Language-Guided Transformer Tokenizer for Human Motion Generation

Sheng Yan[1], Yong Wang[2], Xin Du[3], Junsong Yuan[4], and Mengyuan Liu[5†]

[1]Transsion Ltd., [2]Chongqing University of Technology
[3]Afari Intelligence Drive, [4]State University of New York at Buffalo
[5]State Key Laboratory of General Artificial Intelligence, Peking University, Shenzhen Graduate School

Abstract. In this paper, we focus on motion discrete tokenization, which converts raw motion into compact discrete tokens—a process proven crucial for efficient motion generation. In this paradigm, increasing the number of tokens is a common approach to improving motion reconstruction quality, but more tokens make it more difficult for generative models to learn. To maintain high reconstruction quality while reducing generation complexity, we introduce Language-Guided Tokenization (LG-Tok) for efficient motion tokenization. LG-Tok aligns natural language with motion at the tokenization stage, yielding compact, high-level semantic representations. This approach not only strengthens both tokenization and detokenization but also simplifies the learning of generative models. Furthermore, existing tokenizers predominantly adopt convolutional architectures, whose local receptive fields struggle to support global language guidance. To this end, we propose a Transformer-based Tokenizer that leverages attention mechanisms to enable effective alignment. Additionally, we design a language-drop scheme, in which language conditions are randomly removed during training. This scheme prevents shortcut learning over text and enables the detokenizer to support language-free guidance. On three generation benchmarks, LG-Tok outperforms state-of-the-art methods (e.g., achieving an FID score of 0.057 vs. MARDM's 0.114 on HumanML3D). LG-Tok-mini uses only half the tokens while maintaining competitive performance, validating the efficiency of our semantic representations. Code and checkpoints are available at `https://eanson023.github.io/LG-Tok/`.

Keywords: Motion Tokenizer · Human Motion Generation · Multimodal Learning · Text to Motion

1 Introduction

Text-driven human motion generation [4, 19, 20, 26, 36, 48, 61, 88, 89] enables the synthesis of realistic human motions based on natural language descriptions, with widespread applications in game animation, virtual reality, and video motion editing, etc. In recent years, the two-stage paradigm, combining motion tokenization with generative models [23, 52, 68, 75, 80, 84], has demonstrated a

† Corresponding author: `liumengyuan@pku.edu.cn`.

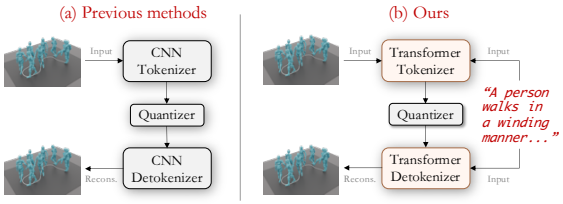


Fig. 1: Comparison between previous CNN-based tokenizers and our Language-Guided Transformer Tokenizer (LG-Tok). Our method aligns language and motion during tokenization, leveraging the transformer's flexibility.

pivotal role in efficiently synthesizing high-fidelity motions. These methods tokenize continuous motion representations into discrete tokens, thereby enabling the direct utilization of well-established generative transformer training and sampling techniques [8, 10, 39, 56, 63] with minimal modifications.

As the core of this paradigm, the properties of motion tokenization significantly influence the performance of generative transformers. Despite various motion tokenization studies focusing on improving training objectives [25, 80] or optimizing quantization [23, 43, 45], current methods still suffer from a fundamental trade-off between reconstruction and generation quality. As shown in Table 1, increasing the number of tokens typically improves motion reconstruction quality (rFID). However, the generation process exhibits a different behavior—more tokens increase learning difficulty, leading to suboptimal results (gFID). Based on this observation, we raise the question: *Can we maintain sufficient tokens for high-quality reconstruction while simultaneously reducing the learning difficulty of token sequences for generative models?*

#Tokens	104	160	236
rFID↓	0.143	0.110	0.090
gFID↓	0.230	0.205	0.257

Table 1: criptsize Performance against #Tokens. rFID/gFID measure reconstruction/generation quality.

Our key insight is that language descriptions naturally provide high-level semantic abstractions—*e.g.*, *"a person walks forward"* encapsulates core motion intent. By introducing language as auxiliary guidance during tokenization, tokens are relieved from redundantly encoding high-level semantics and can instead focus on capturing the fine-grained motion details that language alone cannot fully convey. To this end, we propose Language-Guided Tokenization (LG-Tok), which aligns natural language with motion during tokenization, aiming to provide compact, high-level semantic representations. This approach, typically reserved for the generation phase, is extended to the tokenization stage in our work. It achieves three benefits: *1)* enables latent tokens to learn motion representations while simultaneously absorbing linguistic semantic knowledge during the tokenizing process; *2)* allows language conditions to assist in effectively reconstructing the input motion during the detokenizing process. *3)* simplifies the learning of generative models. For instance, on the Motion-X dataset [42], our compact semantic representations reduce model perplexity from 146.5 to 103.1 and improve FID from 0.257 to 0.088, indicating easier and more effective learning.

Notably, existing tokenizers [23,25,43,80] predominantly adopt convolutional tokenizer-detokenizer architectures, whose local receptive fields struggle to support global language guidance. To address this, we propose a Transformer-based Tokenizer that leverages flexible attention mechanisms to enable effective alignment between language and motion, while enhancing the global contextual awareness of token representations. Specifically, we predefine a learnable sequence of latent tokens that are concatenated with both motion and language representations. Following feature encoding by the transformer-based tokenizer, these latent tokens form semantically informed representations. After vector quantization, a transformer-based detokenizer reconstructs the input motion from the masked token sequences.

Furthermore, we observe that naively incorporating language conditions risks shortcut learning [18], where the tokenizer-detokenizer tends to over-rely on well-trained text embeddings for encoding and reconstruction, undermining the tokens' ability to capture fine-grained motion details. To mitigate this, we propose a language-drop scheme that randomly removes language conditions during training, compelling the model to develop robust token representations independent of linguistic input. Beyond preventing shortcut learning, this scheme further enables language-free guidance decoding during generation. Specifically, drawing on insights from Classifier-Free Guidance [28] for generative models operating in the logit or noise space, our approach extends this principle directly to the *motion space*, providing a complementary guidance mechanism. Fig. 2 shows our LG-Tok outperforming representative baselines. We evaluate LG-Tok on three generation benchmarks: HumanML3D [24], KIT-ML [53], and Motion-X. Across all datasets, LG-Tok consistently surpasses state-of-the-art methods, *e.g.*, achieving a Top-1 R-Precision of **0.542** and FID of **0.057** on HumanML3D, outperforming MARDM [47] (0.500 and 0.114). Meanwhile, LG-Tok-mini, using only half the tokens while maintaining competitive performance (Top-1: 0.521, FID: 0.085 on HumanML3D), validates the efficiency of our semantic representations. In summary, our contributions are as follows:

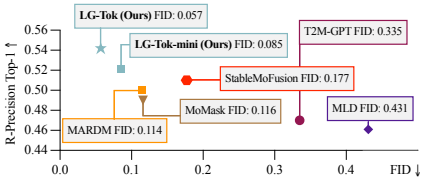


Fig. 2: Generation quality on HumanML3D.

- We propose Language-Guided Tokenization (LG-Tok), which achieves alignment between natural language and motion during the motion tokenization phase.
- We introduce a Transformer-based Tokenizer that leverages attention mechanisms to enable effective language-motion alignment and enhance global contextual awareness of token representations.
- We propose a language-drop scheme that prevents shortcut learning and further enables language-free guidance decoding in the motion space during generation.

2 Related Work

Motion tokenization. Motion tokenizers [23, 25, 26, 34, 49, 51, 52, 71, 72, 80] play a crucial role in efficiently synthesizing high-fidelity motion by reducing computational costs while improving generation quality and efficiency. This has been achieved through autoencoders (AEs) [34, 62]. This general design compresses motion into low-dimensional representations, which are decoded back to the original space. Variational autoencoders (VAEs) [32, 33, 50] extend this paradigm by introducing probability distributions, enabling stochastic sampling and generative capabilities. TEMOS [49] is a representative work, aligning a text encoder's output distribution with the VAE's latent space to enable text-conditioned motion generation. Latent diffusion models [5, 6, 12] further exploit compressed spaces—avoiding the cost of diffusion in the raw motion space. More recently, MARDM [47] performs masked autoregressive diffusion directly on deterministic AE latents, mitigating the sampling noise inherent to VAE representations.

Alternatively, Vector Quantized VAEs (VQ-VAE) [64] insert a quantization step between AE encoders (tokenizers) and decoders (detokenizers), replacing continuous latent distributions with discrete codes that partition the motion space into categorical units. TM2T [25] first applied VQ-VAE to motion, introducing discrete motion tokens. This discrete tokenization facilitates the use of powerful sequence-based generative models [7, 10, 15, 56]. Subsequent variants [59, 80, 84], *e.g.* RQ-VAE [23, 83] and FSQ [45, 68], further refined the quantization and enabled the integration of advanced generative architectures. In this paper, we focus on this discrete (*i.e.*, VQ-based) tokenization pipeline. Based on our observation of the fundamental trade-off between reconstruction and generation in current methods, we propose leveraging language guidance during tokenization to provide compact semantic representations, thereby maintaining high reconstruction quality while reducing the learning difficulty for generative models.

Text-driven motion generation. Natural language provides rich semantic cues for specifying actions, velocities, and directions, making it a key conditioning modality for human motion synthesis [4, 19, 20, 26, 36, 48, 61, 78, 88]. Early GAN-based work, such as Text2Action [1], generated diverse motions from textual descriptions, while JL2P [2] employed GRU-based encoders and decoders to map text to movement. Inspired by advances in text-to-image generation, recent approaches predominantly adopt diffusion or VQ-based paradigms. Diffusion models [30, 31, 44, 66, 81, 82, 86] simulate a forward noising process and train networks to reverse it; MLD [12] improves efficiency by operating in latent space, while PhysDiff [76] enforces physical constraints during generation. VQ-based methods [41, 45, 51, 59, 68, 75, 83–85], *e.g.*, T2M-GPT [80], quantize motion into discrete tokens via VQ-VAE and model them autoregressively using GPT-style [8, 56] transformers. MoMask [23] and MMM [52] adopt MaskGIT-style [10] masked training and non-autoregressive sampling, and MoSa [43] adapts scale-wise autoregressive modeling from the image domain [63]. In this paper, we adopt MoSa as our generative model given its notable gains in quality and efficiency. To

demonstrate the generalizability of our approach, we also apply LG-Tok to MoMask, showing consistent improvements across different generative paradigms.

Image tokenization has emerged as a fundamental technique bridging various vision tasks. This field has diverged into two main branches: understanding-oriented and generation-oriented. Understanding-oriented approaches [3, 38, 46] leverage large language models (LLMs) to create semantic representations for tasks such as classification [16] and segmentation [67]. Meanwhile, generation-oriented methods, such as VQ-GAN [17, 57], emphasize learning latent spaces for detail-preserving compression. These methods typically employ 2D latent grids with fixed downsampling factors [10, 37, 40, 54, 63, 88]. TikTok [74] first introduces a transformer-based 1D tokenizer that generates global tokens as more compact representations of images. Subsequently, numerous variants have refined it [11, 70, 73, 87]. *e.g.*, TxtTok [77] extends this paradigm by incorporating image captions during tokenization to achieve higher compression rates. While we share the high-level concept of language guidance, our technical design fundamentally departs from TxtTok. Naive text conditioning risks representation collapse (*i.e.*, shortcut learning) and restricts decoding flexibility. Instead, a key design of ours is a language-drop scheme—absent in TxtTok—that not only robustifies discrete representations but uniquely empowers the detokenizer to execute classifier-free guidance directly in the *motion space* during generation. Furthermore, we propose a fully transformer-based architecture, while been explored in continuous VAE representations [12, 49], their application in VQ-based tokenizers remains limited—existing attempts [22, 43] only stack shallow self-attention layers after CNN-based residual blocks, which may be insufficient for effective cross-modal alignment. Moreover, while M2DM [35] also adopts a Transformer-based VQ-VAE, it shares with prior motion tokenizers a 1:1 frame-to-latent alignment trained on 64-frame short crops. In contrast, LG-Tok encodes 196-frame sequences via N learnable queries fully decoupled from frame count (N=25 in LG-Tok-mini), enabling each token to capture sentence-level semantics rather than local per-frame features.

3 Method

3.1 Preliminary

Motion tokenizer. Human motion tokenization converts continuous motion representations into discrete tokens to facilitate the generative models. Traditional approaches employ vanilla VQ-VAE [25, 52, 80], where a tokenizer (encoder) $\mathcal{E}(\cdot)$ compresses motion m into latent features $z = \mathcal{E}(m) \in \mathbb{R}^{T \times d}$, followed by a quantizer $\mathcal{Q}(z)$ that maps each latent feature to its nearest codebook entry, producing discrete tokens $x \in [V]^T$, where V is the codebook size. Then, a detokenizer (decoder) $\mathcal{D}(\cdot)$ reconstructs motion from the dequantized embeddings $\hat{m} = \mathcal{D}(\hat{z})$. To address quantization errors, Residual VQ-VAE [23, 83] performs iterative residual quantization with N quantizers, creating the same-scale token sets $(x^{(1)}, \ldots, x^{(N)})$ where each $x^{(n)} \in [V]^T$. Recent work, MoSa [43], introduces interpolations before each residual quantization to construct tokens at different scales $S = (s_1, s_2, \ldots, s_N)$ through downsampling: $x^{(n)} = \mathcal{Q}^{(n)}(\Downarrow(z^{(n)}, s_n))$

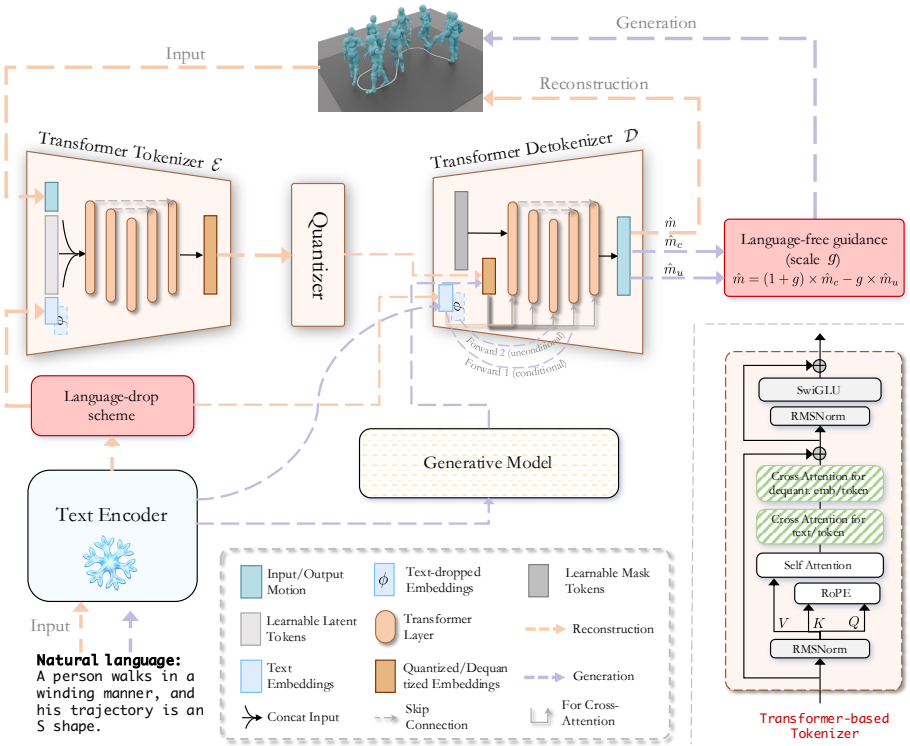


Fig. 3: Overview of LG-Tok. Given a motion sequence and its natural language description, a frozen text encoder extracts text embeddings, which are concatenated with learnable latent tokens and linearly projected motion as inputs to the Transformer-based tokenizer $\mathcal{E}$ (self-attention only), producing discrete motion tokens via the quantizer. A language-drop scheme randomly discards text input with probability p, preventing shortcut learning. In the Transformer-based detokenizer $\mathcal{D}$, learnable mask tokens interact with text embeddings and dequantized embeddings via two dedicated cross-attention layers (shown on the right), respectively, to reconstruct the motion sequence. For Generation, motion tokens sampled by the generative model are dequantized and fed into $\mathcal{D}$ with language-free guidance—two forward passes (conditional and unconditional) are performed and extrapolated in motion space to synthesize diverse, high-fidelity human motion.

where the symbol $\Downarrow(\cdot, s_n)$ denotes downsampling latent features to specific granularities s_n. This process produces compact multi-scale tokens $x^{(n)} \in [V]^{s_n}$ rather than same-scale representations.

Generative model. Leveraging multi-scale tokens, MoSa enables Scale-wise Autoregressive (SAR) modeling that reformulates traditional token-by-token prediction [80] into scale-by-scale generation. The SAR likelihood is defined as:

$$p(x^{(1)}, \ldots, x^{(N)} \mid c) = \prod_{n=1}^{N} p(x^{(n)} \mid x^{(1)}, \ldots, x^{(n-1)}, c) \tag{1}$$

where $x^{(n)} = (x_1^{(n)}, \ldots, x_{s_n}^{(n)})$ represents all tokens at scale s_n, and will be predicted simultaneously at step n. This approach generates multiple tokens in parallel at each autoregressive step, conditioned on previous scales and condition c. Given this framework's notable gains in quality and efficiency, we adopt

it as our generative model. In the following sections, we use LG-Tok to refer to this complete generation framework by default. We also apply LG-Tok to MoMask to demonstrate its generalizability. The complete tokenization-generation-detokenization pipeline is detailed in Appendix B.

3.2 Transformer-based Tokenizer

To facilitate understanding of our approach, we introduce our Transformer-based Tokenizer first.

Existing discrete tokenizers have achieved substantial results but still exhibit a limitation in their standard workflow: The 196-frame motion is compressed into 49 latent tokens via 1D convolutional encoding with 4× downsampling, whose local receptive fields struggle to support global language guidance and to enhance the global contextual awareness of token representations. To this end, we propose a Transformer-based Tokenizer. As depicted in Fig. 3, both our tokenizer and detokenizer are attention-based [65]. We predefine a learnable sequence of latent tokens and use these tokens for reconstruction and subsequent generation. With the self-attention mechanism, token representations can enhance global contextual awareness. After feature encoding, these latent tokens constitute a representation of motion. Specifically, we concatenate a set of learnable tokens $z_l \in \mathbb{R}^{T\times d}$ of length T with linearly transformed motion $m \in \mathbb{R}^{F\times d}$ as the input to the tokenizer, and retain only the output corresponding to the learnable latent tokens as the tokenizer output:

$$z = \mathcal{E}([z_l; m]) \tag{2}$$

subsequently, the latent tokens undergo vector quantization to obtain multi-scale discrete tokens (as mentioned in Sec. 3.1) to support subsequent SAR modeling, and yield dequantized embeddings $\hat{z} \in \mathbb{R}^{T\times d}$. For the detokenizer, we reconstruct motion from a sequence of learnable mask tokens $\hat{m}_l \in \mathbb{R}^{F\times d}$ [15, 27]:

$$\hat{m} = \mathcal{D}(\hat{m}_l, \hat{z}) \tag{3}$$

unlike the tokenizer, the mask tokens interact with the dequantized embeddings through cross-attention. This design is inspired by object queries [9] in object detection, and we provide detailed ablations in the experimental section. At the detokenizer end, we use a linear layer to regress from mask tokens to motion space. Notably, we do not apply patchify processing [16] to motion throughout process, since 1D motion (196 frames) incurs a significantly lower computational cost than 2D images (256×256), and this operation also also avoids information loss.

Our architecture is closely integrated with LLaMA. We incorporate RMSNorm [79], SwiGLU activation [58], and advanced rotary position embedding (RoPE) [60]. The RoPE *base* is set to 100 to accommodate short sequence tasks. We further enhance the transformer networks of both tokenizer $\mathcal{E}$ and detokenizer $\mathcal{D}$ with UNet-like long skip connections for higher-fidelity motion reconstruction.

3.3 Language-Guided Tokenization

Introducing language as auxiliary guidance during tokenization, tokens are relieved from redundantly encoding high-level semantics and can instead focus on capturing the fine-grained motion details that language alone cannot fully convey. Building on the flexibility of our transformer-based tokenizer, we introduce Language-Guided Tokenization, which aligns natural language with motion during tokenization. This approach, typically reserved for the generation phase, is extended to the tokenization stage in our work. Given textual descriptions of motion, we use a frozen LLaMA [21] as the text encoder to extract text embeddings. These embeddings are injected into both the tokenizer and detokenizer, providing semantic guidance throughout the tokenization process. As illustrated in Fig. 3, the tokenizer input is further concatenated with linearly projected text embeddings $t \in \mathbb{R}^{W \times d}$. The Eq. 2 is updated to $z = \mathcal{E}([t; z_l; m])$, yielding compact, high-level semantic representations. For the detokenizer, the mask tokens additionally interact with text embeddings through another cross-attention: $\hat{m} = \mathcal{D}(\hat{m}_l, \hat{z}, t)$. We train LG-Tok using smooth L1 loss for motion reconstruction, without involving text reconstruction. In the generation phase, the provided text descriptions are used for both generation and detokenization. The text embeddings and the latent tokens sampled by the generative model are fed into the detokenizer to produce the final motion.

Despite its simplicity, we emphasize that this approach achieves three benefits: *1)* enables latent tokens to learn motion representations while simultaneously absorbing linguistic semantic knowledge during the tokenizing process; *2)* allows language conditions to assist masked token sequences in effectively reconstructing the input motion during the detokenizing process. *3)* simplifies the learning of generative models by more compact semantic representations. We demonstrated these views in our experiments.

3.4 Language-Drop Scheme

While language guidance benefits tokenization, naively incorporating it risks shortcut learning: the tokenizer-detokenizer may over-rely on well-trained text embeddings for encoding and reconstruction, undermining the tokens' ability to capture fine-grained motion details. To mitigate this, we propose a language-drop scheme that randomly removes language conditions during training. Specifically, we drop the text input ($t = \emptyset$) with a probability of p, compelling the model to develop robust token representations independent of linguistic input. This ensures that the learned tokens remain informative and self-sufficient, rather than collapsing into mere pointers to semantic text features.

Beyond preventing shortcut learning, this scheme enables language-free guidance decoding during the token-to-motion generation phase. Since the detokenizer is trained to operate both with and without text, we can apply Classifier-Free Guidance directly at the final detokenizing stage. The final generated motion $\hat{m}$ is obtained by extrapolating from the unconditional motion $\hat{m}_u$ to the conditional motion $\hat{m}_c$ with guidance scale g:

$$\hat{m} = (1 + g) \times \hat{m}_c - g \times \hat{m}_u \tag{4}$$

This provides a complementary guidance mechanism to existing methods that operate in the logits space. Our approach further applies guidance at the motion space, offering a lightweight yet effective boost to generation quality.

4 Experiment

In this section, we present the results of our experiments. We introduce our experimental setup in Sec. 4.1. Subsequently, we compare our results with competing methods in Sec. 4.2 and 4.3, followed by an in-depth analysis of the language-drop scheme in Sec. 4.4 and related ablation experiments in Sec. 4.5. Finally, we provide reconstruction–generation trade-off analysis in Sec. 4.7 and tokenizer analysis in Sec. 4.6. Additional discussions on the extended applications, inference time, limitations, and more are provided in the Appendix.

4.1 Experimental Setup

We conduct experiments on three text-to-motion benchmarks: HumanML3D [24], KIT-ML [53] and the larger-scale Motion-X dataset [42]. We follow the most evaluation protocol proposed in [24, 47].

Datasets. HumanML3D dataset includes 14,616 high-quality motions paired with 44,970 text descriptions, where three different captions describe each motion. KIT-ML comprises 3,911 motions with 6,278 text annotations. Motion-X is the larger motion-text dataset, featuring greater diversity. Following the protocol of the first dataset, we filter out motion-text pairs exceeding 200 frames, resulting in 37,751 motion sequences and 61,637 text captions. The datasets are split into training, validation, and test sets with ratios of 80%, 5%, and 15%, respectively.

For standardization, all datasets adopt the `meng` (64-67 dim) representation [47], as their research uncovered the redundancy of the original features. That is, Motion-X's whole-body representation is also converted to `meng`. Since most textual descriptions primarily focus on body actions, we choose to ignore finger and facial information to avoid unnecessary modal discrepancies. We provide supplementary training details for the Motion-X feature extractor in Appendix D.

Evaluation metrics. We adopt the following evaluation metrics: (1) the *Frechet Inception Distance (FID)*, which assesses the overall action quality by measuring the distributional difference between the high-level features of generated and real actions; (2) *R-Precision* and *Multimodal distance*, which are used to measure the semantic consistency between the input text and the generated actions; (3) *Multimodality*, which is used to evaluate the diversity of actions generated from the exact text. (4) *CLIP-score*, which measures the compatibility of motion-text pairs by calculating the cosine similarity.

Implementation details. We train LG-Tok using the AdamW optimizer with a batch size of 128 for 200 epochs. The learning rate is reduced from 2×10^{-4} to 2×10^{-5} at the 180th epoch. We set the gradient clipping factor to 0.01 to prevent gradient explosion. No velocity loss is employed during optimization, as

Table 2: Quantitative evaluation on the HumanML3D and Motion-X test set. Each experiment is evaluated 20 times, and ± indicates a 95% confidence interval. **Bold** and underline indicate the best and the second best result, respectively. '†' denotes our reimplementation on the `meng` representation [47].

Datasets	Methods	Venues	#Tokens	R Precision↑			FID↓	MultiModal Dist↓	MultiModality↑	CLIP-score↑
				Top 1	Top 2	Top 3				
Human ML3D	**Real motions**	-	-	$0.501^{\pm.002}$	$0.696^{\pm.003}$	$0.792^{\pm.002}$	$0.000^{\pm.000}$	$3.251^{\pm.010}$	-	$0.639^{\pm.001}$
	MotionDiffuse [81]	TPAMI'24	-	$0.450^{\pm.006}$	$0.641^{\pm.005}$	$0.753^{\pm.005}$	$0.778^{\pm.005}$	$3.490^{\pm.023}$	$\mathbf{3.179}^{\pm.046}$	$0.606^{\pm.004}$
	T2M-GPT [80]	CVPR'23	49	$0.470^{\pm.003}$	$0.659^{\pm.002}$	$0.758^{\pm.002}$	$0.335^{\pm.003}$	$3.505^{\pm.017}$	$2.018^{\pm.053}$	$0.607^{\pm.005}$
	MMM [52]	CVPR'24	49	$0.487^{\pm.003}$	$0.683^{\pm.003}$	$0.782^{\pm.001}$	$0.132^{\pm.004}$	$3.359^{\pm.009}$	$1.241^{\pm.073}$	$0.635^{\pm.003}$
	MoMask [23]	CVPR'24	294	$0.490^{\pm.004}$	$0.687^{\pm.003}$	$0.786^{\pm.003}$	$0.116^{\pm.006}$	$3.353^{\pm.010}$	$1.263^{\pm.079}$	$0.637^{\pm.003}$
	StableMoFusion† [29]	ACM MM'24	-	$0.510^{\pm.002}$	$0.710^{\pm.003}$	$0.810^{\pm.003}$	$0.177^{\pm.006}$	$3.182^{\pm.009}$	$1.969^{\pm.053}$	$0.654^{\pm.001}$
	DisCoRD† [13]	ICCV'25	-	$0.492^{\pm.003}$	$0.690^{\pm.003}$	$0.790^{\pm.003}$	$0.117^{\pm.004}$	$3.306^{\pm.008}$	$1.369^{\pm.061}$	$0.640^{\pm.001}$
	MARDM [47]	CVPR'25	-	$0.500^{\pm.004}$	$0.695^{\pm.003}$	$0.795^{\pm.003}$	$0.114^{\pm.007}$	$3.270^{\pm.009}$	$\underline{2.231}^{\pm.071}$	$0.642^{\pm.002}$
	MoSa† [43]	Arxiv'25	236	$0.518^{\pm.002}$	$0.712^{\pm.002}$	$0.809^{\pm.002}$	$\underline{0.064}^{\pm.004}$	$3.150^{\pm.008}$	$1.789^{\pm.056}$	$0.657^{\pm.001}$
	LG-Tok-mini	-	104	$0.521^{\pm.003}$	$0.715^{\pm.003}$	$0.811^{\pm.003}$	$0.085^{\pm.004}$	$3.113^{\pm.001}$	$1.728^{\pm.053}$	$0.655^{\pm.001}$
	LG-Tok-mid	-	160	$\underline{0.537}^{\pm.003}$	$\underline{0.729}^{\pm.002}$	$\underline{0.821}^{\pm.002}$	$0.109^{\pm.005}$	$\underline{3.061}^{\pm.010}$	$1.674^{\pm.052}$	$\underline{0.664}^{\pm.001}$
	LG-Tok	-	236	$\mathbf{0.542}^{\pm.003}$	$\mathbf{0.736}^{\pm.002}$	$\mathbf{0.830}^{\pm.002}$	$\mathbf{0.057}^{\pm.003}$	$\mathbf{2.997}^{\pm.010}$	$1.540^{\pm.059}$	$\mathbf{0.669}^{\pm.001}$
Motion-X	**Real motions**	-	-	$0.595^{\pm.002}$	$0.787^{\pm.001}$	$0.868^{\pm.001}$	$0.000^{\pm.000}$	$3.717^{\pm.007}$	-	$0.672^{\pm.000}$
	MotionDiffuse† [81]	TPAMI'24	-	$0.559^{\pm.003}$	$0.752^{\pm.002}$	$0.839^{\pm.001}$	$0.954^{\pm.014}$	$4.167^{\pm.023}$	$2.476^{\pm.098}$	$0.660^{\pm.001}$
	T2M-GPT† [80]	CVPR'23	49	$0.470^{\pm.003}$	$0.644^{\pm.002}$	$0.735^{\pm.003}$	$1.085^{\pm.032}$	$5.488^{\pm.023}$	$\mathbf{12.807}^{\pm.405}$	$0.622^{\pm.001}$
	MMM† [52]	CVPR'24	49	$0.424^{\pm.002}$	$0.600^{\pm.002}$	$0.695^{\pm.002}$	$2.918^{\pm.036}$	$6.098^{\pm.019}$	$2.342^{\pm.193}$	$0.607^{\pm.001}$
	MoMask† [23]	CVPR'24	294	$0.502^{\pm.003}$	$0.694^{\pm.002}$	$0.790^{\pm.002}$	$0.247^{\pm.008}$	$4.832^{\pm.009}$	$2.715^{\pm.091}$	$0.644^{\pm.001}$
	StableMoFusion† [29]	ACM MM'24	-	$0.474^{\pm.003}$	$0.682^{\pm.002}$	$0.787^{\pm.002}$	$0.213^{\pm.008}$	$4.888^{\pm.014}$	$2.985^{\pm.111}$	$0.607^{\pm.001}$
	DisCoRD† [13]	ICCV'25	-	$0.518^{\pm.002}$	$0.714^{\pm.002}$	$0.808^{\pm.002}$	$0.164^{\pm.008}$	$4.614^{\pm.014}$	$2.417^{\pm.112}$	$0.649^{\pm.000}$
	MARDM† [47]	CVPR'25	-	$0.528^{\pm.003}$	$0.727^{\pm.001}$	$0.820^{\pm.002}$	$0.147^{\pm.009}$	$4.433^{\pm.018}$	$3.077^{\pm.066}$	$0.643^{\pm.001}$
	MoSa† [43]	Arxiv'25	236	$0.513^{\pm.002}$	$0.703^{\pm.002}$	$0.796^{\pm.002}$	$0.210^{\pm.009}$	$4.783^{\pm.013}$	$\underline{3.167}^{\pm.122}$	$0.654^{\pm.001}$
	LG-Tok-mini	-	104	$\underline{0.588}^{\pm.002}$	$\underline{0.776}^{\pm.002}$	$0.857^{\pm.002}$	$\mathbf{0.071}^{\pm.004}$	$\underline{3.835}^{\pm.012}$	$2.235^{\pm.111}$	$\underline{0.681}^{\pm.001}$
	LG-Tok-mid	-	160	$\mathbf{0.591}^{\pm.002}$	$\mathbf{0.781}^{\pm.002}$	$\mathbf{0.864}^{\pm.001}$	$\underline{0.076}^{\pm.005}$	$\mathbf{3.758}^{\pm.010}$	$2.245^{\pm.091}$	$\mathbf{0.682}^{\pm.000}$
	LG-Tok	-	236	$0.582^{\pm.002}$	$0.775^{\pm.001}$	$\underline{0.858}^{\pm.001}$	$0.088^{\pm.006}$	$3.844^{\pm.009}$	$2.294^{\pm.088}$	$\mathbf{0.682}^{\pm.000}$

the `meng` representation is sufficiently compact. We set the probability $p = 0.1$ of our language-drop scheme. For the Transformer, we stack 9 layers for both the tokenizer and detokenizer, each with 4 heads, 256 latent dimensions, and a SwiGLU dimension of 1024. To accelerate training and reduce memory usage, we enable mixed-precision training and utilize PyTorch 2.2.0 to support flash attention. For text guidance, *LLaMA-3.2-1B* serves as our text encoder, and the maximum text length is set to 77, with more extended sequences truncated. The guidance scale g is set to 2.0, 1.0 and 2.0 on the HumanML3D, Motion-X and KIT-ML, respectively. During generation, the generative model's configuration remains unchanged, except that `PAD` masks are no longer required. Note that the generative model and the evaluation retrieval metrics independently use *CLIP-ViT-B/32* for text embedding, consistent with all baselines; *LLaMA-3.2-1B* features are exclusive to LG-Tok's tokenizer and detokenizer. LG-Tok training can be completed on a single RTX 4090 GPU.

Model variants. To validate the high-level semantic representations of LG-Tok, we study three variants with different token scales: LG-Tok-mini (25 tokens), LG-Tok-mid (36 tokens), and LG-Tok (49 tokens). The interpolation scales mentioned in Sec. 3.1 are set to $S = (1, 2, \ldots, 25)$, $S = (2, 4, \ldots, 36)$, and $S = (3, 6, \ldots, 49)$ respectively. All variants use $N = 10$ scales (quantizers), resulting in 104, 160, and 236 total tokens, respectively.

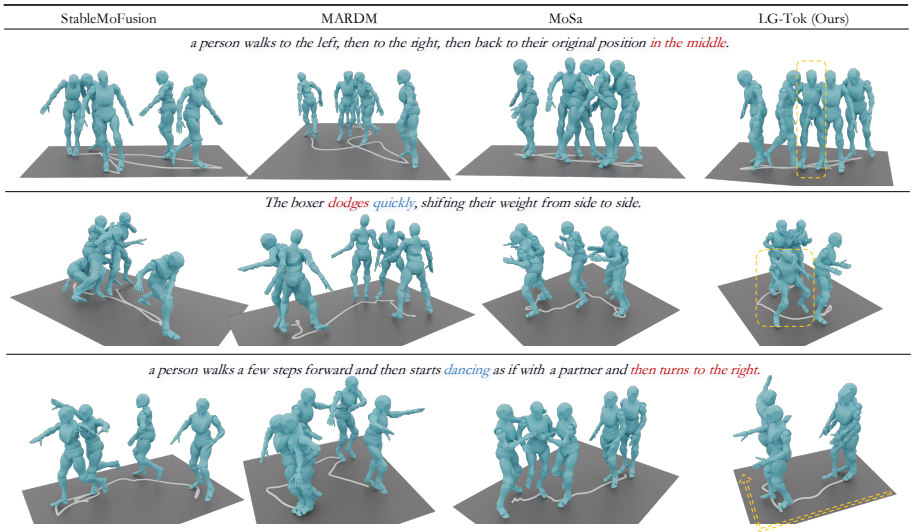


Fig. 4: Qualitative comparisons on HumanML3D dataset. Our LG-Tok demonstrates superior semantic understanding compared to existing methods. The examples show better spatial awareness ("*in the middle*"), more realistic posture synthesis ("*dodges quickly*"), and improved directional control ("*then turns to the left*").

4.2 Text-driven Motion Generation Comparison

Quantitative comparisons. We evaluate our approach against state-of-the-art methods. Results on HumanML3D are primarily sourced from [47], while results on Motion-X are our reproduction based on the `meng` representation. We provide reproduction details in Appendix E. As shown in Table 2, our LG-Tok variants (using MoSa as the generative backbone) outperform existing methods across multiple metrics. On HumanML3D, LG-Tok achieves the best performance with a Top-1 R-Precision of **0.542** and FID of **0.057**, surpassing the previous best method, MoSa (0.518 and 0.064, respectively). On the larger and more diverse Motion-X dataset, our methods show substantial improvements, with LG-Tok-mid achieving the best Top-1 R-Precision of **0.591**. The consistent performance gains validate that transformer-based, language-guided tokenization effectively captures motion semantics. Remarkably, on both datasets, LG-Tok-mini maintains or surpasses state-of-the-art performance even after pruning more than half of the tokens (104 vs. 236), further validating the effectiveness of our high-level semantic representations. Results on the KIT-ML dataset are provided in Appendix C.

Qualitative comparisons. Fig. 4 displays qualitative comparisons of our approach against StableMoFusion [29], MARDM [47], and MoSa [43] based on HumanML3D checkpoints. We randomly select several generated samples for comparison. These three examples demonstrate that LG-Tok can understand more complex semantics. *e.g.*, in the first row, our method successfully understands "*in the middle*" (competitors fail to return to the middle position). In the second row, LG-Tok synthesizes more realistic crouching postures in the

"*dodges quickly*" motion. In the final row, our approach exhibits better directional awareness for "*then turns to the left*". More dynamic results can be found on our supplementary materials.

4.3 Comparison of Discrete Tokenizers

We compare our approach against existing discrete (VQ-based) tokenizers, which are predominantly CNN-based. Table 3 demonstrates our LG-Tok consistent improvements in both reconstruction and generation stages across HumanML3D and Motion-X. Notably, by incorporating language guidance, our compact semantic representations significantly simplify the learning of generative models. On Motion-X, text-guided LG-Tok achieves substantial generation performance gains (FID: 0.257→0.088), and the perplexity we recorded decreases from 146.5 to 103.1; similar improvements on HumanML3D (160.6→155.9). These results confirm that language as auxiliary guidance relieves tokens from encoding high-level semantics, enabling focus on fine-grained motion details.

Table 3: Reconstruction and generation performance comparison. We evaluate both reconstruction quality and generation performance across different discrete tokenizers. "*w/o* text guidance" denotes our approach without language guidance, *i.e.*, using only the transformer-based tokenization from Sec. 3.2. "⋆" denotes report from [47], and the others are our reproductions.

Methods	Reconstruction			Generation	
	FID↓	Top 1↑	MPJPE↓	FID↓	MM-Dist↓
Evaluation on HumanML3D dataset					
T2M-GPT⋆ [80]	$0.081^{\pm.001}$	$0.483^{\pm.003}$	$72.6^{\pm.001}$	$0.335^{\pm.003}$	$3.505^{\pm.017}$
MoMask⋆ [23]	$0.029^{\pm.001}$	$0.497^{\pm.002}$	$31.5^{\pm.001}$	$0.116^{\pm.006}$	$3.353^{\pm.010}$
MoMask-reprod.	$0.029^{\pm.000}$	$0.499^{\pm.003}$	$30.9^{\pm.000}$	$0.180^{\pm.005}$	$3.332^{\pm.009}$
MoMask (LG-Tok)	$\mathbf{0.019}^{\pm.000}$	$0.501^{\pm.003}$	$\mathbf{26.4}^{\pm.000}$	$0.111^{\pm.005}$	$3.300^{\pm.012}$
MoSa [43]	$0.023^{\pm.000}$	$0.496^{\pm.003}$	$43.0^{\pm.000}$	$0.064^{\pm.004}$	$3.150^{\pm.008}$
LG-Tok	$0.022^{\pm.000}$	$\mathbf{0.502}^{\pm.002}$	$39.0^{\pm.000}$	$\mathbf{0.057}^{\pm.003}$	$\mathbf{2.997}^{\pm.010}$
w/o text guidance	$0.025^{\pm.000}$	$0.494^{\pm.002}$	$39.0^{\pm.000}$	$0.062^{\pm.003}$	$3.129^{\pm.008}$
Evaluation on Motion-X dataset					
T2M-GPT [80]	$0.826^{\pm.011}$	$0.497^{\pm.002}$	$59.6^{\pm.000}$	$1.102^{\pm.000}$	$5.515^{\pm.000}$
MoMask [23]	$0.394^{\pm.005}$	$0.554^{\pm.002}$	$24.9^{\pm.000}$	$0.247^{\pm.008}$	$4.832^{\pm.009}$
MoMask (LG-Tok)	$0.076^{\pm.002}$	$\mathbf{0.580}^{\pm.002}$	$\mathbf{23.0}^{\pm.000}$	$0.157^{\pm.006}$	$4.259^{\pm.008}$
MoSa [43]	$0.072^{\pm.001}$	$0.558^{\pm.002}$	$39.0^{\pm.000}$	$0.210^{\pm.009}$	$4.783^{\pm.013}$
LG-Tok	$\mathbf{0.041}^{\pm.001}$	$0.577^{\pm.002}$	$31.0^{\pm.000}$	$\mathbf{0.088}^{\pm.006}$	$\mathbf{3.844}^{\pm.009}$
w/o text guidance	$0.090^{\pm.002}$	$0.568^{\pm.002}$	$33.0^{\pm.000}$	$0.257^{\pm.010}$	$4.274^{\pm.008}$

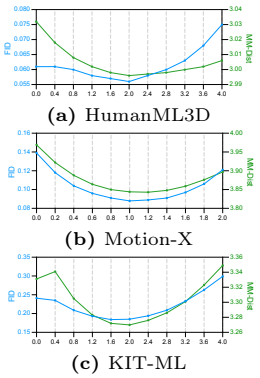
Furthermore, without text guidance (*i.e.*, using only transformer-based tokenization from Sec. 3.2), our method outperforms most competitors, demonstrating that our transformer-based architecture provides superior global context modeling. More importantly, we integrate LG-Tok into a representative motion tokenization–plus–generation framework, MoMask. The enhanced variant, MoMask (LG-Tok), consistently surpasses the original across all evaluation metrics, demonstrating strong generalization.

4.4 In-depth Analysis of Language-Drop Scheme

Table 4 validates the language-drop scheme. A revealing pattern emerges consistently across all three datasets: without the language-drop scheme, reconstruction Top-1 and MM-Dist are marginally *better*, yet reconstruction FID degrades substantially. This discrepancy suggests a shortcut effect: without language-drop, the tokenizer may over-rely on well-trained text embeddings, causing residual

Method	Reconstruction			Generation		
	FID↓	Top 1↑	MM-Dist.↓	FID↓	Top 1↑	MM-Dist.↓
Evaluation on HumanML3D dataset						
LG-Tok ($g = 2.0$)	$\mathbf{0.022}^{\pm.000}$	$0.502^{\pm.002}$	$3.250^{\pm.010}$	$\mathbf{0.057}^{\pm.003}$	$\mathbf{0.542}^{\pm.003}$	$\mathbf{2.997}^{\pm.010}$
LG-Tok ($g = 0.0$)				$0.061^{\pm.003}$	$0.534^{\pm.003}$	$3.032^{\pm.011}$
w/o lang-drop scheme	$0.039^{\pm.000}$	$\mathbf{0.505}^{\pm.002}$	$\mathbf{3.233}^{\pm.009}$	$0.132^{\pm.005}$	$0.533^{\pm.003}$	$3.068^{\pm.007}$
Evaluation on Motion-X dataset						
LG-Tok ($g = 1.0$)	$\mathbf{0.041}^{\pm.001}$	$0.577^{\pm.002}$	$3.890^{\pm.008}$	$\mathbf{0.088}^{\pm.006}$	$\mathbf{0.582}^{\pm.002}$	$\mathbf{3.844}^{\pm.009}$
LG-Tok ($g = 0.0$)				$0.139^{\pm.008}$	$0.576^{\pm.002}$	$3.969^{\pm.012}$
w/o lang-drop scheme	$0.057^{\pm.002}$	$\mathbf{0.580}^{\pm.002}$	$\mathbf{3.846}^{\pm.007}$	$0.136^{\pm.007}$	$0.574^{\pm.002}$	$3.918^{\pm.012}$
Evaluation on KIT-ML dataset						
LG-Tok ($g = 2.0$)	$\mathbf{0.108}^{\pm.003}$	$0.373^{\pm.005}$	$3.395^{\pm.011}$	$\mathbf{0.185}^{\pm.008}$	$\mathbf{0.401}^{\pm.006}$	$\mathbf{3.270}^{\pm.013}$
LG-Tok ($g = 0.0$)				$0.241^{\pm.014}$	$0.390^{\pm.006}$	$3.331^{\pm.017}$
w/o lang-drop scheme	$0.139^{\pm.004}$	$\mathbf{0.378}^{\pm.006}$	$\mathbf{3.324}^{\pm.011}$	$0.251^{\pm.015}$	$0.375^{\pm.005}$	$3.469^{\pm.013}$

Table 4: In-depth analysis of language-drop scheme. Symbol g denotes the guidance scale for language-free decoding during generation. "*w/o* lang-drop scheme" denotes the tokenizer trained with language guidance but without randomly dropping text.



(a) HumanML3D

(b) Motion-X

(c) KIT-ML

Fig. 5: Evaluation sweep over guidance scale g. $g = 2.0$, $g = 1.0$ and $g = 2.0$ yield the best performance for HumanML3D, Motion-X and KIT-ML, respectively.

text priors to be embedded into the reconstructed motions, which could artificially inflate text-motion alignment scores in the joint embedding space. In contrast, FID—which directly measures the distributional distance between motion features—appears more sensitive to the degraded motion fidelity. Our language-drop scheme is designed to mitigate this shortcut by compelling both the tokenizer and detokenizer to operate without text, encouraging discrete tokens to be self-sufficient representations of motion, thereby restoring FID to healthy levels. The resulting tokens, which captured fine-grained motion details, also stabilize generation quality. This scheme further enables language-free guidance at inference: sweeping g in Fig. 5 shows consistent gains across all datasets.

4.5 Ablation Studies

The remaining ablation studies were conducted on a smaller version (LG-Tok-tiny) with reduced model size and training data for efficient hyperparameter tuning. LG-Tok-tiny uses 3-layer transformers trained on 5,000 samples.

Transformer architecture. Tables 5a- 5d demonstrate the advantages of our transformer design over vanilla transformers [65]. RMSNorm, SwiGLU activation, and skip connections all contribute to improved performance. Notably, RoPE with *base*=100 is better suited for our short sequence task (∼10s, 196 frames) compared to other bases.

Guidance location. Table 5e validates our method's benefits. Injecting guidance into both tokenizer and detokenizer achieves the best results, confirming that language guidance: 1) enables latent tokens to learn motion representations while simultaneously absorbing linguistic semantic knowledge during tokenizing; 2) allows language conditions to assist masked token sequences in effectively reconstructing input motion during detokenizing.

Table 5: Ablation studies on LG-Tok-tiny. We ablate key design choices affecting LG-Tok's reconstruction performance on HumanML3D dataset. Default setting: RMSNorm, SwiGLU activation, skip connections, RoPE with *base*=100, LLaMA text encoder, with in-context conditioning for tokenizer and cross-attention for detokenizer. Injecting natural language to both tokenizer and detokenizer obtains the best results.

Normalization	FID↓	MPJPE↓
LayerNorm	$0.050^{\pm.001}$	58.7
RMSNorm	$0.049^{\pm.000}$	56.1

(a) Normalization

Architecture	FID↓	MPJPE↓
GeLU	$0.050^{\pm.001}$	57.4
SwiGLU	$0.049^{\pm.000}$	56.1

(b) Activation

Activation	FID↓	MPJPE↓
Skip Conn.	$0.049^{\pm.000}$	56.1
w/o Skip Conn.	$0.059^{\pm.001}$	61.7

(c) Skip connections

Position embedding	FID↓	MPJPE↓
Learnable	$0.065^{\pm.001}$	54.7
RoPE (*base*=10)	$0.061^{\pm.001}$	54.8
RoPE (*base*=100)	$0.049^{\pm.000}$	56.1
RoPE (*base*=1000)	$0.042^{\pm.001}$	56.3
RoPE (*base*=10000)	$0.052^{\pm.001}$	56.6

(d) Position embedding

Guidance location	FID↓	MPJPE↓
None	$0.064^{\pm.001}$	64.1
Tokenizer only	$0.063^{\pm.001}$	57.5
Detokenizer only	$0.055^{\pm.000}$	58.7
Tokenizer & Detokenizer	$0.049^{\pm.000}$	56.1

(e) Guidance location

Frozen text encoder	FID↓	MPJPE↓
CLIP	$0.051^{\pm.000}$	59.1
T5	$0.053^{\pm.000}$	58.9
BERT	$0.055^{\pm.001}$	58.9
LLaMA	$0.049^{\pm.000}$	56.1

(f) Text encoder chosen

Tokenizer		Detokenizer		FID↓	MPJPE↓
In-Context	Cross-Attn.	In-Context	Cross-Attn.		
✓	✗	✓	✗	$0.053^{\pm.001}$	66.1
✗	✓	✗	✓	$1.120^{\pm.004}$	111.2
✓	✗	✗	✓	$0.049^{\pm.000}$	56.1
✗	✓	✓	✗	$2.049^{\pm.007}$	127.7

(g) Latent-tokens/motion/text interaction in tokenizer & Mask-tokens/latent/text interation in detokenizer

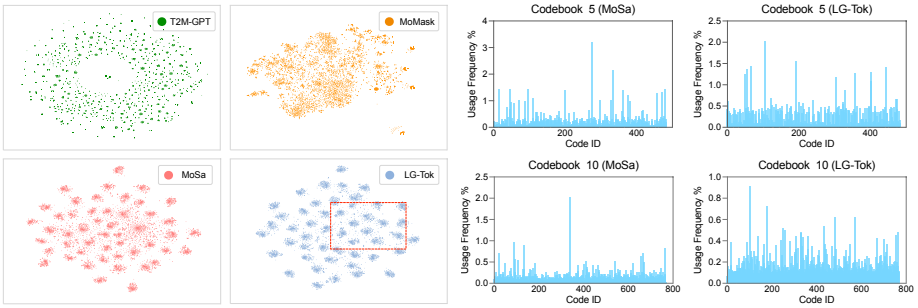


Fig. 6: Tokenizer analysis. Left: t-SNE visualization of dequantized embedding space representations. LG-Tok demonstrates well-structured clusters with clear boundaries, indicating effective capture of distinct motion patterns. Right: code usage comparison on the HumanML3D test-set. LG-Tok achieves more uniform codebook utilization.

Text encoder choice. We compare popular pretrained text encoders [14, 21, 55, 69] in Table 5f. LLaMA achieves the best performance among the tested encoders.

Interaction methods. Table 5g ablates how tokens interact in both the tokenizer and detokenizer. We compare two interaction strategies: *In-context* (concatenating all tokens for joint self-attention) and *Cross-attention* (two separate cross-attention between different token types). The results show that In-context concatenation performs better in the tokenizer, while cross-attention excels in the detokenizer.

4.6 Tokenizer Analysis

We provide further analysis of the learned representations of our tokenizer. As shown in Fig. 6 (Left), the t-SNE representations of dequantized embeddings from different discrete tokenizers reveal that our LG-Tok demonstrates well-structured clusters with clear boundaries (*e.g.*, in the red box), indicating that our compact, high-level semantic representations effectively capture distinct motion patterns. Fig. 6 (Right) compares codebook usage frequencies across different scales on the HumanML3D test-set. LG-Tok achieves more uniform codebook utilization across all scales, indicating better coverage of the motion space. This uniform usage demonstrates that our language-guided, transformer-based approach possesses learned comprehensive motion representations without redundancy.

4.7 Reconstruction–Generation Trade-off Analysis

Table 6 makes concrete the trade-off raised in Sec. 1, where generation quality reflects two competing forces: more tokens raise the representational ceiling and benefit reconstruction, while longer sequences burden the autoregressive model. Which force prevails may hinge on the amount of training data, we think. On the smaller HumanML3D (14k sequences), the nearly fourfold drop in reconstruction error (rFID from 0.094 to 0.022) lets the former dominate, and the full LG-Tok with 236 tokens wins on every metric. On the larger Motion-X (37k), the gain is far milder (roughly twofold, rFID from 0.094 to 0.041), and beyond 160 tokens the latter takes over: generation degrades (gFID from 0.071 to 0.088), leaving LG-Tok-mid as the sweet spot.

Table 6: Trade-off across LG-Tok variants.

	HumanML3D			Motion-X		
# Tokens	LG-Tok-mini 104	LG-Tok-mid 160	LG-Tok 236	LG-Tok-mini 104	LG-Tok-mid 160	LG-Tok 236
Reconstruction						
rFID↓	$0.094^{\pm.001}$	$0.052^{\pm.000}$	$\mathbf{0.022}^{\pm.000}$	$0.094^{\pm.002}$	$0.051^{\pm.001}$	$\mathbf{0.041}^{\pm.001}$
MPJPE↓	$50.2^{\pm.0}$	$40.6^{\pm.0}$	$\mathbf{39.3}^{\pm.0}$	$36.4^{\pm.0}$	$34.2^{\pm.0}$	$\mathbf{31.7}^{\pm.0}$
Generation						
gFID↓	$0.085^{\pm.004}$	$0.109^{\pm.005}$	$\mathbf{0.057}^{\pm.003}$	$\mathbf{0.071}^{\pm.004}$	$0.076^{\pm.005}$	$0.088^{\pm.006}$
Top-1↑	$0.521^{\pm.003}$	$0.537^{\pm.003}$	$\mathbf{0.542}^{\pm.003}$	$0.588^{\pm.002}$	$\mathbf{0.591}^{\pm.002}$	$0.582^{\pm.002}$
MM-Dist↓	$3.113^{\pm.010}$	$3.061^{\pm.010}$	$\mathbf{2.997}^{\pm.010}$	$3.835^{\pm.012}$	$\mathbf{3.758}^{\pm.010}$	$3.844^{\pm.009}$
CLIP↑	$0.655^{\pm.001}$	$0.664^{\pm.001}$	$\mathbf{0.669}^{\pm.001}$	$0.681^{\pm.001}$	$\mathbf{0.682}^{\pm.000}$	$\mathbf{0.682}^{\pm.000}$

5 Conclusion

In this paper, we present Language-Guided Transformer Tokenization (LG-Tok) for text-driven human motion generation. By aligning natural language with motion during tokenization and employing a Transformer-based tokenizer-detokenizer architecture, LG-Tok provides semantically informed latent representations that improve reconstruction fidelity and simplify generative learning. LG-Tok achieves state-of-the-art results on three benchmarks.

Acknowledgements

This work was supported by National Natural Science Foundation of China (No. 62473007), Guangdong Outstanding Youth Fund (No. 2026B1515020015), Shenzhen Innovation in Science and Technology Foundation for The Excellent Youth Scholars (No. RCYX20231211090248064).

References

1. Ahn, H., Ha, T., Choi, Y., Yoo, H., Oh, S.: Text2action: Generative adversarial synthesis from language to action. In: 2018 IEEE International Conference on Robotics and Automation (ICRA). pp. 5915–5920. IEEE (2018)
2. Ahuja, C., Morency, L.P.: Language2pose: Natural language grounded pose forecasting. In: 2019 International Conference on 3D Vision (3DV). pp. 719–728. IEEE (2019)
3. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: a visual language model for few-shot learning. Advances in neural information processing systems **35**, 23716–23736 (2022)
4. Alexanderson, S., Nagy, R., Beskow, J., Henter, G.E.: Listen, denoise, action! audio-driven motion synthesis with diffusion models. ACM Transactions on Graphics (TOG) **42**(4), 1–20 (2023)
5. Andreou, N., Wang, X., Abrevaya, V.F., Cani, M.P., Chrysanthou, Y., Kalogeiton, V.: Lead: Latent realignment for human motion diffusion. In: Computer Graphics Forum. p. e70093. Wiley Online Library (2025)
6. Azadi, S., Shah, A., Hayes, T., Parikh, D., Gupta, S.: Make-an-animation: Large-scale text-conditional 3d human motion generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15039–15048 (2023)
7. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473 (2014)
8. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in neural information processing systems **33**, 1877–1901 (2020)
9. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European conference on computer vision. pp. 213–229. Springer (2020)
10. Chang, H., Zhang, H., Jiang, L., Liu, C., Freeman, W.T.: Maskgit: Masked generative image transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11315–11325 (2022)
11. Chen, H., Wang, Z., Li, X., Sun, X., Chen, F., Liu, J., Wang, J., Raj, B., Liu, Z., Barsoum, E.: Softvq-vae: Efficient 1-dimensional continuous tokenizer. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 28358–28370 (2025)
12. 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)
13. Cho, J., Kim, J., Kim, J., Kim, M., Kang, M., Hong, S., Oh, T.H., Yu, Y.: Discord: Discrete tokens to continuous motion via rectified flow decoding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14602–14612 (2025)
14. Chung, H.W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S., et al.: Scaling instruction-finetuned language models. Journal of Machine Learning Research **25**(70), 1–53 (2024)
15. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers). pp. 4171–4186 (2019)

16. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
17. Esser, P., Rombach, R., Ommer, B.: Taming transformers for high-resolution image synthesis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12873–12883 (2021)
18. Geirhos, R., Jacobsen, J.H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., Wichmann, F.A.: Shortcut learning in deep neural networks. Nature Machine Intelligence **2**(11), 665–673 (2020)
19. Ghosh, A., Cheema, N., Oguz, C., Theobalt, C., Slusallek, P.: Synthesis of compositional animations from textual descriptions. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1396–1406 (2021)
20. Ginosar, S., Bar, A., Kohavi, G., Chan, C., Owens, A., Malik, J.: Learning individual styles of conversational gesture. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3497–3506 (2019)
21. Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., et al.: The llama 3 herd of models. arXiv preprint arXiv:2407.21783 (2024)
22. Guo, C., Hwang, I., Wang, J., Zhou, B.: Snapmogen: Human motion generation from expressive texts. arXiv preprint arXiv:2507.09122 (2025)
23. 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)
24. 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)
25. 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)
26. 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)
27. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16000–16009 (2022)
28. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022)
29. 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)
30. Karunratanakul, K., Preechakul, K., Suwajanakorn, S., Tang, S.: Guided motion diffusion for controllable human motion synthesis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2151–2162 (2023)
31. 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)
32. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)

33. Kingma, D.P., Welling, M., et al.: An introduction to variational autoencoders. Foundations and Trends® in Machine Learning **12**(4), 307–392 (2019)
34. Komura, T., Habibie, I., Holden, D., Schwarz, J., Yearsley, J.: A recurrent variational autoencoder for human motion synthesis. In: The 28th British Machine Vision Conference (2017)
35. Kong, H., Gong, K., Lian, D., Mi, M.B., Wang, X.: Priority-centric human motion generation in discrete latent space. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14806–14816 (2023)
36. Le, N., Pham, T., Do, T., Tjiputra, E., Tran, Q.D., Nguyen, A.: Music-driven group choreography. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8673–8682 (2023)
37. Lee, D., Kim, C., Kim, S., Cho, M., Han, W.S.: Autoregressive image generation using residual quantization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11523–11532 (2022)
38. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: International conference on machine learning. pp. 19730–19742. PMLR (2023)
39. Li, T., Tian, Y., Li, H., Deng, M., He, K.: Autoregressive image generation without vector quantization. Advances in Neural Information Processing Systems **37**, 56424–56445 (2024)
40. Li, X., Qiu, K., Chen, H., Kuen, J., Gu, J., Raj, B., Lin, Z.: Imagefolder: Autoregressive image generation with folded tokens. arXiv preprint arXiv:2410.01756 (2024)
41. Li, Z., Yuan, W., He, Y., Qiu, L., Zhu, S., Gu, X., Shen, W., Dong, Y., Dong, Z., Yang, L.T.: Lamp: Language-motion pretraining for motion generation, retrieval, and captioning. arXiv preprint arXiv:2410.07093 (2024)
42. Lin, J., Zeng, A., Lu, S., Cai, Y., Zhang, R., Wang, H., Zhang, L.: Motion-x: A large-scale 3d expressive whole-body human motion dataset. Advances in Neural Information Processing Systems **36**, 25268–25280 (2023)
43. Liu, M., Yan, S., Wang, Y., Li, Y., Bian, G.B., Liu, H.: Mosa: Motion generation with scalable autoregressive modeling. arXiv preprint arXiv:2511.01200 (2025)
44. Lou, Y., Zhu, L., Wang, Y., Wang, X., Yang, Y.: Diversemotion: Towards diverse human motion generation via discrete diffusion. arXiv preprint arXiv:2309.01372 (2023)
45. 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)
46. Ma, C., Jiang, Y., Wu, J., Yang, J., Yu, X., Yuan, Z., Peng, B., Qi, X.: Unitok: A unified tokenizer for visual generation and understanding. arXiv preprint arXiv:2502.20321 (2025)
47. Meng, Z., Xie, Y., Peng, X., Han, Z., Jiang, H.: Rethinking diffusion for text-driven human motion generation: Redundant representations, evaluation, and masked autoregression. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 27859–27871 (2025)
48. Petrovich, M., Black, M.J., Varol, G.: Action-conditioned 3d human motion synthesis with transformer vae. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10985–10995 (2021)
49. 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)

50. 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)
51. Pinyoanuntapong, E., Saleem, M.U., Wang, P., Lee, M., Das, S., Chen, C.: Bamm: Bidirectional autoregressive motion model. In: European Conference on Computer Vision. pp. 172–190. Springer (2024)
52. 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)
53. Plappert, M., Mandery, C., Asfour, T.: The kit motion-language dataset. Big data **4**(4), 236–252 (2016)
54. Qiu, K., Li, X., Kuen, J., Chen, H., Xu, X., Gu, J., Luo, Y., Raj, B., Lin, Z., Savvides, M.: Robust latent matters: Boosting image generation with sampling error synthesis. arXiv preprint arXiv:2503.08354 (2025)
55. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
56. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al.: Language models are unsupervised multitask learners. OpenAI blog **1**(8), 9 (2019)
57. Razavi, A., Van den Oord, A., Vinyals, O.: Generating diverse high-fidelity images with vq-vae-2. Advances in neural information processing systems **32** (2019)
58. Shazeer, N.: Glu variants improve transformer. arXiv preprint arXiv:2002.05202 (2020)
59. Shi, J., Liu, L., Sun, Y., Zhang, Z., Zhou, J., Nie, Q.: Genm3: Generative pretrained multi-path motion model for text conditional human motion generation. arXiv preprint arXiv:2503.14919 (2025)
60. Su, J., Ahmed, M., Lu, Y., Pan, S., Bo, W., Liu, Y.: Roformer: Enhanced transformer with rotary position embedding. Neurocomputing **568**, 127063 (2024)
61. Tang, T., Jia, J., Mao, H.: Dance with melody: An lstm-autoencoder approach to music-oriented dance synthesis. In: Proceedings of the 26th ACM international conference on Multimedia. pp. 1598–1606 (2018)
62. Tevet, G., Gordon, B., Hertz, A., Bermano, A.H., Cohen-Or, D.: Motionclip: Exposing human motion generation to clip space. In: European Conference on Computer Vision. pp. 358–374. Springer (2022)
63. Tian, K., Jiang, Y., Yuan, Z., Peng, B., Wang, L.: Visual autoregressive modeling: Scalable image generation via next-scale prediction. arXiv preprint arXiv:2404.02905 (2024)
64. Van Den Oord, A., Vinyals, O., et al.: Neural discrete representation learning. Advances in neural information processing systems **30** (2017)
65. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
66. Wan, W., Huang, Y., Wu, S., Komura, T., Wang, W., Jayaraman, D., Liu, L.: Diffusionphase: Motion diffusion in frequency domain. arXiv preprint arXiv:2312.04036 (2023)
67. Wang, H., Zhu, Y., Adam, H., Yuille, A., Chen, L.C.: Max-deeplab: End-to-end panoptic segmentation with mask transformers. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5463–5474 (2021)

68. Wang, Z., Zhang, J., Chen, Y., Jia, B., Liang, W., Huang, S.: Spatial-temporal multi-scale quantization for flexible motion generation. arXiv preprint arXiv:2508.08991 (2025)
69. Warner, B., Chaffin, A., Clavié, B., Weller, O., Hallström, O., Taghadouini, S., Gallagher, A., Biswas, R., Ladhak, F., Aarsen, T., et al.: Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. arXiv preprint arXiv:2412.13663 (2024)
70. Wu, P., Zhu, K., Liu, Y., Tang, L., Yang, J., Peng, Y., Zhai, W., Cao, Y., Zha, Z.J.: Alitok: Towards sequence modeling alignment between tokenizer and autoregressive model. arXiv preprint arXiv:2506.05289 (2025)
71. Wu, Q., Zhao, Y., Wang, Y., Tai, Y.W., Tang, C.K.: Motionllm: Multimodal motion-language learning with large language models. arXiv e-prints pp. arXiv–2405 (2024)
72. Yan, S., Liu, Y., Wang, H., Du, X., Liu, M., Liu, H.: Cross-modal retrieval for motion and text via droptriple loss. In: Proceedings of the 5th ACM International Conference on Multimedia in Asia. pp. 1–7 (2023)
73. Yang, J., Li, T., Fan, L., Tian, Y., Wang, Y.: Latent denoising makes good visual tokenizers. arXiv preprint arXiv:2507.15856 (2025)
74. Yu, Q., Weber, M., Deng, X., Shen, X., Cremers, D., Chen, L.C.: An image is worth 32 tokens for reconstruction and generation. Advances in Neural Information Processing Systems **37**, 128940–128966 (2024)
75. Yuan, W., He, Y., Shen, W., Dong, Y., Gu, X., Dong, Z., Bo, L., Huang, Q.: Mogents: Motion generation based on spatial-temporal joint modeling. Advances in Neural Information Processing Systems **37**, 130739–130763 (2024)
76. Yuan, Y., Song, J., Iqbal, U., Vahdat, A., Kautz, J.: Physdiff: Physics-guided human motion diffusion model. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 16010–16021 (2023)
77. Zha, K., Yu, L., Fathi, A., Ross, D.A., Schmid, C., Katabi, D., Gu, X.: Language-guided image tokenization for generation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 15713–15722 (2025)
78. Zhai, Y., Huang, M., Luan, T., Dong, L., Nwogu, I., Lyu, S., Doermann, D., Yuan, J.: Language-guided human motion synthesis with atomic actions. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 5262–5271 (2023)
79. Zhang, B., Sennrich, R.: Root mean square layer normalization. Advances in neural information processing systems **32** (2019)
80. Zhang, J., Zhang, Y., Cun, X., Zhang, Y., Zhao, H., Lu, H., Shen, X., Shan, Y.: Generating human motion from textual descriptions with discrete representations. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14730–14740 (2023)
81. 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)
82. Zhang, M., Guo, X., Pan, L., Cai, Z., Hong, F., Li, H., Yang, L., Liu, Z.: Remodiffuse: Retrieval-augmented motion diffusion model. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 364–373 (2023)
83. Zhang, Z., Gao, H., Liu, A., Chen, Q., Chen, F., Wang, Y., Li, D., Zhao, R., Li, Z., Zhou, Z., et al.: Kmm: Key frame mask mamba for extended motion generation. arXiv preprint arXiv:2411.06481 (2024)
84. Zhang, Z., Wang, Y., Mao, W., Li, D., Zhao, R., Wu, B., Song, Z., Zhuang, B., Reid, I., Hartley, R.: Motion anything: Any to motion generation. arXiv preprint arXiv:2503.06955 (2025)

85. Zhong, C., Hu, L., Zhang, Z., Xia, S.: Attt2m: Text-driven human motion generation with multi-perspective attention mechanism. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 509–519 (2023)
86. Zhou, W., Dou, Z., Cao, Z., Liao, Z., Wang, J., Wang, W., Liu, Y., Komura, T., Wang, W., Liu, L.: Emdm: Efficient motion diffusion model for fast and high-quality motion generation. In: European Conference on Computer Vision. pp. 18–38. Springer (2024)
87. Zhu, L., Wei, F., Lu, Y., Chen, D.: Scaling the codebook size of vq-gan to 100,000 with a utilization rate of 99%. Advances in Neural Information Processing Systems **37**, 12612–12635 (2024)
88. Zhu, L., Liu, X., Liu, X., Qian, R., Liu, Z., Yu, L.: Taming diffusion models for audio-driven co-speech gesture generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10544–10553 (2023)
89. Zhu, W., Ma, X., Ro, D., Ci, H., Zhang, J., Shi, J., Gao, F., Tian, Q., Wang, Y.: Human motion generation: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence (2023)