adapt to arbitrary objective functions at inference time in a zero-shot manner.

To address these challenges, we introduce the first method that integrates joint control into generative masked motion models, enabling simultaneous high-fidelity motion generation and precise control across a broad range of tasks, as illustrated in Fig. 1. In contrast to diffusion-based approaches, masked models generate motion sequences by training a multi-category token classifier and subsequently sampling from the learned categorical distributions conditioned on input signals, such as text and joint trajectories. Building upon this insight, we propose MaskControl—a straightforward yet powerful control mechanism for generative masked motion models—that implicitly and explicitly manipulates the logits of the token classifier to align token distributions closely with input control signals.

Our contributions can be summarized as follows.

- We propose MaskControl, the first approach to introduce controllability to generative masked motion models.
- We introduce two novel control components of MaskControl: (1) *Logits Regularizer* that implicitly perturbs logits at training time to align the distribution of motion tokens with the controlled joint positions. (2) *Logits Optimization* explicitly optimizes the predicted logits during inference time, directly reshaping the token distribution to minimize the residual errors between the generated motion and the target joint position. Moreover, we propose *Differentiable Expectation Sampling (DES)* to overcome the challenge of non-differentiable probabilistic token selection in logits regularization and optimization.
- We show that Logits Optimization can be generalized to solve unseen control tasks in a zero-shot manner.
- We conduct extensive qualitative and quantitative evaluations on multiple tasks, demonstrating that our approach outperforms current SOTA in both motion generation quality and control precision while supporting multiple applications *i.e.* any-joint-any-frame control, body part timeline control, and zero-shot objective control.

Table 1. Comparison of text-conditioned motion generation with joint control signals. Our MaskControl SOTA achieves performance by leveraging Masked Motion Model, demonstrating high-precision control (low Average Error) while maintaining high generation quality (low FID). Previous SOTA methods utilize diffusion in motion space, latent space, or simple feed-forward models. '✓' indicates the ability to control motion using zero-shot objective functions, while '✗' denotes the lack of this capability. '-' signifies control limited to the pelvis.

Method	FID ↓	Average Error (cm) ↓	Base Model	Zero-shot Objective
GMD [21]	0.576	14.39	Motion Diffusion	-
OmniControl [55]	0.218	3.38	Motion Diffusion	✗
MotionLCM [7]	0.531	18.97	Latent Diffusion	✗
TLControl [51]	0.271	1.08	Feed Forward	✗
Ours	**0.061**	**0.98**	**Masked Model**	✓

2. Related Work

Text-driven Motion Generation. Early methods for text-to-motion generation primarily focus on aligning the latent distributions of motion and language, typically by employing loss functions such as Kullback-Leibler (KL) divergence and contrastive losses. Representative works in this domain include Language2Pose [1], TEMOS [37], T2M [15], MotionCLIP [47], and DropTriple [56]. However, the inherent discrepancy between the distribution of text and motion often results in suboptimal generation quality when using these latent space alignment techniques.

Recently, diffusion models have become a widespread choice for text-to-motion generation, operating directly in the motion space [23, 48, 60], VAE latent space [5], or quantized space [24, 31]. In these works, the model gradually denoises the whole motion sequence to generate the output in the reverse diffusion process. Another line of work explores the token-based models in the human motion domain, for example, autoregressive GPTs [14, 20, 58, 61] and *masked motion modeling* [16, 39, 40]. These methods learn to generate discrete motion token sequences that are obtained from a pretrained motion VQVAE [12, 53]. While GPT models usually predict the next token from history tokens, masked motion models utilize the bidirectional context to decode the masked motion tokens. By predicting multiple tokens at once, the masked modeling methods can generate motion sequences in as few as 10 steps, achieving state-of-the-art performance on generation quality and efficiency. Despite the performance gains of masked motion

models, supporting spatial controllability in these models remains unexploited. This paper is the first work that proposes controllable masked motion model to simultaneously achieve high-quality motion generation with high-precision spatial control.

Controllable Motion Synthesis. In addition to text prompts, synthesizing motion based on other control signals has also been a topic of interest. Example control modalities include music [26, 27, 29, 44, 45, 49], interacting object [2, 9, 25, 28], tracking sensors [11], scene [17, 52] programmable motion [30], style [62], goal-reaching task [10], and multi-track timeline control [38]. [32, 33, 35, 36, 46, 54, 57] incorporate physics into motion generation. To control the trajectory, PriorMDM [43] finetunes MDM to enable control over the locations of end effectors. CondMDI [6] generates motion in-betweening from arbitrarily placed dense or sparse keyframes. GMD [21] and Trace and Pace [42] incorporates spatial control into the diffusion process by guiding the root joint location. OmniControl [55] extends the control framework to any joint, while MotionLCM [7] applies this control in the latent space, both leveraging ControlNet [59]. DNO [22] introduces an optimization process on the diffusion noise to generate motion that minimizes a differentiable objective function. Recent approaches [18, 51] model each body part separately to achieve fine-grained control but are limited to dense trajectory objectives. The existing frameworks for controllable motion generation predominantly rely on diffusion models; however, these models typically suffer from low-quality motion generation.

3. MaskControl

The objective of MaskControl is to enable controllable text-to-motion generation based on a masked motion model that generates high-precision and high-quality motion. In particular, given a text prompt and an additional joint control signal, our goal is to generate a physically plausible human motion sequence that closely aligns with the textual descriptions, while following the joint control conditions, i.e., (x, y, z) positions of each human joint at each frame in the motion sequence. Towards this goal, in Sec. 3.1, we first introduce the background of conditional motion synthesis based on the generative masked motion model. We then describe two key components of MaskControl, including *Logits Regularizer* in Sec. 3.2 and inference-time *Logits Optimization* in Sec. 3.3. The first component aims to learn the categorical distribution of motion tokens, conditioned on joint control signals during training time. The second component aims to improve control precision by optimally modifying learned motion distribution via *Logits Optimization* during inference time. Lastly, we introduce *Differentiable Expectation Sampling* to overcome non-differentiability of categorical sampling during Logits Regularizer and Optimization.

3.1. Preliminary: Masked Motion Model

Generative masked motion models [16, 39, 40] generally consist of two stages: *Motion Tokenizer* and *Text-conditioned Masked Transformer*. The objective of the Motion Tokenizer is to learn a discrete representation of motion by quantizing the encoder's output embedding z into a codebook $\mathcal{C}$. For a given motion sequence $\mathcal{P} = [p_1, p_2, ..., p_F]$, where each frame p represents a 3D pose, *Motion Tokenizer outputs* a discrete motion tokens $X = [x_1, x_2, ..., x_L]$. Specifically, the encoder compresses $\mathcal{P}$ into a latent embedding $z \in \mathbb{R}^{t \times d}$ with a downsampling rate of F/L. The embedding z is quantized into codes $c \in \mathcal{C}$ from the codebook $\mathcal{C} = \{c_k\}_{k=1}^{K}$, which contains K codes. The nearest code is selected by minimizing the Euclidean distance between z and the codebook entries, computed as $\hat{z} = \text{argmin}_j \|\mathbf{z} - c_j\|_2^2$. The vector quantization loss L_{VQ} is defined as:

$$L_{VQ} = \|\,\text{sg}(\mathbf{z}) - \mathbf{c_j}\|_2^2 + \beta\|\mathbf{z} - \text{sg}(\mathbf{c_j})\|_2^2, \qquad (1)$$

where $\text{sg}(\cdot)$ is the stop-gradient operator and β is a hyperparameter for commitment loss.

During the second stage, the quantized motion token sequence $X = [x_1, x_2, ..., x_L]$ is updated with `[MASK]` tokens to form the corrupted motion sequence $X_{\overline{\mathbf{M}}}$. This corrupted sequence along with text embedding W are fed into a text-conditioned masked transformer parameterized by θ to reconstruct input motion token sequence with reconstruction probability equal to $p_\theta\left(x_i \mid X_{\overline{\mathbf{M}}}, W\right)$, which is obtained by the motion token classifier. The objective is to minimize the negative log-likelihood of the predicted masked tokens conditioned on text:

$$\mathcal{L}_{\text{mask}} = -\mathop{\mathbb{E}}_{\mathbf{X}\in\mathcal{D}}\left[\sum_{\forall i\in[1,L]} \log p\left(x_i \mid X_{\overline{\mathbf{M}}}, W\right)\right]. \qquad (2)$$

During inference, the transformer masks out the tokens with the least confidence and re-samples these tokens according to their respective distributions $p_\theta\left(x_i \mid X_{\overline{\mathbf{M}}}, W\right)$in the subsequent step. The number of masked tokens n_M is controlled by a masking schedule, a decaying function of the step t. Early steps use a large masking ratio due to high uncertainty, and as the process continues, the ratio decreases as more context is available from previous predictions.

3.2. Logits Regularizer

MaskControl aims to generate a human motion sequence based on the text prompt W and joint control signals S. Towards this goal, we introduce *Logits Regularizer* to conditioned masked transformer, which aims to learn the motion token distribution jointly conditioned both on W and S.

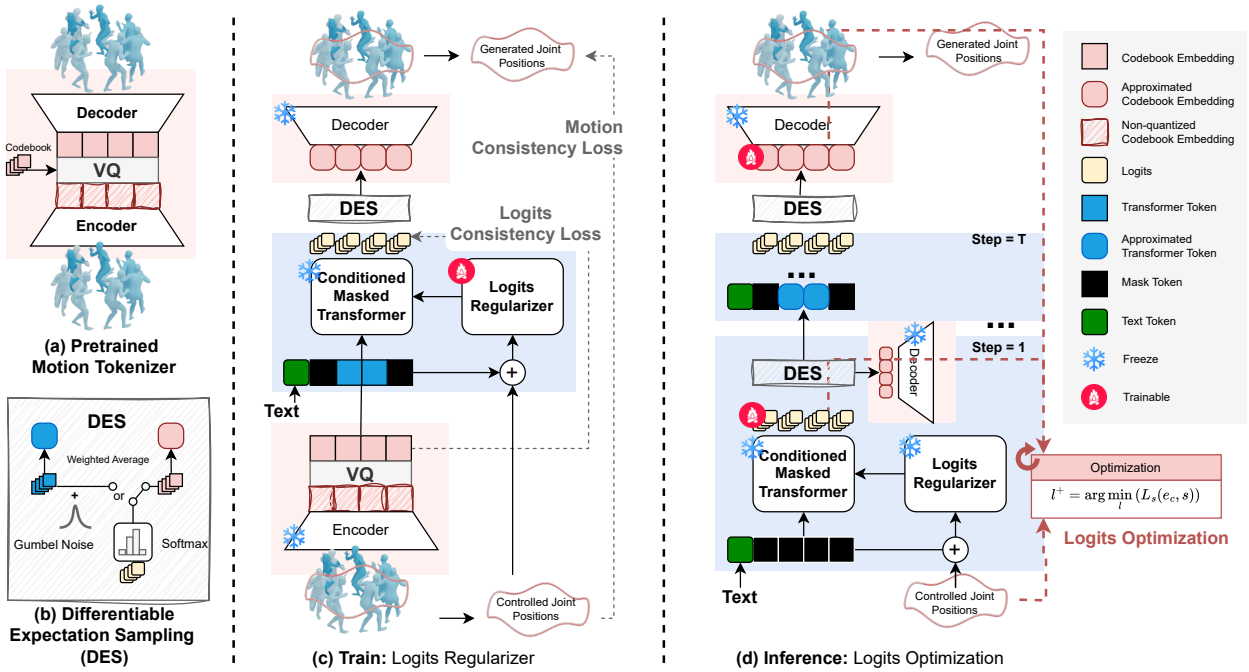


Figure 2. Overall architecture of MaskControl. ***(a) Motion Tokenizer*** transforms the motion sequence into discrete motion tokens. ***(b) Differentiable Expectation Sampling (DES)*** is a differentiable sampling from logits enabling differentiable conversion between discrete tokens in codebook space and transformer token space. ***(c) Training: Logits Regularizer*** ensures high-quality motion by generating embedding closely aligns with joint control signals during an unmasking process. ***(d) Inference: Logits Optimization*** guides logits during the unmasking process at inference time based on the objective function.

Logits Regularizer implicitly alters the output logits of pre-trained text-to-motion model, thus changing the distribution of motion tokens, toward joint control positions.

Model Architecture. Diffusion ControlNet has shown its excellence in adding additional control signals to pre-trained image diffusion model [59]. We demonstrate for the first time that similar design principle can be applied to introduce controllability to generative masked models [3, 4, 8, 50]. Our architecture consists of a pre-trained text-conditioned masked motion model and a *Logits Regularizer*. The pre-trained model provides a strong motion prior based on text prompts, while the *Logits Regularizer* introduces additional joint control signals. Specifically, the *Logits Regularizer* is a trainable replica of the pre-trained masked motion model, as shown in Fig 2. Each Transformer layer in the original model is paired with a corresponding layer in the trainable copy, connected via a zero-initialized linear layer. This initialization ensures that the layers have no effect at the start of training. Unlike the original masked motion model, the *Logits Regularizer* incorporates two conditions: the text prompt W from the pre-trained CLIP model [41] and the joint control signals S. The text prompt W influences the motion tokens through attention, while the joint control signal S is directly added to the motion token sequence via a projection layer.

Joint Control Signal. The conditioned masked transformer is trained to learn the conditional distribution $p_\theta\left(x_k \mid X, X_{\overline{\mathbf{M}}}, W, S\right)$ by reconstructing the masked motion tokens $X_{\overline{\mathbf{M}}}$, conditioned on the unmasked tokens X, text prompt W, and joint control signals S. The joint control condition is a sequence of joint control signals $S = [s_1, s_2, ..., s_F]$ with $s_i \in \mathbb{R}^{j\times 3}$. Each control signal s_i specifies the targeted 3D coordinates of the joints to be controlled, among the total j joints, while joints that are not controlled are zeroed out. Since the semantics of the generated motion are primarily influenced by the textual description, to guarantee the controllability of joint control signals, we extract the joint control signals from the generated motion sequence and directly optimize the consistency loss between input control signals and those extracted from the output. Note that zero-valued in 3D joint coordinates can be ambiguous. Please refer to Sec. A.11 for more details.

Motion Consistency Loss evaluates the alignment between the generated motion and the input joint control signals s:

$$L_s(e_c, s) = \frac{\sum_n \sum_j \sigma_{nj} \odot \|s_{nj} - R(D(e_c))\|}{\sum_n \sum_j \sigma_{nj}}, \quad (3)$$

where σ_{nj} is a binary value indicating whether the joint

control signals s contains a control value at frame n for joint j. The motion tokenizer decoder $D(\cdot)$ converts motion embedding into relative position in local coordinate system and $R(\cdot)$ further transforms the joint's local positions to global absolute locations. The global location of the pelvis at a specific frame can be calculated from the cumulative aggregation of rotations and translations from all previous frames. The locations of the other joints can also be computed by the aggregation of the relative positions of the other joints to the pelvis position.

Logits Consistency Loss extends the objective function of masked motion model in Eq. 2 by applying negative log-likelihood to all positions, including unmasked ones, conditioned on both the text W and joint control signals S.

$$\mathcal{L}_{\text{logits}} = -\sum_{\forall i \in [1,L]} \log p\left(x_i \mid X_{\overline{\mathbf{M}}}, W, S\right). \tag{4}$$

The final loss is the weighted combination *Logits Consistency Loss* and *Motion Consistency Loss*:

$$\mathcal{L} = \alpha \mathcal{L}_{\text{logits}} + (1-\alpha) L_s(e_c, s). \tag{5}$$

3.3. Logits Optimization

The goal of inference-time *Logits Optimization* is to enhance control precision by further reducing the discrepancy between the generated motion and the desired control objectives. This approach does not require pretraining on specific tasks, allowing the model to handle arbitrary objective functions during inference, enabling new control tasks in a zero-shot manner.

The core idea behind *Logits Optimization* is to update the learned logits through gradient-guided optimization during inference, perturbing the motion token distribution. Optimizing logits makes optimization during unmasking process possible since the process requires conversion between transformer token for unmask sampling and codebook embedding for motion reconstruction, as shown in Fig. 2. The optimization process is initialized with the logits obtained from conditioned masked transformer with *Logits Regularizer*, and these logits are iteratively updated to minimize the motion consistency loss L_s.

$$l^{+} = \arg\min_{l} \left(L_s(e_c, s)\right). \tag{6}$$

At each iteration m, the logits l_m are updated using the following gradient-based approach:

$$l_{m+1} = l_m - \eta \nabla_{l_m} L_s(l_m, s), \tag{7}$$

where η controls the magnitude of the updates to the logits, while $L_s(l_m, s)$ represents the gradient of the objective function w.r.t. the logits l_m. This refinement process continues over I iterations. Similarly, in the last unmask step, optimizing embeddings from the codebook space directly is possible. We can directly optimize the embedding:

$$e_{m+1} = e_m - \eta \nabla_{e_m} L_s(e_m, s), \tag{8}$$

where e represents the embedding in the codebook space. For zero-shot objective control, loss function L_s can be replaced with any arbitrary objective function.

3.4. Differentiable Expectation Sampling (DES)

Differentiable Sampling. Both logits regularizer and optimization update the logits by computing the gradient of the disparity between spatial-temporal control signals and generated motion sequence defined in eq. 3. This requires sampling motion tokens according to the categorical token distribution during training/optimization time, which is inherently non-differentiable. To address this, we apply *Straight-Through Gumbel-Softmax Estimator* [19], a reparameterization trick that allows differentiable sampling from a categorical distribution, i.e.,

$$p_\theta\left(x_k \mid X_{\overline{\mathbf{M}}}, W, S\right) = \frac{\exp\left((\ell_k + g_k)/\tau\right)}{\sum_{j=1}^{K} \exp\left(\ell_j + g_j/\tau\right)}, \tag{9}$$

where l is logits, τ refers to temperature, and g represents Gumbel noise with $g_1, \ldots, g_K$ being independent and identically distributed (i.i.d.) samples from a Gumbel$(0,1)$ distribution. The Gumbel$(0,1)$ distribution can be sampled via inverse transform sampling by first drawing $u \sim$ Uniform$(0,1)$ and then computing $g = -\log(-\log(u))$.

Token Expectation. Masked models reconstruct an embedding from the reconstruction probability $p_\theta(x_k \mid \cdot)$ by querying the discrete token embedding using the index with the highest probability, $\text{argmax}_k\, p_\theta(x_k \mid \cdot)$. However, argmax is non-differentiable, and cannot be controlled by perturbation from a control signal. To address this, we relax the quantization constraint by utilizing the weighted average of the codebooks (or transformer tokens) w.r.t. the reconstruction probability $p_\theta(x_k \mid \cdot)$ to approximate continuous tokens. This can be expressed as the expectation of the reconstructed embedding X_{recon}:

$$\mathbb{E}[X_{recon}] = \sum_{k=1}^{K} p_\theta\left(x_k \mid X_{\overline{\mathbf{M}}}, W, S\right) \cdot c_k. \tag{10}$$

4. Applications

Any-Joints-Any-Frame Control. To control specific joints at particular frames, the joint control signal can be directly applied to the desired joint and frame in the global position, as the loss function during training is specifically designed for this task.

Body-Part Timeline Control. MaskControl supports motion generation conditioned on multiple joints, enabling

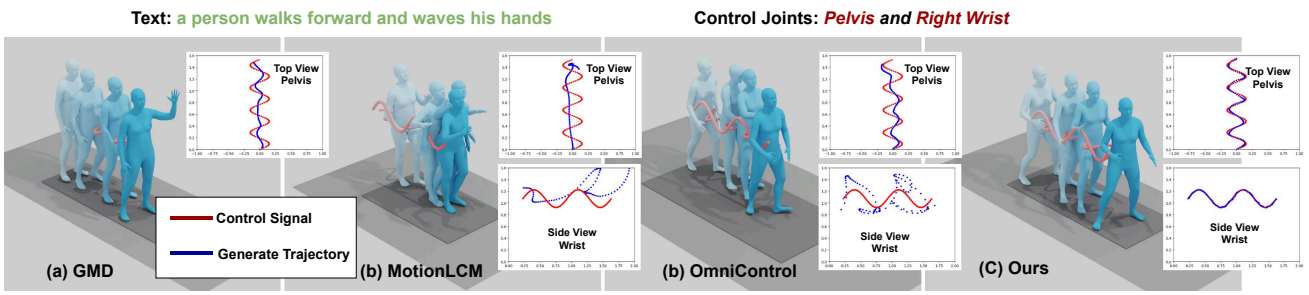


Figure 3. Visualization comparisons to state-of-the-art methods for any-joint any-frame control. The plots on the top display the top view of pelvis control (root trajectory), while the bottom plot shows the side view of the right wrist. Red represents the control signal, and Blue represents the generated joint motion.

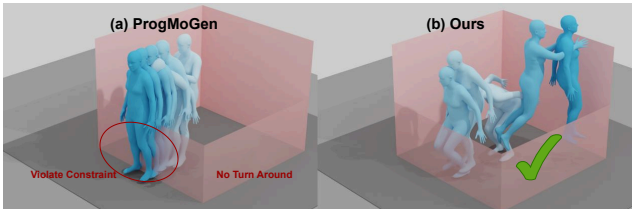


Figure 4. Visualization comparisons to state-of-the-art methods for zero-shot objective control. Objective: constrain a human to walk inside a square area.

control over body parts. To handle multiple prompts corresponding to various body parts and timelines, MaskControl processes each prompt sequentially. Initially, it generates motion without any body part control, then iteratively refines the motion by incorporating prompts conditioned on the specified body parts and timeline constraints from the prior generation. Since MaskControl allows joint control signals to target any joint and frame, partial body or temporal frame control is applicable within this framework. The detail of this process is described in A.12.

Zero-shot Objective Control. In many human motion generation tasks, motion generation requires adaptation to dynamic constraints at inference time, such as human-scene interaction, human-object interaction, or human self-contact. *Logits Optimization* allows optimization with arbitrary loss functions that take joints and frames as inputs, enabling zero-shot adaptation at inference time. However, direct optimization to satisfy arbitrary objective functions can often lead to unrealistic motion. To mitigate this, *Logits Optimization* perturbs the logits during the unmasking process, allowing the masked transformer to re-predict them. By doing so, it ensures that the generated motion remains close to the learned distribution while still achieving the desired control.

5. Experiment

Datasets. We conduct comprehensive experiments on the HumanML3D dataset [15] HumanML3D covers a wide variety motions. It includes 14,616 motion sequences accompanied by 44,970 text descriptions. The textual data contains 5,371 unique words. The motion sequences are sourced from AMASS [34] and HumanAct12 [13].

Evaluation. We follow the evaluation protocol from OmniControl [55] which combines evaluation of quality from HumanML3D[15] and trajectory error from GMD [21].

5.1. Quantitative Comparison to State-of-the-art

GMD [21] only addresses the pelvis location on the ground plane (xz coordinates). To ensure a fair comparison, we follow OmniControl [55] and compare GMD in managing the full 3D position of the pelvis (xyz coordinates). The first section of Tab. 2 resents results for models trained on the pelvis alone to ensure a fair comparison with previous state-of-the-art methods on the HumanML3D [15] dataset. $\rightarrow$ means closer to real data is better. Our model demonstrates significant improvements across all evaluation metrics. When compare to TLControl, the FID score notably decreased from 0.271 to 0.061, the R-Precision increased from 0.779 to 0.809, indicating superior generation quality. In terms of joint control accuracy, both Trajectory Error and Location Error dropped to zero, while the average error decreased to 0.91 cm, indicating highly precise joint control. Furthermore, our model outperforms existing methods in both Diversity and Foot Skating Ratio metrics. In the second section, *Train on All Joints*, we follow the evaluation from OmniControl [55], as our model supports control of any joint, not just the root (pelvis). We train the model to control multiple joints, specifically the pelvis, left foot, right foot, head, left wrist, and right wrist. Our model constantly outperform SOTA in all metrics.

5.2. Qualitative Comparison to State-of-the-art

Any-joint-Any-Frame. We visualize the generated motion using GMD [21], MotionLCM [7] and OmniControl [55] in Fig. 3. The motion is generated based on the prompt "a person walks forward and waves his hands," with the pelvis and right wrist controlled in a zigzag pattern. Since GMD can only control the pelvis, we apply control only to the pelvis

Table 2. Comparison of text-condition motion generation with joint control signal on the HumanML3D. The first section, "Train on Pelvis Only," evaluates our model that was trained solely on the pelvis. The last section, "Train on All Joints", is trained on all joints and reports the average evaluation for each joint. → indicates the closer to the real value, the better.

Method	Joint	R-Precision Top-3 ↑	FID ↓	Diversity →	Foot Skating Ratio ↓	Trajectory Error (>50 cm) (%) ↓	Location Error (>50 cm) (%) ↓	Average Error (cm) ↓
Real	-	0.797	0.002	9.503	-	0.00	0.00	0.00
Train on Pelvis Only								
MDM	Pelvis	0.602	0.698	9.197	0.1019	40.22	30.76	59.59
PriorMDM		0.583	0.475	9.156	0.0897	34.57	21.32	44.17
GMD		0.665	0.576	9.206	0.1009	9.31	3.21	14.39
OmniControl (on pelvis)		0.687	0.218	9.422	0.0547	3.87	0.96	3.38
MotionLCM		0.752	0.531	9.253	-	18.87	7.69	18.97
TLControl		0.779	0.271	9.569	-	0.00	0.00	1.08
Ours (on pelvis)		**0.809**	**0.061 (-77%)**	**9.496**	**0.0547**	**0.00**	**0.00**	**0.98**
Train on All Random Joints								
OmniControl	Average	0.693	0.310	**9.502**	0.0608	6.17	1.07	4.04
TLControl		0.782	0.256	9.719	-	0.00	0.00	1.11
Ours		**0.805**	**0.083 (-68%)**	9.395	**0.0545**	**0.00**	**0.00**	**0.72**

for GMD. However, it fails to follow the zigzag pattern, tending instead to move in a straight line. OmniControl receives control signals for both the pelvis and right wrist. Yet, it not only fails to follow the root trajectory (pelvis) but also does not adhere to the zigzag pattern for the right wrist. In contrast, our MaskControl demonstrates realistic motion with precise joint control for both the pelvis and the right wrist, accurately following the intended zigzag pattern.

Zero-shot Objective Control. Fig. 4 shows the visualization results with the unseen objective function, 'Constrain a human to walk inside a square area,' compared to ProgMoGen [30]. ProgMoGen not only violates the constraint but also generates motion unrelated to the text, as it lacks any 'turns around' motion.

5.3. Body Part Editing

Table 3. Quantitative result of upper body editing task on HumanML3D dataset.

Method	R-precision ↑			FID ↓	MM-Dist ↓	Diversity →
	Top1	Top2	Top3			
MDM [48]	0.298	0.462	0.571	4.827	4.598	7.010
OmniControl [55]	0.374	0.550	0.656	1.213	5.228	9.258
MMM [40]	0.500	0.694	0.798	0.103	2.972	9.254
MotionLCM [7]	0.512	0.685	0.798	0.311	2.948	9.736
Ours	**0.517**	**0.708**	**0.804**	**0.074**	**2.945**	**9.380**

With joint signal controls, our model is capable of conditioning on multiple joints, which can be treated as distinct body parts, while generating the remaining body parts based on text input. In Tab. 3 We quantitatively compare our approach to existing methods designed for this task, including MDM [48] and MMM [40]. Additionally, we compare it with OmniControl [55] and MotionLCM [7], which also supports joint signal control. However, our evaluation demonstrates that OmniControl performs poorly in this task. Following the evaluation protocol from [40], we condition the lower body parts on ground truth for all frames and generate the upper body based on text descriptions using the HumanML3D dataset [15]. Our model is evaluated without retraining, using the same model as in the *Train on All Joints* setup, ensuring a fair comparison with OmniControl, which is trained on a subset of joints. Specifically, we condition only on the pelvis, left foot, and right foot as the lower body signals.

The results show that MDM struggles significantly when conditioned on multiple joints, with the FID score increasing to 4.827. Although OmniControl supports multiple joint control, our experiments reveal that it also suffers under these conditions, with its FID score rising to 1.213. This is consistent with the Cross-Joint evaluation in Tab. 7, which evaluate on multiple joint combination, where OmniControl's FID score deteriorates considerably. MMM performs well in this task but requires retraining with separate codebooks for upper and lower body parts. In contrast, our model outperforms all other methods across all metrics without any retraining. When comparing to the 'Train on Pelvis Only' setup in Tab. 2, our model achieves similar FID and R-Precision scores, highlighting its robustness in handling multiple joint control signals.

5.4. Zero-shot Objective Control

We evaluate three zero-shot objective control tasks using three Human-Scene Interaction evaluations from ProgMoGen [30], a programmable motion control model, *i.e.* 'Head Height Constraint', 'Avoiding Barrier', and 'Walking Inside a Square'. MDM [48] and MoMask [16] serve as baselines without any control mechanisms. Since ProgMoGen builds on MDM as its base model. Similarly MoMask is shown as a baseline for the unconstrained Mask Motion Model. Tab. 4 shows that our MaskControl outperform ProgMoGen in all metrics. **Max Acc.** represents the maximum acceleration of joints which indicates the amount of joint jittering. **Constraint Error** measures how much the head height exceeds the specified constraint. **Unsucc. Rate** represents the rate of motion sequences that violate the height constraint.

Table 4. Comparison of zero-shot objective control. Three *Human-Scene Interaction* objectives are adopted from the programmable motion model (ProgMoGen [30]). Both ProgMoGen and MaskControl are able to control motion during inference by arbitrary loss functions, while MDM and MoMask serve as uncontrollable baseline models.

Task HSI-1: Head Height Constraint

Method	Foot Skate ↓	Max Acc. ↓	Constraint Error ↓	Unsucc. Rate ↓	FID ↓	Diversity →	R-prec. (Top3) ↑
MDM [48] (Unconstrained)	0.086	0.097	0.118	0.718	0.545	9.656	0.610
MoMask [16] (Unconstrained)	0.067	0.072	0.092	0.601	0.354	9.505	0.683
ProgMoGen [30]	0.075	0.094	0.012	0.088	0.556	**9.611**	0.597
Ours	**0.066**	**0.074**	**0.000**	**0.000**	**0.246**	9.393	**0.695**

	Task HSI-2: Avoiding Barrier			Task HSI-3: Walking Inside a Square		
Method	**Foot Skate ↓**	**Max Acceleration ↓**	**Constraint Error ↓**	**Foot Skate ↓**	**Max Acceleration ↓**	**Constraint Error ↓**
MDM [48] (Unconstrained)	0.096	0.126	0.454	0.096	0.126	0.301
MoMask [16] (Unconstrained)	0.072	0.117	0.464	0.072	0.117	0.270
ProgMoGen [30]	0.189	0.150	0.097	0.125	0.093	0.012
Ours	**0.146**	**0.126**	**0.000**	**0.092**	**0.078**	**0.000**

6. Ablation Study

6.1. Component Analysis

Table 5. Ablation results of components analysis and different densities of joint control signal.

Method	R-Prec. Top-3 ↑	FID ↓	Foot Skat. ↓	Traj. Err. ↓	Loc. Err. ↓	Avg. Err. ↓
Component Analysis						
No Control	0.807	0.095	**0.0527**	50.66	35.11	63.18
w/o Logits Regularizer	0.795	0.142	0.0577	0.32	0.02	2.18
w/o Logits Optimization	0.802	0.128	0.0594	39.14	24.00	40.41
Full model	**0.809**	**0.061**	0.0547	**0.00**	**0.00**	**0.98**
Density of Spatial Control Signal						
Density: 1	0.804	0.077	0.0551	0.00	0.00	0.10
Density: 2	0.806	0.087	0.0553	0.00	0.00	0.34
Density: 5	0.811	0.078	0.0553	0.00	0.00	0.98
Density: 25%	0.812	0.055	0.0536	0.01	0.00	1.68
Density: 100%	0.814	0.054	0.0543	0.02	0.00	1.64

To understand how each component impact the quality and joint control error. We conduct an ablation study on each component in sec. *Component Analysis* of Tab. 5, using same evaluation as Tab. 2. Without any control, **No Control**, the model achieves the highest diversity and the lowest Foot Skating Ratio, indicating strong realism in the generated motion. The FID score is also on par. However, all spatial errors are poor due to the absence of joint control components in the model. **Without the Logits Regularizer**, the average error remains relatively low, but the FID score is the worst, highlighting the importance of the *Logits Regularizer* for generation quality. **Without Logits Optimization**, the average error significantly worsens, although the FID remains acceptable, suggesting that the *Logits Regularizer* helps maintain generation quality. In the **Full model**, both the *Logits Regularizer* and *Logits Optimization* complement each other, improving both generation quality (FID) and control accuracy (Average Error).

6.2. Density of Joint Control Signal

In second part of table 5, we provide a detailed analysis of MaskControl's performance across five different joint control density levels, where the model is trained for pelvis control using the HumanML3D dataset. The values 1, 2, and 5 refer to controlling the motion using exactly 1, 2, or 5 frames. The values 25% and 100% indicate the percentage of the total ground-truth motion length for each sample, which can range from 40 to 196 frames depending on the sample. The results show that increasing the joint control improves the quality: the FID score decreases from 0.077 with 1-frame control to 0.054 with full 196-frame (100%) control. Similarly, R-Precision improves from 0.804 at 1-frame density to 0.814 at 196-frame (100%) density. However, the Average Error shows the opposite trend—more joint control leads to higher error, as the model is required to target more specific points.

7. Conclusion

MaskControl is the first model to introduce controllability to the generative masked motion model, enabling precise control while maintaining high-quality motion generation, consistently outperforming diffusion-based controllable frameworks. MaskControl introduces *Differentiable Expectation Sampling (DES)* to relax the quantization constraint, enabling two key innovations: *Logits Regularizer* uses random masking and reconstruction to ensure that the generated motions are of high fidelity, while also reducing inconsistencies between the input control signals and the motions produced. *Inference-Time Logit Optimization* fine-tunes the predicted motion distribution during the unmasking process to the input control signals, enhancing precision and making MaskControl adaptable for unseen tasks. MaskControl has a wide range of applications, including any-joint-any-frame control, body-part timeline control, and zero-shot objective control.

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, 2019. 2

[2] Junuk Cha, Jihyeon Kim, Jae Shin Yoon, and Seungryul Baek. Text2hoi: Text-guided 3d motion generation for hand-object interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1577–1585, 2024. 3

[3] Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T. Freeman. Maskgit: Masked generative image transformer. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11305–11315, 2022. 4

[4] Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, José Lezama, Lu Jiang, Ming Yang, Kevin P. Murphy, William T. Freeman, Michael Rubinstein, Yuanzhen Li, and Dilip Krishnan. Muse: Text-to-image generation via masked generative transformers. *ArXiv*, abs/2301.00704, 2023. 4

[5] 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. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18000–18010, 2022. 2

[6] Setareh Cohan, Guy Tevet, Daniele Reda, Xue Bin Peng, and Michiel van de Panne. Flexible motion in-betweening with diffusion models. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–9, 2024. 3

[7] Wenxun Dai, Ling-Hao Chen, Jingbo Wang, Jinpeng Liu, Bo Dai, and Yansong Tang. Motionlcm: Real-time controllable motion generation via latent consistency model. *arXiv preprint arXiv:2404.19759*, 2024. 2, 3, 6, 7

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*, 2019. 4

[9] Christian Diller and Angela Dai. Cg-hoi: Contact-guided 3d human-object interaction generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19888–19901, 2024. 3

[10] Markos Diomataris, Nikos Athanasiou, Omid Taheri, Xi Wang, Otmar Hilliges, and Michael J. Black. WANDR: Intention-guided human motion generation. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 3

[11] Yuming Du, Robin Kips, Albert Pumarola, Sebastian Starke, Ali Thabet, and Artsiom Sanakoyeu. Avatars grow legs: Generating smooth human motion from sparse tracking inputs with diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 481–490, 2023. 3

[12] Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12868–12878, 2020. 2

[13] Chuan Guo, Xinxin Zuo, Sen Wang, Shihao Zou, Qingyao Sun, Annan Deng, Minglun Gong, and Li Cheng. Action2motion: Conditioned generation of 3d human motions. *Proceedings of the 28th ACM International Conference on Multimedia*, 2020. 6

[14] Chuan Guo, Xinxin Xuo, Sen Wang, and Li Cheng. Tm2t: Stochastic and tokenized modeling for the reciprocal generation of 3d human motions and texts. *ArXiv*, abs/2207.01696, 2022. 2

[15] 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 *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5142–5151, 2022. 2, 6, 7

[16] Chuan Guo, Yuxuan Mu, Muhammad Gohar Javed, Sen Wang, and Li Cheng. Momask: Generative masked modeling of 3d human motions. 2023. 2, 3, 7, 8, 13, 15

[17] Siyuan Huang, Zan Wang, Puhao Li, Baoxiong Jia, Tengyu Liu, Yixin Zhu, Wei Liang, and Song-Chun Zhu. Diffusion-based generation, optimization, and planning in 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16750–16761, 2023. 3

[18] Yiming Huang, Weilin Wan, Yue Yang, Chris Callison-Burch, Mark Yatskar, and Lingjie Liu. Como: Controllable motion generation through language guided pose code editing, 2024. 3

[19] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparametrization with gumble-softmax. In *International Conference on Learning Representations (ICLR 2017)*. OpenReview. net, 2017. 5

[20] Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. Motiongpt: Human motion as a foreign language. *ArXiv*, abs/2306.14795, 2023. 2

[21] Korrawe Karunratanakul, Konpat Preechakul, Supasorn Suwajanakorn, and Siyu Tang. Guided motion diffusion for controllable human motion synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2151–2162, 2023. 1, 2, 3, 6, 13

[22] 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. 3

[23] Jihoon Kim, Jiseob Kim, and Sungjoon Choi. Flame: Free-form language-based motion synthesis & editing. In *AAAI Conference on Artificial Intelligence*, 2022. 2

[24] 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. 2

[25] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 947–957, 2024. 3

[26] Hsin-Ying Lee, Xiaodong Yang, Ming-Yu Liu, Ting-Chun Wang, Yu-Ding Lu, Ming-Hsuan Yang, and Jan Kautz. Dancing to music. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2019. 3

[27] Buyu Li, Yongchi Zhao, Zhelun Shi, and Lu Sheng. Danceformer: Music conditioned 3d dance generation with parametric motion transformer. In *AAAI Conference on Artificial Intelligence*, 2021. 3

[28] Jiaman Li, Alexander Clegg, Roozbeh Mottaghi, Jiajun Wu, Xavier Puig, and C Karen Liu. Controllable human-object interaction synthesis. *arXiv preprint arXiv:2312.03913*, 2023. 3

[29] Ruilong Li, Sha Yang, David A. Ross, and Angjoo Kanazawa. Ai choreographer: Music conditioned 3d dance generation with aist++. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13381–13392, 2021. 3

[30] Hanchao Liu, Xiaohang Zhan, Shaoli Huang, Tai-Jiang Mu, and Ying Shan. Programmable motion generation for open-set motion control tasks. *CVPR*, 2024. 3, 7, 8

[31] Yunhong Lou, Linchao Zhu, Yaxiong Wang, Xiaohan Wang, and Yezhou Yang. Diversemotion: Towards diverse human motion generation via discrete diffusion. *ArXiv*, abs/2309.01372, 2023. 2

[32] Zhengyi Luo, Jinkun Cao, Kris Kitani, Weipeng Xu, et al. Perpetual humanoid control for real-time simulated avatars. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10895–10904, 2023. 3

[33] Zhengyi Luo, Jinkun Cao, Josh Merel, Alexander Winkler, Jing Huang, Kris Kitani, and Weipeng Xu. Universal humanoid motion representations for physics-based control. *arXiv preprint arXiv:2310.04582*, 2023. 3

[34] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. Amass: Archive of motion capture as surface shapes. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5441–5450, 2019. 6

[35] Xue Bin Peng, Ze Ma, P. Abbeel, Sergey Levine, and Angjoo Kanazawa. Amp. *ACM Transactions on Graphics (TOG)*, 40:1 – 20, 2021. 3

[36] Xue Bin Peng, Yunrong Guo, Lina Halper, Sergey Levine, and Sanja Fidler. Ase. *ACM Transactions on Graphics (TOG)*, 41:1 – 17, 2022. 3

[37] Mathis Petrovich, Michael J. Black, and Gül Varol. Temos: Generating diverse human motions from textual descriptions. *ArXiv*, abs/2204.14109, 2022. 2

[38] 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, 12

[39] Ekkasit Pinyoanuntapong, Muhammad Usama Saleem, Pu Wang, Minwoo Lee, Srijan Das, and Chen Chen. Bamm: Bidirectional autoregressive motion model. In *Computer Vision – ECCV 2024*, 2024. 2, 3

[40] 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 (CVPR)*, 2024. 2, 3, 7

[41] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021. 4, 13

[42] Davis Rempe, Zhengyi Luo, Xue Bin Peng, Ye Yuan, Kris Kitani, Karsten Kreis, Sanja Fidler, and Or Litany. Trace and pace: Controllable pedestrian animation via guided trajectory diffusion. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 1, 3

[43] Yonatan Shafir, Guy Tevet, Roy Kapon, and Amit H. Bermano. Human motion diffusion as a generative prior. *ArXiv*, abs/2303.01418, 2023. 1, 3

[44] Lian 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. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11040–11049, 2022. 3

[45] Li Siyao, Weijiang Yu, Tianpei Gu, Chunze Lin, Quan Wang, Chen Qian, Chen Change Loy, and Ziwei Liu. Bailando++: 3d dance gpt with choreographic memory. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–15, 2023. 3

[46] Chen Tessler, Yunrong Guo, Ofir Nabati, Gal Chechik, and Xue Bin Peng. Maskedmimic: Unified physics-based character control through masked motion inpainting. *ACM Transactions on Graphics (TOG)*, 2024. 3

[47] Guy Tevet, Brian Gordon, Amir Hertz, Amit H. Bermano, and Daniel Cohen-Or. Motionclip: Exposing human motion generation to clip space. In *European Conference on Computer Vision*, 2022. 2

[48] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H. Bermano. Human motion diffusion model. *ArXiv*, abs/2209.14916, 2022. 2, 7, 8

[49] Jo-Han Tseng, Rodrigo Castellon, and C. Karen Liu. Edge: Editable dance generation from music. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 448–458, 2022. 3

[50] Ruben Villegas, Mohammad Babaeizadeh, Pieter-Jan Kindermans, Hernan Moraldo, Han Zhang, Mohammad Taghi Saffar, Santiago Castro, Julius Kunze, and D. Erhan. Phenaki: Variable length video generation from open domain textual description. *ArXiv*, abs/2210.02399, 2022. 4

[51] Weilin Wan, Zhiyang Dou, Taku Komura, Wenping Wang, Dinesh Jayaraman, and Lingjie Liu. Tlcontrol: Trajectory and language control for human motion synthesis. *arXiv preprint arXiv:2311.17135*, 2023. 1, 2, 3, 13, 16

[52] Zan Wang, Yixin Chen, Baoxiong Jia, Puhao Li, Jinlu Zhang, Jingze Zhang, Tengyu Liu, Yixin Zhu, Wei Liang, and Siyuan Huang. Move as you say, interact as you can: Language-guided human motion generation with scene affordance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 3

[53] Will Williams, Sam Ringer, Tom Ash, John Hughes, David Macleod, and Jamie Dougherty. Hierarchical quantized autoencoders. *ArXiv*, abs/2002.08111, 2020. 2

[54] Kevin Xie, Tingwu Wang, Umar Iqbal, Yunrong Guo, Sanja Fidler, and Florian Shkurti. Physics-based human motion estimation and synthesis from videos. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11512–11521, 2021. 3

[55] Yiming Xie, Varun Jampani, Lei Zhong, Deqing Sun, and Huaizu Jiang. Omnicontrol: Control any joint at any time for human motion generation. 2023. 1, 2, 3, 6, 7, 13, 19

[56] Sheng Yan, Yang Liu, Haoqiang Wang, Xin Du, Mengyuan Liu, and Hong Liu. Cross-modal retrieval for motion and text via doptriple loss. 2023. 2

[57] Ye Yuan, Jiaming Song, Umar Iqbal, Arash Vahdat, and Jan Kautz. Physdiff: Physics-guided human motion diffusion model. *ArXiv*, abs/2212.02500, 2022. 3

[58] Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Shaoli Huang, Yong Zhang, Hongwei Zhao, Hongtao Lu, and Xiaodong Shen. Generating human motion from textual descriptions with discrete representations. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14730–14740, 2023. 2

[59] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 3, 4

[60] 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*, abs/2208.15001, 2022. 2

[61] Chongyang Zhong, Lei Hu, Zihao Zhang, and Shihong Xia. Attt2m: Text-driven human motion generation with multi-perspective attention mechanism. *ArXiv*, abs/2309.00796, 2023. 2

[62] Lei Zhong, Yiming Xie, Varun Jampani, Deqing Sun, and Huaizu Jiang. Smoodi: Stylized motion diffusion model. *arXiv preprint arXiv:2407.12783*, 2024. 3