ScaleMoGen: Autoregressive Next-Scale Prediction for Human Motion Generation

Inwoo Hwang[1*], Hojun Jang[1*], Bing Zhou[2], Jian Wang[2], Young Min Kim[1†], and Chuan Guo[3†]

[1] Seoul National University
[2] Snap Inc.
[3] Meta Reality Labs
https://inwoohwang.me/ScaleMoGen

Abstract. We present ScaleMoGen, a scale-wise autoregressive framework for text-driven human motion generation. Unlike conventional autoregressive approaches that rely on standard next-token prediction, ScaleMoGen frames motion generation as a coarse-to-fine process. We quantize 3D motions into compositional discrete tokens across multiple skeletal-temporal scales of increasing granularity, learning to generate motion by autoregressively predicting next-scale token maps. To maintain structural integrity, our motion tokenizers and quantizers are explicitly designed so that discrete tokens at every scale strictly preserve the skeletal hierarchy. Additionally, we employ bitwise quantization and prediction, which efficiently scale up the tokenizer vocabulary to preserve motion details and stabilize optimization. Extensive experiments demonstrate that ScaleMoGen achieves state-of-the-art performance, establishing an FID of 0.030 (vs. 0.045 for MoMask) on HumanML3D and a CLIP Score of 0.693 (vs. 0.685 for MoMask++) on the SnapMoGen dataset. Furthermore, we demonstrate that our skeletal-temporal multi-scale representation naturally facilitates training-free, text-guided motion editing.

Keywords: Multi-Scale Motion Modeling · Autoregressive Next-Scale Prediction · Text-Driven Motion Generation

1 Introduction

Text-driven human motion generation has witnessed remarkable progress in recent years. While most existing works—such as VAEs [7, 29, 34] and diffusion models [3, 28, 35, 41, 42]—largely utilize continuous motion representations, an emerging line of work focuses on modeling motion in a discrete latent space. By quantizing motion into discrete tokens, the generation task can be reformulated as a token classification problem. This allows the model to better represent complex, multimodal motion distribution and scales effectively with large-scale datasets [25, 37]. Current discrete methods generally fall into two categories:

[*] Indicates equal contribution
[†] Co-corresponding author

autoregressive models [8, 21, 40] that generate motions through *next-token* prediction, and masked models [4–6, 32, 39] that attend bidirectional context to iteratively fill in missing tokens within a sequence. Recent approaches [11, 31] further combine both paradigms to leverage their complementary strengths. In general, these approaches represent motions as temporally connected short contexts, and the central goal is to grow the token sequences along the time axis.

In this work, we propose a new perspective for generative motion modeling, shifting from *temporal* expansion to *next-scale* prediction, dubbed ScaleMoGen. This design is motivated by the inherent multi-scale structure of 3D human motions: sparse key poses already convey global semantics, while progressively denser frames and joints further add local detail [23,43]. This is also aligned with several multi-scale motion VQ designs, such as DuetGen [4] and MoMask++ [5]. In ScaleMoGen, we explicitly embed this coarse-to-fine hierarchy in both the motion representation and the generative process.

Our approach first learns a hierarchy of skeletal-temporal discrete token maps at multiple scales. A dedicated encoder maps an input motion sequence into a structured skeletal-temporal latent map, which is then progressively quantized in a residual manner into token maps of increasing skeletal-temporal granularity. Token maps across all scales are jointly supervised to reconstruct the source motion. To preserve consistent semantics across scales, we design topology-aware upsampling and downsampling operators in the VQ network and quantizers that respect the skeletal kinematic structure. At each scale, we adopt bit-wise quantization [44], representing each latent dimension as a binary label; this yields an effective vocabulary size of 2^d for an embedding dimension d.

On top of the multi-scale tokenization, a transformer autoregressively predicts the *next-scale* token map conditioned on all coarser scales. Importantly, the prediction target is the binary labels rather than a single categorical index in $[0, 2^d)$, which improves optimization stability. At inference time, generation starts from a 1×1 token map and proceeds through a constant number of scale steps, producing a full motion sequence in a coarse-to-fine manner. Beyond generation, our scale-wise skeletal-temporal model naturally supports text-driven motion editing in a zero-shot manner. Given an existing motion, users can flexibly mask tokens corresponding to specific temporal regions or skeletal parts and re-generate them conditioned on the remaining motion context and text.

Our main contributions are as follows: First, we propose ScaleMoGen, next-scale autoregressive framework for text-driven human motion generation. Second, we introduce a skeletal-temporal multi-scale discrete motion representation that enables structured and semantically consistent generation. Third, ScaleMoGen achieves state-of-the-art motion quality and text–motion alignment on HumanML3D [7] and SnapMoGen [5] benchmarks, and further supports intuitive zero-shot motion editing.

2 Related Works

2.1 Human Motion Generation

Text-conditioned human motion generation has been actively studied for controllable animation. Early approaches relied on continuous motion representations and latent-variable generative models such as VAEs [7,9,29], which enabled smooth motion synthesis but often struggled with long-horizon coherence and semantic alignment. More recently, diffusion-based models [2,3,17–20,28,35,41,42] have emerged as the dominant paradigm, achieving high motion realism and strong text–motion alignment through iterative denoising. SALAD [15] improves generation quality by incorporating structural priors, such as skeleton-aware latent spaces, which better respect the kinematic structure of the human body.

Alongside these advances, an emerging line of work explores discrete motion representations, where continuous human motion is quantized into sequences of discrete tokens and generated using token-based generative models. Under this formulation, motion generation becomes a token classification problem, enhancing scalability and multimodal modeling [21].

Discrete Motion Tokenization Most discrete motion generation methods employ vector quantization or learned codebooks to map continuous pose sequences into discrete tokens [37]. Early approaches adopt a flat, one-dimensional tokenization scheme that encodes full-body motion into temporally ordered token streams, as seen in TM2T [8] and T2M-GPT [40], where all joints and time steps are treated uniformly. To improve reconstruction fidelity, MoMask [6] introduces residual vector quantization with multiple quantization layers. While effective in reducing reconstruction error, this design operates at a fixed full temporal scale, producing tokens with uneven information content and limited semantic interpretability.

More recent works move beyond flat token streams by introducing structured token layouts. MoGenTS [39] organizes motion tokens on a joint–time grid to better preserve spatial correlations across body parts, while DuetGen [4], MoMask++ [5], and MoScale [45] extend earlier designs with multi-scale tokenization along the temporal dimension. Despite these advances, most existing approaches define token structures and scales in a heuristic or temporally driven manner, leaving the correspondence between token granularity and the underlying skeletal-temporal hierarchy of human motion largely implicit.

2.2 Token-based Generative Modeling

Given discrete motion tokens, most existing methods generate motion using either unidirectional autoregressive or masked prediction schemes. Unidirectional autoregressive approaches [8,40] predict motion tokens sequentially in temporal order. This strictly sequential generation makes it difficult to capture bidirectional temporal dependencies, increases inference time, and limits overall expressiveness. In contrast, masked prediction methods [5,6,32,39] formulate generation

as iterative token completion, but typically rely on repeated refinement and require many sampling steps at inference to achieve high-quality motion. Moreover, since generation proceeds through local token updates, these methods lack an explicit notion of global-to-local structure, making it difficult to capture global context and long-range motion dependencies. Recent works [11, 31] attempt to unify both paradigms but still operates at a single motion scale.

Recently, advances in visual generative modeling have revisited autoregressive generation from a multi-scale perspective. Visual Autoregressive Modeling (VAR) [10, 36] reformulates autoregression as coarse-to-fine next-scale prediction, replacing raster-scan next-token prediction with scale-wise generation [45]. This paradigm enables high-level structure to be generated first, followed by progressively finer details, improving both efficiency and structural coherence.

2.3 Zero-Shot Motion Editing

Learning-based motion editing has predominantly relied on paired text–motion supervision, which restricts generalization beyond the training distribution and often necessitates additional task-specific training [1,24]. To alleviate this, recent works—drawing inspiration from diffusion-based image editing techniques [14,27] and exploring zero-shot motion editing [15, 22, 43] by modulating diffusion attention maps or leveraging fine-grained spatio-temporal control. However, such methods operate on a single continuous latent and do not explicitly model structural correspondence between the edited content and the original motion.

Motivated by recent next-scale autoregressive image editing approaches [38], our method enables zero-shot motion editing by preserving global motion context through retaining early-level motion codes, while selectively editing fine-grained details at later levels using a target text prompt. Furthermore, our approach allows precise and localized text-guided motion editing via direct replacement of motion codes associated with specific skeletal parts or temporal segments. This design highlights a key advantage of our representation in supporting structured and interpretable motion edits.

3 ScaleMoGen

Our goal is to generate a 3D human pose sequence $\mathbf{m} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ of length N guided by a textual description c, where each pose $\mathbf{x}_t$ is represented as a collection of joint features. ScaleMoGen encodes motions into *multi-scale token maps* that explicitly reflect the *skeletal-temporal hierarchy* of human movement, where coarser scales capture global semantics and finer scales model localized joint dynamics. The model then autoregressively *predicts next-scale token map* conditioned on the accumulated reconstruction from all previous coarser scales, enabling efficient coarse-to-fine generation.

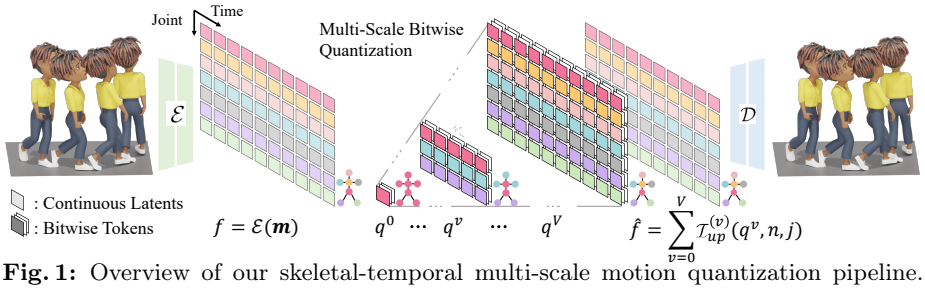

Fig. 1: Overview of our skeletal-temporal multi-scale motion quantization pipeline. Given an input motion sequence $\mathbf{m}$, the encoder $\mathcal{E}$ maps it to a continuous skeletal-temporal latent grid f. The latent is decomposed into a hierarchy of residual components $\{q^v\}_{v=0}^{V}$ via binary multi-scale residual quantization, where each scale has its own temporal resolution and skeletal partition. The quantized residuals are then upsampled and accumulated to form the reconstructed latent $\hat{f}$, which the decoder $\mathcal{D}$ converts back into a full-resolution motion sequence.

3.1 Multi-Scale Skeletal-Temporal Token Map

Our approach first learns a structured skeletal-temporal discrete token map at multiple scales as shown in Fig. 1. We first construct a skeletal-temporal feature map $f = \mathcal{E}(\mathbf{m}) \in \mathbb{R}^{n \times j \times d}$, where n denotes the downsampled temporal length, j the number of atomic skeletal segments, and d the latent feature dimension. We derive this latent space by extending the skeletal-temporal motion encoder $\mathcal{E}(\cdot)$ and decoder $\mathcal{D}(\cdot)$, ensuring that the latent representation captures the fundamental articulated structure of the human body. Specifically, we maintain $j = 7$ articulated atomic skeletal segments—comprising the root, spine, head, and the four extremities (both arms and both legs)—to represent this coarse skeletal structure. Let $\mathcal{J}$ denote the set of the j atomic latent segments.

Skeletal-Temporal Hierarchy To effectively reflect the inherent multi-scale structure of human movements, we define two orthogonal axes of granularity across $(V+1)$ scales, indexed by $v \in \{0, \ldots, V\}$: a sequence of temporal resolutions h^v and a skeletal partition $\mathcal{S}^{(v)}$, where $|\mathcal{S}^{(v)}| = m_v$ denotes the number of semantic segments.

Temporally, we define a sequence of progressively increasing temporal resolutions $\{h^v\}_{v=0}^{V}$ such that $h^0 < \cdots < h^V = n$. This hierarchy ensures that the representation captures temporal motion dynamics ranging from a coarse global context at h^0 to a fine-grained resolution at h^V.

Spatially, to capture structural dependencies, we define a skeletal partition $\mathcal{S}^{(v)} = \{s_1^{(v)}, \ldots, s_{m_v}^{(v)}\}$ that decomposes the atomic joint set $\mathcal{J}$ into m_v semantically coherent segments. For instance, at the coarsest scale ($v = 0$), the partition represents the whole body as a single unit (i.e., $\mathcal{S}^{(0)} = \{\mathcal{J}\}$). At intermediate scales, this decomposes into broader kinematic regions, such as the upper and lower body, to capture regional semantics. Finally, at the finest scale ($v = V$) reaching the atomic resolution where each segment corresponds to a distinct

latent body part (e.g., $\mathcal{S}^{(V)} = \{\{\texttt{root}\}, \{\texttt{spine}\}, \ldots, \{\texttt{right leg}\}\}$). This partitioning scheme satisfies the following properties to effectively preserve consistent semantics across scales:

- Completeness: $\bigcup_{i=1}^{m_v} s_i^{(v)} = \mathcal{J}$ and $s_a^{(v)} \cap s_b^{(v)} = \emptyset$ for $a \neq b$.
- Recursive Refinement: Each segment at a finer scale is a subset of a segment at a coarser scale, i.e., $\forall s \in \mathcal{S}^{(v)}, \exists s' \in \mathcal{S}^{(v-1)}$ s.t. $s \subseteq s'$.

Multi-Scale Residual Token Map Starting from the structured skeletal-temporal continuous feature $f \in \mathbb{R}^{n \times j \times d}$ extracted by the encoder $\mathcal{E}(\cdot)$, this latent f is progressively quantized in a residual manner into $(V+1)$ multi-scale token maps $q^v \in \mathbb{R}^{h^v \times m_v \times d}$. At each level v, the model computes a residual feature r^v in the unified global resolution and quantizes it into a token map q^v with increasing skeletal-temporal granularity.

For any latent feature $z \in \mathbb{R}^d$, we adopt the bit-wise quantization [44] operator $\mathcal{Q}_b(\cdot)$ is defined as:

$$\hat{z} = \mathcal{Q}_b(z) = \frac{1}{\sqrt{d}} \texttt{sign}\left(\frac{z}{|z|}\right) \tag{1}$$

This quantization represents each quantized feature as a sequence of binary codes, resulting in a scalable vocabulary size of 2^d.

At scale level v, the residual feature is first downsampled to the corresponding skeletal-temporal resolution and then quantized as

$$q^v = \mathcal{Q}_b\left(\mathcal{I}_{down}^{(v)}(r^v, h^v, m_v)\right), \quad r^{v+1} = r^v - \mathcal{I}_{up}^{(v)}(q^v, n, j), \quad r^0 = f \tag{2}$$

where $q^v \in \mathbb{R}^{h^v \times m_v \times d}$ denotes the quantized token map at scale level v. Here, $\mathcal{I}_{down}^{(v)}$ and $\mathcal{I}_{up}^{(v)}$ denote the topology-aware downsampling and upsampling operators, respectively, which we describe below. The final structured latent space $\hat{f}$ is reconstructed by accumulating these multi-scale residual contributions:

$$\hat{f} = \sum_{v=0}^{V} \mathcal{I}_{up}^{(v)}(q^v, n, j) \quad \in \mathbb{R}^{n \times j \times d} \tag{3}$$

This additive hierarchical structure allows ScaleMoGen to successively refine motion details, starting from coarse global semantic intent to localized joint dynamics.

Topology-Aware Scaling Each scale token map has a different level of skeletal-temporal granularity, becoming progressively finer across scales. To accurately preserve semantics across scales while maintaining scale-specificed topology, we introduce topology-aware scaling operators, $\mathcal{I}_{down}^{(v)}$ and $\mathcal{I}_{up}^{(v)}$. This operation synchronizes semantic information between the unified global latent resolution (n, j) and each topology granularity (h^v, m_v) at scale v.

Topology-Aware Downsampling ($\mathcal{I}_{down}^{(v)}$): This operator maps the global residual $r^v \in \mathbb{R}^{n\times j\times d}$ to a scale-specific resolution (h^v, m_v). Spatially, for each skeletal segment $s_i^{(v)} \in \mathcal{S}^{(v)}$, $i \in 1, \ldots, m_v$, it aggregates the features of its constituent joints into a single representative latent vector Temporally, it resamples the sequence to h^v temporal resolution. The resulting representation $\mathcal{I}_{down}^{(v)}(r^v, h^v, m_v)$ therefore captures the collective dynamics of each skeletal segment at the specified temporal resolution.

Topology-Aware Upsampling ($\mathcal{I}_{up}^{(v)}$): This operator upscales the quantized $q^v \in \mathbb{R}^{h^v\times m_v\times d}$ back to the global unified target resolution (n, j). Temporally, the sequence is upsampled to n frames, while spatially each segment-level feature $s_i^{(v)}$ is broadcast to all individual joints. The resulting representation is denoted as $\mathcal{I}_{up}^{(v)}(q^v, n, j)$.

Training Multi-Scale Token Map We optimize ScaleMoGen's multi-scale token map using a compound objective. First, the reconstructed motion is obtained via the skeletal-temporal decoder $\hat{\mathbf{m}} = \mathcal{D}(\hat{f})$. The reconstruction loss measures the fidelity against the ground truth $\mathbf{m}$:

$$\mathcal{L}_{\text{rec}} = \|\mathbf{m} - \hat{\mathbf{m}}\|_2^2 \tag{4}$$

To prevent codebook collapse and encourage diverse code usage, we adopt an entropy penalty $\mathcal{L}_{\text{ent}}$. For a latent feature $z \in \mathbb{R}^d$, let $q(z)$ denote the probability distribution over binary codes induced by the quantization operator. The entropy regularization is defined as:

$$\mathcal{L}_{\text{ent}} = \mathbb{E}[H(q(z))] - H[\mathbb{E}[q(z)]] \tag{5}$$

where $H(\cdot)$ denotes the entropy function. The total objective is the weighted sum: $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{rec}} + \lambda\mathcal{L}_{\text{ent}}$.

3.2 Learning Next-Scale Token Map Prediction

Training Next, we employ a transformer p_ϕ to model the coarse-to-fine generation, autoregressively predicting the token map at the next scale conditioned on all coarser scales. We utilize `T5-base` [33] to transform intricate textual prompt c into a sequence of word-level feature embeddings $\mathbf{c}$. At scale k, the model conditions on both the text embedding $\mathbf{c}$ and a structural prefix sum $\hat{f}_k$ constructed from all preceding coarser scale token maps. This prefix aggregates upsampled bit-tokens from earlier levels and aligns them to the current resolution through topology-aware upsampling/downsampling, providing a consistent structural context for fine-scale prediction.

$$\hat{f}_k = \mathcal{I}_{down}^{(k)}\left(\sum_{v=0}^{k-1} \mathcal{I}_{up}^{(v)}(q^v, n, j), h^k, m_k\right) \quad \in \mathbb{R}^{h^k\times m_k\times d} \tag{6}$$

where q^v represents the token map from coarse levels $v \leq k-1$, and the scaling operators $\mathcal{I}$ synchronize these residuals to the current scale's resolution. At the coarsest level ($k = 0$), the word-level embedding $\mathbf{c}$ is projected to $\mathbb{R}^{h^0 \times m_0 \times d}$ and used as the structural prefix $\hat{f}_0$.

The transformer p_ϕ is trained to autoregressively predict the next-scale token map at each scale by maximizing the likelihood of the ground-truth tokens:

$$L_{next} = \sum_{k=0}^{V} -\log p_\phi\big(q^k \mid \hat{f}_k, \mathbf{c}\big). \tag{7}$$

Since each token is a binary code under our quantization scheme, the likelihood decomposes into independent binary classification losses over the bits.

Our transformer backbone consists of stacked blocks for multi-scale feature aggregation. Each block integrates self-attention, cross-attention, and feed-forward layers, together with RoPE2d [13] to capture relative positional relationships over the skeletal–temporal grid. The world-level text embeddings $\mathbf{c}$ are fused via cross-attention, guiding motion synthesis at each scale. To mitigate error propagation across scales, we incorporate a stochastic bit-perturbation strategy during training. With probability r, we randomly flip a subset of bits to simulate prediction errors, followed by re-quantization of the perturbed residual features. This encourages the model to recover from imperfect coarse predictions, improving robustness and enabling self-correcting behavior during generation. We further apply a block-wise causal attention mask to enforce hierarchical dependencies across scales strictly.

Inference At inference, token maps are generated sequentially from coarse to fine scales. For each scale k, the token map is sampled conditioned on the text embedding and the structural prefix made from previously formed token maps:

$$q^k \sim p_\phi(q^k \mid \hat{f}_k, \mathbf{c}), \quad k = 0, \ldots, V. \tag{8}$$

After generating all token maps, the final latent representation is reconstructed as

$$\hat{f} = \sum_{v=0}^{V} \mathcal{I}_{up}^{(v)}(q^v, n, j), \tag{9}$$

which is then decoded to obtain the output motion sequence. An overview of the generation process is illustrated in Fig. 2 (a).

3.3 Zero-Shot Text-Driven Motion Editing

By leveraging our multi-scale contextual discrete representation, which captures hierarchical skeletal structure and temporal variations, our framework supports zero-shot text-driven motion editing. Given a source motion $\mathbf{m}_s$, we quantize it into multi-scale discrete tokens $\{q_s^v\}_{v=0}^V$ and generate target tokens $\{q_t^v\}_{v=0}^V$ that align with a target text c_t while preserving selected attributes from $\mathbf{m}_s$, which

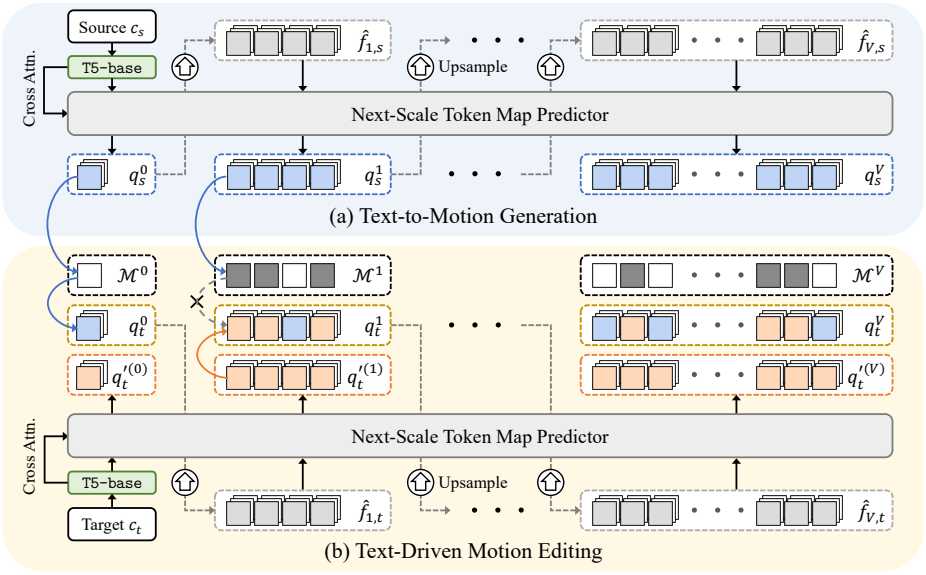

Fig. 2: (a) *Text-to-Motion Generation:* Given a prompt c_s, we autoregressively predict the next-scale token maps $\{q_s^v\}_{v=0}^V$ conditioned on all coarser-scale token maps. (b) *Text-Driven Motion Editing:* With additional target prompt c_t with a source-token preservation mask $\{\mathcal{M}^v\}_{v=0}^V$, we predict edited tokens $q_t'^{(v)}$ conditioned on the remaining source motion context and c_t. The target token maps $\{q_t^v\}_{v=0}^V$ are generated by blending the source tokens q_s^v with the predicted tokens $q_t'^{(v)}$. Note that the 2D skeletal-temporal token maps are flattened into 1D for ease of visualization.

are then decoded into the edited motion $\mathbf{m}_t$. This is achieved using a binary preservation mask $\{\mathcal{M}^v\}_{v=0}^V \in \{0,1\}^{h^v \times m_v}$, where $\mathcal{M}_{t,i}^v = 1$ indicates that a token is retained from the source and 0 indicates it should be resampled based on the c_t. The final mask is constructed according to three distinct criteria:

Global Structure Mask ($\mathcal{M}_{GS}$) : In our multi-scale token maps, coarser scales represent global motion semantics, while finer scales capture localized details. To preserve the global structure of the original motion, we retain tokens up to a specific scale threshold γ:

$$\mathcal{M}_{GS}^v = 1 \quad \text{if} \quad v < \gamma, \quad \text{else} \quad 0.$$

By keeping these coarse tokens, the model ensures that the edited motion $\mathbf{m}_t$ maintains the global movement trajectory of $\mathbf{m}_s$.

Skeletal-Temporal Editing Mask ($\mathcal{M}_{ST}$) : A significant advantage of our method is the explicit organization of the latent space along skeletal and temporal axes. If a user intends to edit a specific set of target joints $\mathcal{J}_e \subseteq \mathcal{J}$ (e.g., the right arm) within a temporal interval $[t_0, t_1]$, the mask is defined as:

$$\mathcal{M}_{ST}^v(t,i) = \begin{cases} 0 & \text{if } s_i^{(v)} \subseteq \mathcal{J}_e \text{ and } t \in [t_0, t_1] \\ 1 & \text{otherwise} \end{cases}$$

This allows for precise, localized editing—such as modifying only the right arm while keeping the rest of the body's trajectory intact.

Semantic-Aware Mask ($\mathcal{M}_{SA}$) : Inspired by [38], we apply a confidence-based replacement where source tokens are resampled if their likelihood under the target distribution $p(\cdot|c_t)$ is below a threshold τ. This selectively updates semantically divergent regions while preserving the original motion's structural continuity.

The final preservation mask is defined as the element-wise sum of the individual masking components: $\mathcal{M}^v = \mathcal{M}^v_{GS} + \mathcal{M}^v_{ST} + \mathcal{M}^v_{SA}$. Since our model generates motion in a scale-wise manner, the target tokens q^v_t at each scale v are synthesized by blending the original attributes with the newly sampled distribution:

$$q^v_t = \mathcal{M}^v \odot q^v_s + (1 - \mathcal{M}^v) \odot q'^{(v)}_t, \tag{10}$$

where $q'^{(v)}_t$ denotes token map sampled from the distribution $p(\cdot|q^{<v}_t, c_t)$. An overview of the editing process is illustrated in Fig. 2 (b).

4 Experiments

In this section, we provide a comprehensive evaluation of ScaleMoGen, including text-to-motion generation (Section 4.1), comprehensive ablation experiments (Section 4.2), and zero-shot motion editing (Section 4.3). Additionally, we show the sampling efficiency of ScaleMoGen compared to the baselines (Section 4.4). Implementation details are provided in the supplementary material.

Datasets We evaluate ScaleMoGen on two benchmarks, HumanML3D [7] and SnapMoGen [5]. HumanML3D is a large-scale dataset built from AMASS [26] and HumanAct12 [9], containing 14,616 motion clips paired with 44,970 single-sentence descriptions, and serves as the standard benchmark for text-to-motion generation. SnapMoGen is a more recent dataset comprising about 20K motion clips paired with 122K expressive and long-form descriptions that average 48 words—roughly four times longer than HumanML3D captions—making it a challenging testbed for fine-grained language conditioning.

Baseline Models We compare ScaleMoGen against a range of state-of-the-art text-to-motion methods covering three distinct training and sampling paradigms. Diffusion-based models such as MDM [35], MLD [3], MotionDiffuse [41], StableMoFusion [16], MARDM [28], and SALAD [15]. Next-token prediction models in an autoregressive manner (T2M [7], TM2T [8], T2M-GPT [40]), generative masked modeling (MMM [32], MoGenTS [39], MoMask [6], and MoMask++ [5]), and next-scale prediction (MoScale [45]).

Evaluation Setup Following standard protocols [5, 7], we report R-Precision, FID, Multimodal Distance (MM Dist), and Multimodality on HumanML3D. For

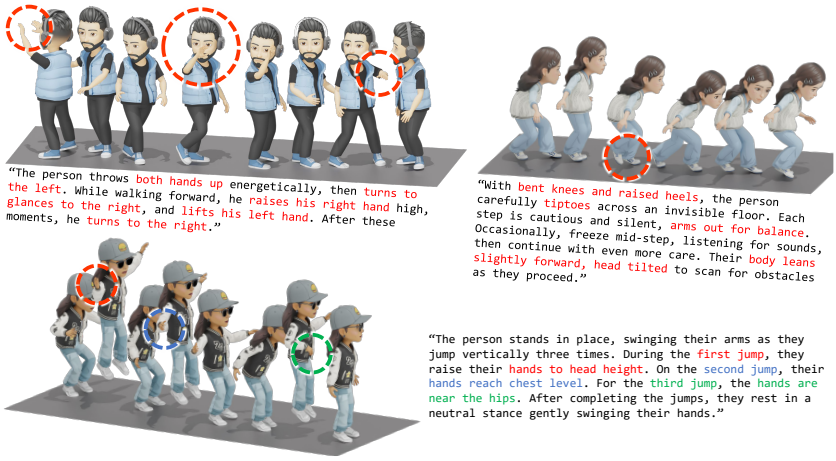

Fig. 3: Qualitative text-to-motion generation results of ScaleMoGen. Given highly descriptive, long-form text prompts, ScaleMoGen accurately synthesizes complex sequences of actions (top-left), fine-grained body-part articulations (top-right), and timely executed motions with precise spatial constraints (bottom).

SnapMoGen, we use a TMR-style retriever [30] and report R-Precision, FID, CLIP Score, and Multimodality. We generate one motion per text, evaluate multiple times, and report the mean with 95% confidence intervals. Real motions are evaluated by the same metrics and reported as an upper bound.

4.1 Text-to-Motion Generation

Tables 1 and 2 show the results on HumanML3D and SnapMoGen. On HumanML3D, ScaleMoGen achieves the best FID and MM Distance. While performing comparably to SALAD [15] in R Precision, ScaleMoGen achieves a substantially lower FID, demonstrating both high fidelity and strong text alignment. On SnapMoGen, ScaleMoGen attains the best R Precision and CLIP Score, outperforming MoMask++ [5] in alignment. Although ScaleMoGen's FID is marginally higher than MoMask++, the improved CLIP Score indicates better adherence to fine-grained descriptions, confirming its generalization to rich, long-form prompts.

Figure 3 visualizes qualitative results on complex prompts. Trained on SnapMoGen, ScaleMoGen excels at interpreting fine-grained skeletal-temporal conditions. For instance, in the top-left example, it seamlessly transitions across distinct sub-actions (throwing hands up, turning, walking). The top-right shows nuance in body-part constraints, capturing a tiptoeing posture with bent knees. The bottom example highlights temporal precision by strictly following chronological instructions to progressively lower hand height across consecutive jumps. These visuals align with our strong quantitative alignment scores. Please refer to our project-page for additional examples with videos.

Table 1: Quantitative results on HumanML3D test set. ± indicates a 95% confidence interval. **Bold** and underline denote the best and second-best results, respectively.

Methods	R Precision↑			FID↓	MM Dist↓	MModality↑
	Top 1	Top 2	Top 3			
TM2T [8]	$0.424^{\pm.003}$	$0.618^{\pm.003}$	$0.729^{\pm.002}$	$1.501^{\pm.017}$	$3.467^{\pm.011}$	$2.424^{\pm.093}$
T2M [7]	$0.455^{\pm.003}$	$0.636^{\pm.003}$	$0.736^{\pm.002}$	$1.087^{\pm.021}$	$3.347^{\pm.008}$	$2.219^{\pm.074}$
MDM [35]	-	-	$0.611^{\pm.007}$	$0.544^{\pm.044}$	$5.566^{\pm.027}$	$\mathbf{2.799}^{\pm.072}$
MLD [3]	$0.481^{\pm.003}$	$0.673^{\pm.003}$	$0.772^{\pm.002}$	$0.473^{\pm.013}$	$3.196^{\pm.010}$	$2.413^{\pm.079}$
MotionDiffuse [41]	$0.491^{\pm.001}$	$0.681^{\pm.001}$	$0.782^{\pm.001}$	$0.630^{\pm.001}$	$3.113^{\pm.001}$	$1.553^{\pm.042}$
T2M-GPT [40]	$0.492^{\pm.003}$	$0.679^{\pm.002}$	$0.775^{\pm.002}$	$0.141^{\pm.005}$	$3.121^{\pm.009}$	$1.831^{\pm.048}$
MMM [32]	$0.515^{\pm.002}$	$0.708^{\pm.002}$	$0.804^{\pm.002}$	$0.089^{\pm.005}$	$2.926^{\pm.007}$	$1.226^{\pm.040}$
MoMask [6]	$0.521^{\pm.002}$	$0.713^{\pm.002}$	$0.807^{\pm.002}$	$0.045^{\pm.002}$	$2.958^{\pm.008}$	$1.241^{\pm.040}$
MoGenTS [39]	$0.529^{\pm.003}$	$0.719^{\pm.002}$	$0.812^{\pm.002}$	$0.033^{\pm.001}$	$2.867^{\pm.006}$	$0.808^{\pm.036}$
SALAD [15]	$\mathbf{0.581}^{\pm.003}$	$0.769^{\pm.003}$	$\mathbf{0.857}^{\pm.002}$	$0.076^{\pm.002}$	$2.649^{\pm.009}$	$1.751^{\pm.062}$
MoMask++ [5]	$0.517^{\pm.002}$	$0.709^{\pm.002}$	$0.803^{\pm.002}$	$0.069^{\pm.003}$	$2.948^{\pm.007}$	$1.192^{\pm.053}$
MoScale [45]	$0.540^{\pm.002}$	$0.727^{\pm.002}$	$0.817^{\pm.002}$	$0.046^{\pm.002}$	$2.830^{\pm.005}$	$0.873^{\pm.044}$
ScaleMoGen	$0.577^{\pm.004}$	$\mathbf{0.769}^{\pm.003}$	$0.856^{\pm.002}$	$\mathbf{0.030}^{\pm.002}$	$\mathbf{2.626}^{\pm.006}$	$1.105^{\pm.057}$

Table 2: Quantitative evaluation on SnapMoGen test set.

Methods	R Precision↑			FID↓	CLIP Score↑	MModality↑
	Top 1	Top 2	Top 3			
Real motions	$0.940^{\pm.001}$	$0.976^{\pm.001}$	$0.985^{\pm.001}$	$0.001^{\pm.000}$	$0.837^{\pm.000}$	-
MDM [35]	$0.503^{\pm.002}$	$0.653^{\pm.002}$	$0.727^{\pm.002}$	$57.783^{\pm.092}$	$0.481^{\pm.001}$	$\mathbf{13.412}^{\pm.231}$
T2M-GPT [40]	$0.618^{\pm.002}$	$0.773^{\pm.002}$	$0.812^{\pm.002}$	$32.629^{\pm.087}$	$0.573^{\pm.001}$	$9.172^{\pm.181}$
StableMoFusion [16]	$0.679^{\pm.002}$	$0.823^{\pm.002}$	$0.888^{\pm.002}$	$27.801^{\pm.063}$	$0.605^{\pm.001}$	$9.064^{\pm.138}$
MARDM [28]	$0.659^{\pm.002}$	$0.812^{\pm.002}$	$0.860^{\pm.002}$	$26.878^{\pm.131}$	$0.602^{\pm.001}$	$9.812^{\pm.287}$
MoMask [6]	$0.777^{\pm.002}$	$0.888^{\pm.002}$	$0.927^{\pm.002}$	$17.404^{\pm.051}$	$0.664^{\pm.001}$	$8.183^{\pm.184}$
MoGenTS [39]	$0.626^{\pm.003}$	$0.771^{\pm.003}$	$0.845^{\pm.002}$	$27.802^{\pm.102}$	$0.586^{\pm.002}$	$7.142^{\pm.191}$
SALAD [15]	$0.742^{\pm.002}$	$0.875^{\pm.003}$	$0.922^{\pm.002}$	$23.248^{\pm.091}$	$0.658^{\pm.002}$	$9.238^{\pm.175}$
MoMask++ [5]	$0.802^{\pm.001}$	$0.905^{\pm.002}$	$0.938^{\pm.001}$	$\mathbf{15.06}^{\pm.065}$	$0.685^{\pm.001}$	$7.259^{\pm.180}$
ScaleMoGen	$\mathbf{0.807}^{\pm.004}$	$\mathbf{0.908}^{\pm.003}$	$\mathbf{0.941}^{\pm.003}$	$16.35^{\pm.086}$	$\mathbf{0.693}^{\pm.001}$	$9.399^{\pm.669}$

4.2 Component Analysis

Effect of Skeletal Topology To validate the importance of structured multi-scale quantization, we ablate the topology constraint by treating motion simply as a monolithic vector sequence rather than a structured hierarchy. As shown in Tab. 3 (A), ScaleMoGen without skeletal topology suffers a severe performance drop across all metrics. This confirms that embedding the inherent human skeletal structure into the latent space is crucial. It not only provides a stronger inductive bias for realistic articulations but also creates a more expressive discrete representation that easily aligns with fine-grained descriptions.

Effect of Code Size Table 3 (B) shows that reducing code size from 24 to 16 bits improves CLIP score but degrades FID, whereas increasing to 32 bits yields the opposite trend. A smaller code dimension reduces the effective code space, aiding the autoregressive predictor in selecting consistent text-conditioned codes (improving CLIP score). However, the limited bit capacity increases quantization

Table 3: Ablation analysis of model configuration on SnapMoGen test set.

	VQ Config.		T2M Config.	T2M Generation		
	Skeleton -Aware	Code Size	# params.	R Precision↑ (Top 3)	FID↓	CLIP Score↑
Base	✓	24-bit	125M	0.941	16.35	0.693
(A) w/o Skeletal Topology	✗			0.902	26.74	0.630
(B) Code Size		16-bit		0.951	19.53	0.705
		32-bit		0.933	15.98	0.683
(C) Larger model			362M	0.934	15.84	0.688
			941M	0.930	16.54	0.686
			2.2B	0.927	16.51	0.682

error, losing fine motion details (worsening FID). Conversely, higher bit capacity eliminates quantization-induced degradation for better FID, but expands the discrete search space, making it harder to predict text-consistent codes. Our base setting (24-bit) provides a good compromise between these two trade-offs.

Effect of Token Map Predictor Model Size We investigate scaling the token map predictor from 125M to 2.2B parameters as in Tab. 3 (C). While motion quality lacks a consistent trend, text-motion alignment degrades with increased capacity. This suggests that blindly scaling the autoregressive model without correspondingly expanding the dataset size leads to overfitting on the training distribution. The over-parameterized model memorizes specific motion sequences rather than learning generalizable text-motion mappings. Unlike LLMs trained on Internet-scale data, current 3D motion datasets have limited volume and diversity. Thus, the 125M model proves more robust in our setting. Full realization of scaling laws [12, 25] will require proportional expansion in future data scales.

VQ Design and Motion Reconstruction We evaluate motion reconstruction metrics (FID, MPJPE) for our vector quantization (VQ) design against several alternative motion quantization and reconstruction schemes in Tab. 4. SALAD [15] uses continuous skeletal latents, while MoGenTS [39] employs 2D residual quantization. We further compare our multi-scale skeletal-temporal bitwise VQ to a categorical codebook and a variant without skeletal topology. Our multi-scale skeletal-temporal bitwise VQ achieves the best reconstruction performance, which directly enables higher-quality motion generation. This demonstrates that combining skeletal structure with a multi-scale bitwise design is critical for preserving motion fidelity and improving generation quality.

4.3 Zero-Shot Text-Driven Motion Editing

Since text-driven motion editing lacks established automatic metrics, we conduct a user study to evaluate editing quality. We compare ScaleMoGen against two baselines: MDM [35] and SALAD [15]. For each method, we first generate a

Table 4: VQ ablation.

Method	FID	MPJPE
SALAD [15]	0.25	9.91
MoGenTS [39]	1.02	15.34
Categorical.	4.83	12.93
w/o Skeletal.	5.60	6.66
MoScale	**0.12**	**2.73**

Table 5: User study on motion editing.

Methods	Preservation	Semantic Align.	Overall Qual.
MDM [35]	$3.75^{\pm 1.35}$	$2.51^{\pm 1.29}$	$2.73^{\pm 1.22}$
SALAD [15]	$3.79^{\pm 1.28}$	$3.10^{\pm 1.45}$	$3.23^{\pm 1.29}$
ScaleMoGen	$\mathbf{4.41}^{\pm 0.78}$	$\mathbf{4.78}^{\pm 0.63}$	$\mathbf{4.65}^{\pm 0.71}$

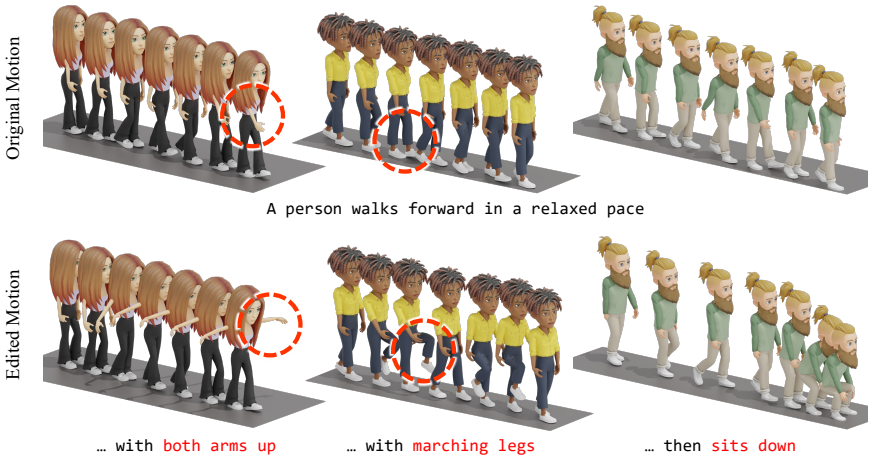

Fig. 4: Qualitative results of text-driven motion editing. Given a source motion and a new target description, ScaleMoGen accurately synthesizes the desired semantic changes, while preserving the identity and unrelated behaviors of the original source motion.

source motion from the source text using its own generation pipeline, then apply the target text to produce the edited motion, yielding 10 editing examples in total. Participants are shown three anonymized results per example and asked to rate each along three criteria on a 5-point scale: (1) *Preservation*—how well the edited motion retains aspects of the original motion unrelated to the edit; (2) *Semantic Alignment*—how accurately the edited motion reflects the target text description; and (3) *Overall Quality*—the naturalness and plausibility of the resulting motion.

As shown in Tab. 5, ScaleMoGen achieves the highest scores across all criteria. Notably, our best preservation score shows that the multi-scale skeletal-temporal discretization enables precise, localized edits while leaving unrelated body parts intact. Additionally, leading semantic alignment confirms our token replacement strategy effectively steers motion toward the target description. These combined strengths yield the highest overall quality, proving ScaleMoGen produces more natural and faithful edits than diffusion baselines. Visual examples in Fig. 4 further demonstrate temporal coherence and fine-grained accuracy. Additional editing examples with videos can be found on our project-page.

Table 6: Sampling efficiency comparison on the SnapMoGen test set.

Methods	Sampling Steps ↓	Inference Time (s) ↓
MDM [35]	1000	10.416
T2M-GPT [40]	40	0.239
SALAD [15]	50	0.519
MoGenTS [39]	18	0.181
MoMask++ [5]	<u>10</u>	**0.051**
ScaleMoGen	**7**	<u>0.071</u>

4.4 Sampling Efficiency

At inference time, the multi-scale token maps are generated autoregressively according to the predefined hierarchy with $(V+1)$ scales, requiring only $(V+1)$ sampling steps starting from the text embedding $\mathbf{c}$. Table 6 compares the number of sampling steps and the average inference time per sample across different methods. We evaluate the computational overhead of each method on a single NVIDIA 4090 GPU. Notably, our method requires the fewest sampling steps, achieving a highly competitive runtime while maintaining superior efficiency.

5 Conclusion

In this work, we presented ScaleMoGen, a novel autoregressive next-scale prediction framework for text-driven human motion generation. In the ScaleMoGen, motion is represented as a multi-scale discrete token map and autoregressively predict next-scale token map conditioned on accumulated coarser-scale token maps. Specifically, our multi-scale design reflects the hierarchical structure of the human skeleton and temporal dynamics, progressively captures global motion semantics and fine-grained joint movements. By shifting the motion generation paradigm into a scale-wise, next-scale prediction process, ScaleMoGen improves text–motion alignment: high-level language governs global motion structure at coarser scales, while finer scales refine localized joint dynamics, resulting in higher overall motion quality. Extensive experiments on HumanML3D and SnapMoGen demonstrate state-of-the-art motion fidelity and text-motion alignment. Furthermore, the explicit correspondence between hierarchical token maps and physical motion components enables intuitive, precise zero-shot motion editing without additional training.

Acknowledgments

This work was partially supported by the National Research Foundation of Korea (NRF) grant (No. RS-2026-25485899) and the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant (RS-2025-25442338, AI Star Fellowship Support Program, Seoul National Univ.) funded by the Korea government (MSIT).

References

1. Athanasiou, N., Ceske, A., Diomataris, M., Black, M.J., Varol, G.: MotionFix: Text-driven 3d human motion editing. In: SIGGRAPH Asia 2024 Conference Papers (2024) 4
2. Bae, J., Hwang, I., Lee, Y.Y., Guo, Z., Liu, J., Ben-Shabat, Y., Kim, Y.M., Kapadia, M.: Less is more: Improving motion diffusion models with sparse keyframes. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 11069–11078 (October 2025) 3
3. Chen, X., Jiang, B., Liu, W., Huang, Z., Fu, B., Chen, T., Yu, G.: Executing your commands via motion diffusion in latent space. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18000–18010 (2023) 1, 3, 10, 12
4. Ghosh, A., Zhou, B., Dabral, R., Wang, J., Golyanik, V., Theobalt, C., Slusallek, P., Guo, C.: Duetgen: Music driven two-person dance generation via hierarchical masked modeling. In: ACM SIGGRAPH (2025) 2, 3
5. Guo, C., Hwang, I., Wang, J., Zhou, B.: Snapmogen: Human motion generation from expressive texts. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025), https://openreview.net/forum?id=pdE9onSn2h 2, 3, 10, 11, 12, 15
6. Guo, C., Mu, Y., Javed, M.G., Wang, S., Cheng, L.: Momask: Generative masked modeling of 3d human motions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1900–1910 (2024) 2, 3, 10, 12
7. Guo, C., Zou, S., Zuo, X., Wang, S., Ji, W., Li, X., Cheng, L.: Generating diverse and natural 3d human motions from text. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5152–5161 (2022) 1, 2, 3, 10, 12
8. Guo, C., Zuo, X., Wang, S., Cheng, L.: Tm2t: Stochastic and tokenized modeling for the reciprocal generation of 3d human motions and texts. In: European Conference on Computer Vision. pp. 580–597. Springer (2022) 2, 3, 10, 12
9. Guo, C., Zuo, X., Wang, S., Zou, S., Sun, Q., Deng, A., Gong, M., Cheng, L.: Action2motion: Conditioned generation of 3d human motions. In: Proceedings of the 28th ACM International Conference on Multimedia. pp. 2021–2029 (2020) 3, 10
10. Han, J., Liu, J., Jiang, Y., Yan, B., Zhang, Y., Yuan, Z., Peng, B., Liu, X.: Infinity: Scaling bitwise autoregressive modeling for high-resolution image synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 15733–15744 (June 2025) 4
11. Han, S.H.K., et al.: Bad: Bidirectional auto-regressive diffusion for text-to-motion generation. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2025) 2, 4
12. Henighan, T., Kaplan, J., Katz, M., Chen, M., Hesse, C., Jackson, J., Jun, H., Brown, T.B., Dhariwal, P., Gray, S., et al.: Scaling laws for autoregressive generative modeling. arXiv preprint arXiv:2010.14701 (2020) 13
13. Heo, B., Park, S., Han, D., Yun, S.: Rotary position embedding for vision transformer. In: European Conference on Computer Vision (ECCV) (2024) 8
14. Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-Or, D.: Prompt-to-prompt image editing with cross attention control (2022) 4
15. Hong, S., Kim, C., Yoon, S., Nam, J., Cha, S., Noh, J.: Salad: Skeleton-aware latent diffusion for text-driven motion generation and editing. In: Proceedings of

the Computer Vision and Pattern Recognition Conference (CVPR). pp. 7158–7168 (June 2025) 3, 4, 10, 11, 12, 13, 14, 15
16. Huang, Y., Yang, H., Luo, C., Wang, Y., Xu, S., Zhang, Z., Zhang, M., Peng, J.: Stablemofusion: Towards robust and efficient diffusion-based motion generation framework. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 224–232 (2024) 10, 12
17. Hwang, I., Bae, J., Lim, D., Kim, Y.M.: Goal-driven human motion synthesis in diverse task. In: Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR) Workshops. pp. 2920–2930 (June 2025) 3
18. Hwang, I., Bae, J., Lim, D., Kim, Y.M.: Motion synthesis with sparse and flexible keyjoint control. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 13203–13213 (October 2025) 3
19. Hwang, I., Lim, D., Jang, H., Kim, Y.M.: Egoforce: Robust online egocentric motion reconstruction via diffusion forcing (2026), https://arxiv.org/abs/2605.13041 3
20. Hwang, I., Zhou, B., Kim, Y.M., Wang, J., Guo, C.: Scenemi: Motion in-betweening for modeling human-scene interaction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 6034–6045 (October 2025) 3
21. Jiang, B., Chen, X., Liu, W., Yu, J., Yu, G., Chen, T.: Motiongpt: Human motion as a foreign language. Advances in Neural Information Processing Systems **36**, 20067–20079 (2023) 2, 3
22. Kim, J., Kim, J., Choi, S.: Flame: Free-form language-based motion synthesis & editing. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 8255–8263 (2023) 4
23. Li, P., Aberman, K., Zhang, Z., Hanocka, R., Sorkine-Hornung, O.: Ganimator: Neural motion synthesis from a single sequence. ACM Transactions on Graphics (TOG) **41**(4), 1–12 (2022) 2
24. Li, Z., Cheng, K., Ghosh, A., Bhattacharya, U., Gui, L., Bera, A.: Simmotionedit: Text-based human motion editing with motion similarity prediction (2025) 4
25. Lu, S., Wang, J., Lu, Z., Chen, L.H., Dai, W., Dong, J., Dou, Z., Dai, B., Zhang, R.: Scamo: Exploring the scaling law in autoregressive motion generation model. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 27872–27882 (2025) 1, 13
26. Mahmood, N., Ghorbani, N., Troje, N.F., Pons-Moll, G., Black, M.J.: Amass: Archive of motion capture as surface shapes. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 5442–5451 (2019) 10
27. Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.Y., Ermon, S.: SDEdit: Guided image synthesis and editing with stochastic differential equations. In: International Conference on Learning Representations (2022) 4
28. Meng, Z., Xie, Y., Peng, X., Han, Z., Jiang, H.: Rethinking diffusion for text-driven human motion generation. arXiv preprint arXiv:2411.16575 (2024) 1, 3, 10, 12
29. Petrovich, M., Black, M.J., Varol, G.: Temos: Generating diverse human motions from textual descriptions. In: European Conference on Computer Vision. pp. 480–497. Springer (2022) 1, 3
30. Petrovich, M., Black, M.J., Varol, G.: Tmr: Text-to-motion retrieval using contrastive 3d human motion synthesis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9488–9497 (2023) 11
31. Pinyoanuntapong, E., Saleem, M.U., Wang, P., Lee, M., Das, S., Chen, C.: Bamm: Bidirectional autoregressive motion model. In: Computer Vision – ECCV 2024 (2024) 2, 4

32. Pinyoanuntapong, E., Wang, P., Lee, M., Chen, C.: Mmm: Generative masked motion model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1546–1555 (2024) 2, 3, 10, 12
33. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research **21**(140), 1–67 (2020), http://jmlr.org/papers/v21/20-074.html 7
34. Sui, K., Ghosh, A., Hwang, I., Wang, J., Guo, C.: A survey on human interaction motion generation (2025), https://arxiv.org/abs/2503.12763 1
35. Tevet, G., Raab, S., Gordon, B., Shafir, Y., Cohen-Or, D., Bermano, A.H.: Human motion diffusion model. arXiv preprint arXiv:2209.14916 (2022) 1, 3, 10, 12, 13, 14, 15
36. Tian, K., Jiang, Y., Yuan, Z., PENG, B., Wang, L.: Visual autoregressive modeling: Scalable image generation via next-scale prediction. In: The Thirty-eighth Annual Conference on Neural Information Processing Systems (2024), https://openreview.net/forum?id=gojL67CfS8 4
37. Van Den Oord, A., Vinyals, O., et al.: Neural discrete representation learning. Advances in neural information processing systems **30** (2017) 1, 3
38. Wang, Y., Guo, L., Li, Z., Huang, J., Wang, P., Wen, B., Wang, J.: Training-free text-guided image editing with visual autoregressive model. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 17577–17586 (October 2025) 4, 10
39. Yuan, W., Shen, W., HE, Y., Dong, Y., Gu, X., Dong, Z., Bo, L., Huang, Q.: Mogents: Motion generation based on spatial-temporal joint modeling. Neural Information Processing Systems (NeurIPS) (2024) 2, 3, 10, 12, 13, 14, 15
40. Zhang, J., Zhang, Y., Cun, X., Huang, S., Zhang, Y., Zhao, H., Lu, H., Shen, X.: T2m-gpt: Generating human motion from textual descriptions with discrete representations. arXiv preprint arXiv:2301.06052 (2023) 2, 3, 10, 12, 15
41. Zhang, M., Cai, Z., Pan, L., Hong, F., Guo, X., Yang, L., Liu, Z.: Motiondiffuse: Text-driven human motion generation with diffusion model. arXiv preprint arXiv:2208.15001 (2022) 1, 3, 10, 12
42. Zhang, M., Guo, X., Pan, L., Cai, Z., Hong, F., Li, H., Yang, L., Liu, Z.: Remodiffuse: Retrieval-augmented motion diffusion model. arXiv preprint arXiv:2304.01116 (2023) 1, 3
43. Zhang, M., Li, H., Cai, Z., Ren, J., Yang, L., Liu, Z.: Finemogen: Fine-grained spatio-temporal motion generation and editing. Advances in Neural Information Processing Systems **36**, 13981–13992 (2023) 2, 4
44. Zhao, Y., Xiong, Y., Krähenbühl, P.: Image and video tokenization with binary spherical quantization. arXiv preprint arXiv:2406.07548 (2024) 2, 6
45. Zheng, Z., Jin, S., Liu, L., Zhao, M.: Next-scale autoregressive models for text-to-motion generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 16376–16386 (June 2026) 3, 4, 10, 12