K-Mask: Kinematic-Aware Masked Modeling for Controllable Text-to-Motion Synthesis

Faisal Ahmed, Chenqiu Zhao, and Anup Basu

University of Alberta, Edmonton, AB, Canada

Abstract. Masked generative models have made significant advances in text-to-motion synthesis. Previous approaches adopt a non-factorized, whole-body tokenization, limiting compositional and fine-grained control. We present K-Mask, a kinematic-aware generative framework that factors motion into anatomically grounded groups and learns temporal and cross-group interactions through masked modeling. In the proposed approach, a kinematic-group residual VQ-VAE (KG-RVQ) encodes motion into disentangled, group-aligned latents. In particular, a latent-aware kinematic dropout (LAKD) loss is proposed to promote within-group reconstruction and suppress inter-group leakage. Then, we introduce two bidirectional spatiotemporal transformers: (i) a base model that generates coarse tokens over a time-group lattice with axis-factorized attention, and (ii) a residual model that predicts deeper quantizer layers for refinement. Finally, We introduce kinematic group masking, which randomly occludes entire group-specific subtrees, to enhance inter-group reasoning. On the HumanML3D and KIT-ML text-to-motion benchmarks, K-Mask obtains competitive FID scores of 0.041 and 0.154, respectively, while providing an anatomically factorized token interface for localized motion control.

Keywords: Text-to-motion · Discrete tokenization · Part-level control

1 Introduction

Text-to-motion generation aims to synthesize physically plausible and temporally coherent 3D human movements that align with free-form language descriptions [17, 34, 44]. Its recent successes are driving adoption in character animation [3, 7], virtual and augmented reality [4, 52], filmmaking [60], and human-robot interaction [17, 23, 32, 35, 56], where practical utility often relies on achieving fine-grained control for specific edits (e.g., walk forward, then raise the right arm) while preserving the global trajectory, coordination, and temporal continuity of the overall sequence [5, 6].

Existing synthesis methods balance fidelity, efficiency, and control. Continuous diffusion-based approaches [11, 13, 45, 57, 58] achieve high fidelity and diversity [23, 32, 57]; however, incur heavy sampling costs, and part-specific edits often require complex inference-time optimization [43, 45, 57]. Discrete methods [16, 18, 36, 42, 54, 56, 59, 61, 62] recast generation as classification over a learned motion vocabulary. While early autoregressive approaches simplify training, their

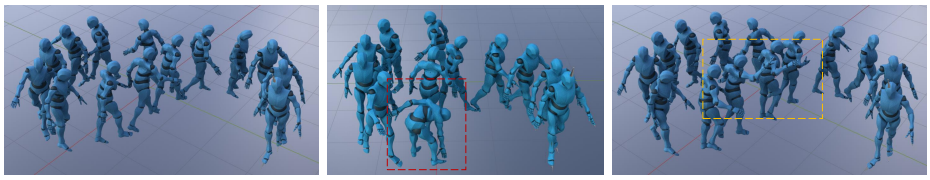

Fig. 1: K-Mask enables both text-to-motion generation and fine-grained spatiotemporal editing. (Left) A base motion is generated from the text prompt "a person walks in a continuous circular pattern." (Middle) We demonstrate temporal inpainting by masking a time window (red box) and applying a new prompt, "a person picks something up from the floor". (Right) We show kinematic (limb-level) editing by masking only the right arm's tokens in a time window (yellow box) and providing a right arm-specific prompt "a person waves with their right arm", our model generates the new arm motion while preserving the original walking motion for all other body parts.

sequential inference leads to error accumulation and cumbersome local edits, while quantization can under-represent limb dynamics [10, 16, 36, 54].

Masked generative models have emerged as a dominant discrete paradigm [16, 35, 36, 54, 62], offering fast parallel sampling and native inpainting [10, 20, 53]. To reduce quantization artifacts, residual VQs (RVQs) are introduced [16, 35, 36, 62], where a residual stack of codebooks first captures the overall structure and subsequent codebooks incrementally add high-frequency motion details. However, most frameworks rely on whole-body, part-agnostic tokenization [16, 36], which is ill-suited to part-level user intents (*e.g.*, raise the left arm) and often causes collateral changes to unrelated limbs during editing. While part-aware tokenization has been explored, existing approaches introduce significant trade-offs. Some methods utilize separate local generators and fusion modules, increasing architectural complexity [62]. Others push granularity to the joint level [54], fragmenting capacity and demanding complex spatiotemporal attention. Furthermore, while prior work has utilized group-specific codebooks (e.g., for trajectory control or ego-centric views [48, 50]), they either deploy part-specific encoder stack or lack explicit mechanisms to prevent information leakage between groups during encoding, which compromises true disentanglement.

Motivated by these observations, we introduce K-Mask, a kinematic-aware masked text-to-motion generation framework that factorizes motion into anatomically meaningful groups and enforces this structure during training. First, a kinematic-group residual VQ-VAE (KG-RVQ) encodes each motion sequence into G group-aligned token streams. Importantly, we obtain part-level tokenization without introducing explicit limb-specific encoder/decoder modules or per-part generator branches. Instead, KG-RVQ operates as a single tokenizer over the full body, and exposes limb-level tokens by (i) allocating the latent representation into kinematic groups and (ii) explicitly discouraging cross-group information sharing. To make this separation reliable under finite latent capacity, we introduce the latent-aware kinematic dropout (LAKD) objective, which promotes within-group reconstruction while penalizing cross-group leakage. Building on

these group-aligned tokens, we train two bidirectional masked transformers: a base model for coarse motion tokens (q=0), and a residual model for refinement (q>0). To maintain efficiency on the resulting time×group lattice, we employ axis-factorized attention [8, 49]. Lastly, we introduce kinematic group masking, which intermittently occludes entire body subtrees (*e.g.*, both legs, left arm and right leg, etc.), explicitly training the model to reason about inter-group dynamics when a subset of groups are masked. In combination, these components support localized token edits while preserving whole-body coordination, narrowing the gap between how users specify actions and how models represent and generate human motion. At inference, the same factorized interface also allows the edit scope to be adjusted, from strict target-group resampling to broader scopes that permit adjacent groups or the root/spine to adapt. Our main contributions are:

- We introduce K-Mask, a unified kinematic-aware generative framework that couples per-group RVQ tokenization (KG-RVQ) with a dedicated anti-leakage objective (LAKD) and a two-stage masked transformer pipeline, enabling part-level control without explicit per-limb encoder/decoder stacks.
- We propose the latent-aware kinematic dropout (LAKD) loss, which encourages within-group fidelity while explicitly penalizing inter-group information leakage in the latent space.
- We design a hybrid masking strategy that combines random token occlusion with structured kinematic group masking, training the model to coordinate across groups while retaining the ability to localize edits.
- On HumanML3D and KIT-ML, K-Mask achieves competitive text-to-motion generation results and supports compositional synthesis and iterative refinement with improved localized control.

2 Related Work

Continuous Generative Methods. Early text-to-motion frameworks focused on regressing continous pose vectors from language, utilizing variational autoencoders (VAEs) [2, 6, 15, 17, 34], recurrent or feed-forward networks [22, 38, 51], transformers [28, 33, 44], or generative adversarial networks (GANs) [9, 27, 46]. While optimizing in the continuous pose space is natural, these models can face challenges in capturing multimodality and supporting fine-grained controllability over long sequences [13, 23, 25, 44, 57]. Diffusion models subsequently became the dominant continuous formulation, offering strong fidelity and diversity, with trade-offs involving expensive sampling process [11, 13, 45, 57, 58]. Among these methods, MDM [45], MoFusion [13], and MotionDiffuse [57] refine text conditioning and long-range modeling, while others incorporate physical [55] and guidance constraints [26] to inject robustness. However, part-level edits often require cross-attention manipulation or additional guidance at inference. Recent work strengthens structure and training signals within diffusion: SALAD [23] builds a skeleton-aware latent space that supports zero-shot editing via attention modulation. A hybridd masked-autoregressive training scheme with refined motion representations has demonstrated consistent gains for text-motion diffusion

under stricter evaluation [32], while retrieval-augmented diffusion [58] injects kinematic neighbors during denoising to improve generalization.

Discrete Quantization and Masked Models. Discrete methods for motion generation compress motion into codes using vector-quantized autoencoders, and then formulates generation as a classification problem over a learned vocabulary. VQ-VAEs [47, 56] and its hierarchical variants [14, 41] established the recipe for learning discrete latents with EMA updates and codebook maintenance. Early two-stage methods (*e.g.*, TM2T [18], T2M-GPT [56], MotionGPT [24]) paired a motion tokenizer with an autoregressive prior, showing competitive quality, exact likelihoods, and fast training, while practical stabilization relied on code reset and large, well-utilized codebooks [18, 56]. Beyond left-to-right decoding, masked token prediction [10, 53]) trains bidirectional transformers to infill random masks, enabling parallel sampling and native inpainting. MoMask [16] adds residual VQ (RVQ) and mask-and-predict decoding for coarse-to-fine synthesis in a few refinements. MMM [36] further simplifies the masked architecture and schedule, yielding strong quality vs. speed trade-offs with built-in editability. BAMM [35] fuses masked modeling with bidirectional autoregression to better exploit context, while offering native variable length generation. Token-space diffusion [12] replaces continuous denoising with categorical transitions, preserving the discrete interface and editability with multi-step sampling.

Part-Aware and Factorized Tokens. Frame-level tokens often obscure limb-level intent and control. Recent part-aware designs aim to address this limitation, however, introduce their own trade-offs. ParCo [62] improves part-conditioned responsiveness via multiple lightweight generators plus a fusion stage, at the cost of extra coordination machinery and specific tokenizers and generator models for each part. MoGenTS [54] pushes granularity to the joint×time grid with 2D masking and attention, however, capacity fragments across many channels and coherence leans on scheduling and attention on joint-speific tokens, significantly increasing computational requirements. KinMo [59] strengthens kinematic-group alignment through group-wise descriptions, yet focuses on aligning motion with more granular text annotations, without rethinking tokenization or masked frameworks. In contrast, we propose to keep a single masked generator but factorize the tokenizer into anatomically grounded groups. By allocating RVQ capacity per kinematic group and training with group-aware masking, our tokens localize edits (e.g., a single arm) without per-part generators, while cross-group attention preserves whole-body coordination. The formulation remains compatible with stronger alignment cues [59] and alternative decoders [12], making part-level intent map directly to localized token edits.

3 Method

Our objective is to learn a conditional distribution $P(\mathbf{X}|\mathbf{c})$ that generates a 3D human pose sequence $\mathbf{X} \in \mathbb{R}^{T \times D}$ from a textual description $\mathbf{c}$, while simul-

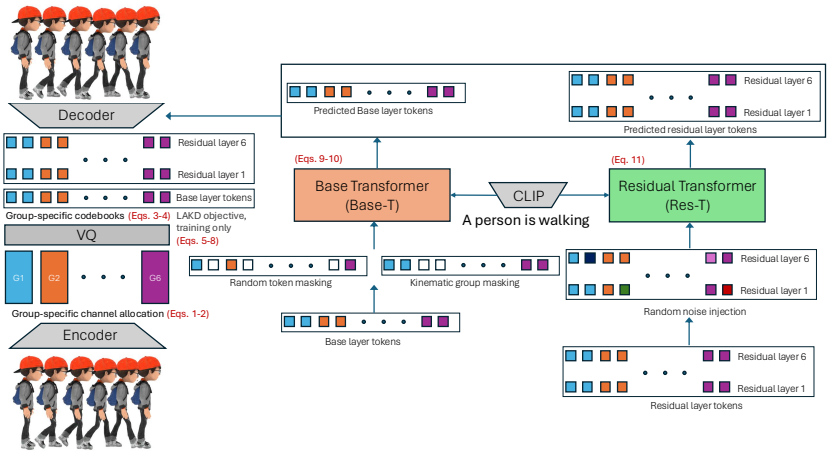

Fig. 2: Overview of the K-Mask framework. First, our kinematic-group RVQ (KG-RVQ) tokenizes motion by partitioning the latent space into G=6 groups, each quantized by independent residual codebooks. Next, a text-conditioned base transformer (Base-T) predicts the coarse base tokens (q=0) using a hybrid strategy of random token masking and kinematic group masking. Finally, a residual transformer (Res-T) progressively refines the motion by predicting the finer residual tokens (q>0) conditioned on the text and all preceding layers.

taneously enabling fine-grained, kinematically-aware control. To achieve this, we introduce K-Mask, a framework that factorizes motion generation onto a spatiotemporal lattice aligned with the human kinematic structure. As illustrated in Figure 2, K-Mask comprises three main stages: (1) a kinematic-group residual VQ-VAE (KG-RVQ) that learns a disentangled, anatomically aligned discrete representation of motion, enforced by a novel latent-aware kinematic dropout (LAKD) objective (Sec. 3.1), (2) a spatiotemporal base transformer (Base-T) that models the coarse motion dynamics (RVQ level q=0) using a hybrid masking strategy (Sec. 3.3), and (3) a spatiotemporal residual transformer (Res-T) that predicts the subsequent RVQ levels (q>0) for refinement (Sec. 3.4).

3.1 Kinematic-Group Residual VQ-VAE

The foundation of K-Mask is a tokenizer capable of representing complex motion in a discrete, structured latent space. We design the KG-RVQ to explicitly enforce this structure by splitting both input motion features and latents into anatomically grounded partitions and quantizing them with separate residual codebooks per group. Consequently, this yields tokens arranged on a time-group lattice that the generator can mask and update with fine granularity.

Kinematic Latent Partitioning. K-Mask achieves kinematic disentanglement by architecturally enforcing a partitioned latent space. We begin by defining an

anatomical factorization of the input motion representation $\mathbf{X} \in \mathbb{R}^{T \times D}$. The motion feature dimension D is partitioned into G disjoint kinematic groups; we adopt $G = 6$ (root, spine, left/right legs, and left/right arms), though the framework is agnostic to kinematic subtree granularity. Let $\mathcal{I}_g$ denote the feature indices for group g, satisfying $\bigcup_g \mathcal{I}_g = \{1, \ldots, D\}$ and $\mathcal{I}_i \cap \mathcal{I}_j = \emptyset$ for $i \neq j$, with $D_g = |\mathcal{I}_g|$ representing the input feature dimension of group g. A convolutional encoder f_θ maps the input $\mathbf{X}$ to a latent representation $\mathbf{Z} \in \mathbb{R}^{T' \times d}$, where T' is the temporally downsampled length. We impose the kinematic structure onto this latent space by explicitly partitioning the latent vector $\mathbf{Z}_t \in \mathbb{R}^d$ at time t along the channel dimension:

$$\mathbf{Z}_t = \left[\mathbf{Z}_t^{(1)} \,\|\, \cdots \,\|\, \mathbf{Z}_t^{(G)} \right], \qquad \mathbf{Z}_t^{(g)} \in \mathbb{R}^{d_g}, \tag{1}$$

with the motivation to impose an inductive bias toward disentanglement, supporting the specialized and independent processing of each kinematic stream. We draw our inspiration from parametric audio coding [21, 39], where complex signals are partitioned for efficient, high-fidelity quantization. To fully exploit this factorization, the latent capacity d should reflect the non-uniform structure of human motion (e.g., differing degrees of freedom in the spine vs. the root). Hence, instead of employing a uniform split ($d_g = d/G$), we adopt the concept of efficient bit allocation principles in perceptual coding [29], utilizing proportional allocation:

$$d_g \propto D_g, \quad \text{s.t.} \sum_{g=1}^{G} d_g = d, \quad d_g \geq d_{\min}. \tag{2}$$

Here, the latent dimension d_g for each group is set proportional to its input feature dimension D_g with a minimum capacity $d_{\min}$ (e.g., D/16) and efficiency constraints (e.g., multiples of 8). This ensures that representational capacity is concentrated where the signal complexity lies, yielding G latent sub-streams $\mathbf{Z}^{(g)}$ for group-wise quantization.

Grouped Residual Quantization. We employ G independent residual vector quantization (RVQ) stacks, one for each latent sub-stream. Each stack g consists of Q codebooks, $\mathcal{C}_{g,q} = \{\mathbf{e}_{g,q,k}\}_{k=1}^{K}$, where K is the codebook size. The quantization process for group g proceeds iteratively. Starting with the residual $\mathbf{r}_{t,0}^{(g)} = \mathbf{Z}_t^{(g)}$, at each level q:

$$k_{t,q}^{(g)} = \arg\min_k \left\| \mathbf{r}_{t,q}^{(g)} - \mathbf{e}_{g,q,k} \right\|_2^2, \tag{3}$$

$$\hat{\mathbf{Z}}_{t,q}^{(g)} = \mathbf{e}_{g,q,k_{t,q}^{(g)}}, \quad \mathbf{r}_{t,q+1}^{(g)} = \mathbf{r}_{t,q}^{(g)} - \hat{\mathbf{Z}}_{t,q}^{(g)}. \tag{4}$$

The quantized representation for group g is the sum: $\hat{\mathbf{Z}}_t^{(g)} = \sum_{q=0}^{Q-1} \hat{\mathbf{Z}}_{t,q}^{(g)}$. The final quantized latent $\hat{\mathbf{Z}}_t$ is the concatenation of the group latents, $\hat{\mathbf{Z}}_t = [\hat{\mathbf{Z}}_t^{(1)} \| \cdots \| \hat{\mathbf{Z}}_t^{(G)}]$, which is passed to a convolutional decoder g_ϕ to reconstruct the motion $\hat{\mathbf{X}}$. The resulting token indices are structured as a tensor $\mathbf{I} \in \{1, \ldots, K\}^{T' \times G \times Q}$.

Latent-Aware Kinematic Dropout (LAKD). While the architecture encourages separation, an explicit objective is required to enforce disentanglement. We introduce the latent-aware kinematic dropout (LAKD) loss. LAKD operates directly on the quantized latent space $\hat{\mathbf{Z}}$, actively encouraging the encoder f_θ to produce disentangled features. Let $\mathbf{M}_g$ be a binary mask in the latent space that selects only the channels corresponding to group g. During training, we randomly sample a group g and construct two complementary partial latent representations:

$$\hat{\mathbf{Z}}_{\text{only } g} = \hat{\mathbf{Z}} \odot \mathbf{M}_g, \quad \hat{\mathbf{Z}}_{\text{except } g} = \hat{\mathbf{Z}} \odot (1 - \mathbf{M}_g). \tag{5}$$

Here, $\odot$ represents element-wise product. Next, let $P_g(\cdot)$ denote the projection operator that extracts features for group g from the reconstructed motion space, and $P_{\bar{g}}(\cdot)$ be its complement. We define the LAKD loss as a combination of three terms designed to ensure the self-sufficiency of group latents and block information leakage:

$$\mathcal{L}_{\text{LAKD}}(g) = \mathcal{L}_{\text{pos}}(g) + \mathcal{L}_{\text{neg-in}}(g) + \mathcal{L}_{\text{neg-out}}(g). \tag{6}$$

The terms are defined as follows (using the ℓ_1 norm, averaged over the original sequence length T):

$$\begin{aligned} \mathcal{L}_{\text{pos}}(g) &= \|P_g(g_\phi(\hat{\mathbf{Z}}_{\text{only } g})) - P_g(\mathbf{X})\|_1, \\ \mathcal{L}_{\text{neg-in}}(g) &= \|P_g(g_\phi(\hat{\mathbf{Z}}_{\text{except } g}))\|_1, \\ \mathcal{L}_{\text{neg-out}}(g) &= \|P_{\bar{g}}(g_\phi(\hat{\mathbf{Z}}_{\text{only } g}))\|_1. \end{aligned} \tag{7}$$

In particular, positive reconstruction ($\mathcal{L}_{\text{pos}}$) ensures that the latents for group g are sufficient to reconstruct the motion of that group. In-group suppression ($\mathcal{L}_{\text{neg-in}}$) penalizes other groups' latents from contributing to the motion of group g. Out-group suppression ($\mathcal{L}_{\text{neg-out}}$) penalizes group g's latents from contributing to the motion of other groups. The overall KG-RVQ objective combines the LAKD loss with standard reconstruction ($\mathcal{L}_{\text{rec}}$, smooth-ℓ_1) and commitment losses ($\mathcal{L}_{\text{commit}}$):

$$\mathcal{L}_{\text{KG-RVQ}} = \mathcal{L}_{\text{rec}} + \lambda_{\text{com}}\mathcal{L}_{\text{commit}} + \lambda_{\text{LAKD}}\mathbb{E}_g[\mathcal{L}_{\text{LAKD}}(g)]. \tag{8}$$

3.2 Spatiotemporal Transformers for Generation

We employ a transformer architecture tailored to the factorized structure of the KG-RVQ tokens, which exist on a time-group lattice (t, g). Both the Base-T and Res-T utilize this architecture. To maximize parameter efficiency and facilitate the sharing of motion primitives across limbs, we employ a shared token embedding layer for all groups. Spatial awareness is injected via learned group embeddings $\mathbf{P}_g$. The input representation $\mathbf{H} \in \mathbb{R}^{T' \times G \times d_m}$ is constructed by combining token embeddings, temporal positional encodings $\mathbf{P}_t$, group embeddings, and the CLIP [40] text condition embedding:

$$\mathbf{H}_{t,g} = \text{Embed}(\mathbf{I}_{t,g}) + \mathbf{P}_t + \mathbf{P}_g + \text{Proj}(\text{CLIP}(\mathbf{c})), \tag{9}$$

modeling the full $T' \times G$ sequence with standard self-attention incurs a complexity of $\mathcal{O}((T'G)^2)$. We adopt a factorized temporal-then-spatial (T→S) attention architecture, which first models the dynamics of individual limbs and then coordinates them. In the temporal block, attention is applied independently for each group across the time axis. In the spatial block, attention is applied across the group axis for each timestep. This reduces the complexity to $\mathcal{O}(T'^2G + T'G^2)$.

3.3 Base Transformer (Base-T): Coarse Generation

The Base-T models the joint distribution of the coarse motion tokens (level q=0), denoted $\mathbf{I}^{(0)} \in \{1, \ldots, K\}^{T' \times G}$, conditioned on the text $\mathbf{c}$. We train the Base-T using the MaskGIT objective [10]. A key contribution is a hybrid masking strategy that combines random token masking with structured, kinematically-aware masking to explicitly teach the model limb coordination. During training, for each sequence, we select a masking strategy:

- **Kinematic masking** (with probability p_{kin}): We mask the tokens corresponding to a specific subset of kinematic groups for the entire sequence duration. To ensure the model learns diverse coordination patterns, we randomly sample the subset from various regimes, including: (i) symmetric groups: masking symmetric pairs (e.g., both legs), (ii) upper/lower body: masking the entire upper or lower body, and (iii) random subset: masking a random combination of groups.
- **Random token masking** (with probability $1 - p_{kin}$): We sample a masking ratio γ from a cosine schedule and randomly mask $\gamma \cdot (T'G)$ tokens across the entire spatiotemporal lattice.

Let $\mathcal{M} \in \{0, 1\}^{T' \times G}$ be the resulting binary mask, where $\mathbf{M}_{t,g}$=1 indicates a masked position, and $\bar{\mathbf{M}}_{t,g}$ is its complement, with $\bar{\mathbf{M}}_{t,g} \cup \mathbf{M}_{t,g} = \mathbf{1}$. . The model takes a corrupted sequence $\mathbf{I} \odot \bar{\mathbf{M}}_{t,g}$ as input. The objective is to predict the original tokens for these positions:

$$\mathcal{L}_{\text{Base-T}} = \mathbb{E}_{\mathbf{I}}\big[-\sum_{t,g} \log P(\mathbf{I} \odot \mathbf{M}_{t,g} | \mathbf{I} \odot \bar{\mathbf{M}}_{t,g}, \mathbf{c}) \big], \tag{10}$$

where $P(\mathbf{I} \odot \mathbf{M}_{t,g} | \mathbf{I} \odot \bar{\mathbf{M}}_{t,g}, \mathbf{c})$ is the probability distribution predicted by the Base-T model, $\odot$ denotes element-wise product. Lastly, we also employ classifier-free guidance by randomly dropping the condition $\mathbf{c}$ during training.

3.4 Residual Transformer (Res-T): Refinement

The Res-T generates the finer details by predicting the remaining RVQ levels $q \in \{1, \ldots, Q-1\}$. To predict the tokens at level q, $\mathbf{I}^{(q)}$, the Res-T conditions on the text $\mathbf{c}$ and the tokens from all preceding levels, $\mathbf{I}^{(<q)}$. We dequantize the history tokens back into the latent space to provide a continuous representation:

$$\hat{\mathbf{Z}}_t^{(g,<q)} = \sum_{p=0}^{q-1} \mathbf{e}_{g,p}[\mathbf{I}_{t,g}^{(p)}] \tag{11}$$

where $\mathbf{e}$ denotes the embedding vector in codebook, with the input of $\mathbf{I}_{t,g}^{(p)}$. These history latents are projected and provided as input to the transformer along with a learned embedding for the target level q. We train the Res-T by randomly sampling the target level q, with random noise injection into the base and residual history tokens to make Res-T robust against exposure bias during real inference task.

3.5 Inference and Kinematic Control

Generation. Inference proceeds hierarchically. First, the Base-T generates $\hat{\mathbf{I}}^{(0)}$ using iterative refinement over S steps, guided by the CFG scale w. Second, the Res-T sequentially generates the remaining levels $\hat{\mathbf{I}}^{(q)}$ for $q = 1$ to $Q - 1$. Finally, the complete token tensor $\hat{\mathbf{I}}$ is decoded by the KG-RVQ decoder g_ϕ.

Kinematic Editing. The factorization of the latent space into the $T' \times G$ lattice natively enables fine-grained editing. To edit a specific part (e.g., the left arm, group g), we can freeze the tokens for all other groups $\bar{g}$ and resample only the tokens for group g using the Base-T and Res-T, conditioned on the original text or a modified prompt. This allows localized changes while preserving the overall context. In practice, this strict target-only setting can be relaxed when an edit requires biomechanical compensation: the edit scope can be expanded to include the root, spine, or adjacent groups, allowing these parts to adapt coherently to the edit prompt.

4 Experiments

We experimentally validate the K-Mask framework for high-fidelity, controllable text-to-motion synthesis. Following a description of the setup (Sec. 4.1), we provide comprehensive quantitative and qualitative comparisons against state-of-the-art methods on established benchmarks and a matched localized editing setting (Sec. 4.2). We then conduct detailed ablation studies to analyze the contribution of each component (Sec. 4.3).

4.1 Experimental Setup

Datasets and Evaluation Metrics. We evaluate on two publicly-available motion-language benchmarks: HumanML3D [18] and KIT-ML [37]. HumanML3D aggregates 14,616 motions from AMASS [30] and HumanAct12 [19], each paired with ≈3 textual descriptions (44,970 in total). Motions are standardized to 20 FPS with a maximum duration of 10s. KIT-ML provides 3,911 motions and 6,278 descriptions sourced from the KIT [31] and the CMU [1] databases, with a mean duration of 10.33 ± 13.38 s. All sequences are resampled to 12.5 FPS. For both datasets, we use the T2M pose representation [18], apply left-right mirroring for augmentation, and adopt a 0.8/0.15/0.05 train/test/val split. We

adopt the standard evaluation protocol [16, 32, 45], using the pre-trained motion feature extractor of [17] to compute all metrics. We assess generation fidelity and realism with *Fréchet Inception Distance (FID)*. Text-motion alignment is measured by *R-Precision (Top-k)* and *Multimodal Distance*. Finally, we report *Multimodality (MM)*, the average feature distance between outputs generated from the same text prompt, and *Diversity (Div.)*, the average feature distance between generated motions sampled from different text prompts.

To evaluate controllable editing, we define a fixed 200-case HumanML3D editing evaluation set from the test split, covering eight edit classes with 25 cases each: left/right arm raise, right-hand wave, left/right knee lift, right-leg kick, bend down, and sit down. Each case contains a real source motion, its original caption, an edit prompt, a predefined edit window, and a fixed random seed. The edit window covers approximately 35–60% of the sequence, with boundaries rounded when needed to match the temporal resolution used by the motion tokenizers. All methods are evaluated on the same source motions, edit windows, prompts, seeds, and scoring code, while retaining each method's native editing procedure and sampling budget. Whole-body tokenizers such as MoMask [16] and BAMM [35] inpaint all motion tokens within the edit window, whereas K-Mask uses strict group-local editing by resampling only the kinematic group(s) relevant to the edit prompt and freezing non-target groups. ParCo [62] does not provide a matched localized-editing implementation, hence we evaluate it as editing-as-regeneration and interpret preservation metrics accordingly. MoGenTS is evaluated using spatial-temporal masked editing as presented in [54]. We report *Target Success Rate (TSR)*, *R@3*, *Spillover*, *Drift*, and *Skate*. *TSR* is the fraction of cases that pass a class-specific deterministic detector on recovered 22-joint motions; for example, an arm-raise edit must lift the target wrist above the head/neck by a body-normalized margin for a sustained portion of the edit window, while wrong-side and no-op edits are counted as failures. Limb actions are evaluated in a body-local frame to factor out global translation and turning, whereas bending and sitting use source-relative world-frame measurements. The detectors are fixed before comparison and reverse-checked on unrelated source clips, yielding a 2.5% false-positive floor. *R@3* is computed with the standard HumanML3D evaluator in intra-edit batches using other edited motions as negatives. *Spillover* measures the mean body-local displacement of non-target joints within the edit window: $spillover = \frac{1}{|\mathcal{W}||\bar{\mathcal{J}}_a|} \sum_{t\in\mathcal{W}} \sum_{j\in\bar{\mathcal{J}}_a} \|\tilde{\mathbf{p}}^{\text{loc}}_{t,j} - \mathbf{p}^{\text{loc}}_{t,j}\|_2$, where $\mathcal{W}$ is the edit window, $\bar{\mathcal{J}}_a$ denotes non-target joints for edit class a, and $\mathbf{p}^{\text{loc}}$ and $\tilde{\mathbf{p}}^{\text{loc}}$ are source and edited joint positions in the body-local frame. *Drift* measures the mean world-frame pelvis trajectory deviation from the source within the edit window and *Skate* measures horizontal foot displacement during detected ground contacts.

Implementation Details. We implement our framework in PyTorch. The KG-RVQ is trained for 50 epochs (batch size =256) and utilizes 1D temporal convolutions with $G = 6$ kinematic groups, $d = 512$, and $Q = 6$ residual quantization layers (codebook $K = 512$). The 8-layer Base-T and Res-T transformers use axis-factorized attention (temporal-spatial, $d_m = 512$), are conditioned on CLIP

Table 1: Quantitative evaluation on the HumanML3D and KIT-ML test set. ± indicates a 95% confidence interval. Metrics were calculated with 20 repeats and the average is reported. **Bold** face indicates the best result, while underscore shows the second best.

Datasets	Methods	R Precision ↑			FID ↓	MultiModal Dist ↓	MultiModality ↑	Div. →
		Top 1	Top 2	Top 3				
HumanML3D	TM2T [18]	$0.424^{\pm.003}$	$0.618^{\pm.003}$	$0.729^{\pm.002}$	$1.501^{\pm.017}$	$3.467^{\pm.011}$	$2.424^{\pm.093}$	$8.589^{\pm.076}$
	T2M [17]	$0.455^{\pm.003}$	$0.636^{\pm.003}$	$0.736^{\pm.002}$	$1.087^{\pm.021}$	$3.347^{\pm.008}$	$2.219^{\pm.074}$	$9.175^{\pm.083}$
	MDM [45]	-	-	$0.611^{\pm.007}$	$0.544^{\pm.044}$	$5.566^{\pm.027}$	$\mathbf{2.799^{\pm.072}}$	$9.559^{\pm.086}$
	MLD [11]	$0.481^{\pm.003}$	$0.673^{\pm.003}$	$0.772^{\pm.002}$	$0.473^{\pm.013}$	$3.196^{\pm.010}$	$2.413^{\pm.079}$	$9.724^{\pm.082}$
	MotionDiffuse [57]	$0.491^{\pm.001}$	$0.681^{\pm.001}$	$0.782^{\pm.001}$	$0.630^{\pm.001}$	$3.113^{\pm.001}$	$1.553^{\pm.042}$	$9.410^{\pm.049}$
	T2M-GPT [56]	$0.492^{\pm.003}$	$0.679^{\pm.002}$	$0.775^{\pm.002}$	$0.141^{\pm.005}$	$3.121^{\pm.009}$	$1.831^{\pm.048}$	$9.761^{\pm.081}$
	ReMoDiffuse [58]	$0.510^{\pm.005}$	$0.698^{\pm.006}$	$0.795^{\pm.004}$	$0.103^{\pm.004}$	$2.974^{\pm.016}$	$1.795^{\pm.043}$	$9.018^{\pm.075}$
	MotionGPT [42]	$0.492^{\pm.003}$	$0.681^{\pm.003}$	$0.778^{\pm.002}$	$0.232^{\pm.008}$	$3.096^{\pm.008}$	$2.008^{\pm.084}$	9.006
	MoMask [16]	$0.521^{\pm.002}$	$0.713^{\pm.002}$	$0.807^{\pm.002}$	$0.045^{\pm.002}$	$2.958^{\pm.008}$	$1.241^{\pm.040}$	-
	BAMM [35]	$0.525^{\pm.002}$	$0.720^{\pm.003}$	$0.814^{\pm.003}$	$0.055^{\pm.002}$	$2.919^{\pm.008}$	$1.687^{\pm.051}$	$9.717^{\pm.089}$
	MMM [36]	$0.504^{\pm.003}$	$0.696^{\pm.003}$	$0.794^{\pm.002}$	$0.080^{\pm.003}$	$2.998^{\pm.007}$	$1.164^{\pm.041}$	$9.411^{\pm.058}$
	ParCo [62]	$0.515^{\pm.003}$	$0.706^{\pm.003}$	$0.801^{\pm.002}$	$0.109^{\pm.005}$	$2.927^{\pm.008}$	$1.382^{\pm.060}$	$9.576^{\pm.088}$
	MARDM-SiT [32]	$0.500^{\pm.004}$	$0.695^{\pm.003}$	$0.795^{\pm.003}$	$0.114^{\pm.007}$	-	$2.231^{\pm.071}$	-
	K-Mask (ours)	$\mathbf{0.531}^{\pm.003}$	$\mathbf{0.735}^{\pm.003}$	$\mathbf{0.824}^{\pm.002}$	$\mathbf{0.041}^{\pm.003}$	$\mathbf{2.917}^{\pm.006}$	$1.415^{\pm.057}$	$9.671^{\pm.074}$
KIT-ML	TM2T [18]	$0.280^{\pm.005}$	$0.463^{\pm.006}$	$0.587^{\pm.005}$	$3.599^{\pm.153}$	$4.591^{\pm.026}$	$\mathbf{3.292}^{\pm.081}$	$9.473^{\pm.117}$
	T2M [17]	$0.361^{\pm.005}$	$0.559^{\pm.007}$	$0.681^{\pm.007}$	$3.022^{\pm.107}$	$3.488^{\pm.028}$	$2.052^{\pm.107}$	$10.72^{\pm.145}$
	MDM [45]	-	-	$0.396^{\pm.004}$	$0.497^{\pm.021}$	$9.191^{\pm.022}$	$1.907^{\pm.214}$	$10.85^{\pm.109}$
	MLD [11]	$0.390^{\pm.008}$	$0.609^{\pm.008}$	$0.734^{\pm.007}$	$0.404^{\pm.027}$	$3.204^{\pm.027}$	$2.192^{\pm.071}$	$10.80^{\pm.117}$
	MotionDiffuse [57]	$0.417^{\pm.004}$	$0.621^{\pm.004}$	$0.739^{\pm.004}$	$1.954^{\pm.062}$	$2.958^{\pm.005}$	$0.730^{\pm.013}$	$11.10^{\pm.143}$
	T2M-GPT [56]	$0.416^{\pm.006}$	$0.627^{\pm.006}$	$0.745^{\pm.006}$	$0.514^{\pm.029}$	$3.007^{\pm.023}$	$1.570^{\pm.039}$	$10.86^{\pm.094}$
	ReMoDiffuse [58]	$0.427^{\pm.014}$	$0.641^{\pm.004}$	$0.765^{\pm.005}$	$0.155^{\pm.006}$	$2.814^{\pm.012}$	$1.239^{\pm.028}$	$10.80^{\pm.105}$
	MotionGPT [42]	$0.415^{\pm.010}$	$0.634^{\pm.011}$	$0.760^{\pm.005}$	$0.372^{\pm.020}$	$2.931^{\pm.041}$	$1.097^{\pm.054}$	10.54
	MoMask [16]	$0.433^{\pm.007}$	$0.656^{\pm.005}$	$0.781^{\pm.005}$	$0.204^{\pm.011}$	$2.779^{\pm.022}$	$1.131^{\pm.043}$	-
	BAMM [35]	$0.438^{\pm.009}$	$0.661^{\pm.009}$	$0.788^{\pm.005}$	$0.183^{\pm.013}$	$2.723^{\pm.026}$	$1.609^{\pm.065}$	$11.008^{\pm.094}$
	MMM [36]	$0.381^{\pm.005}$	$0.590^{\pm.006}$	$0.718^{\pm.005}$	$0.429^{\pm.019}$	$3.146^{\pm.019}$	$1.105^{\pm.026}$	$10.633^{\pm.097}$
	ParCo [62]	$0.430^{\pm.004}$	$0.649^{\pm.007}$	$0.772^{\pm.006}$	$0.453^{\pm.027}$	$2.820^{\pm.028}$	$1.245^{\pm.022}$	$10.95^{\pm.094}$
	MARDM-SiT [32]	$0.387^{\pm.006}$	$0.610^{\pm.006}$	$0.749^{\pm.006}$	$0.242^{\pm.014}$	-	$1.312^{\pm.053}$	-
	K-Mask (ours)	$\mathbf{0.455}^{\pm.006}$	$\mathbf{0.682}^{\pm.006}$	$\mathbf{0.793}^{\pm.006}$	$\mathbf{0.154}^{\pm.025}$	$\mathbf{2.671}^{\pm.020}$	$1.188^{\pm.038}$	$10.993^{\pm.096}$

(ViT-B/32), and are trained for 2,000 epochs (batch size = 64). All models use the AdamW optimizer with a cosine learning rate schedule, starting at 2×10^{-4}. For training, we use a kinematic masking probability $p_{kin} = 0.25$, and at inference, we employ 18 refinement steps and a classifier-free guidance (CFG) scale as recommended in [16].

4.2 Comparison with Existing Methods

Quantitative Evaluation for Text-to-Motion. We compare K-Mask with prior text-to-motion methods, including continuous diffusion-based and discrete masked/quantized approaches. Table 1 summarizes the results on HumanML3D and KIT-ML. Overall, K-Mask achieves strong and consistent performance across metrics: it obtains the best FID on HumanML3D (0.041) and KIT-ML (0.154), and it ranks highest in R-Precision across Top-k on both datasets (e.g., 0.531 Top-1 on HumanML3D). We also observe the best MultiModal Distance on HumanML3D (2.917) and KIT-ML (2.671). Taken together, these results suggest that the proposed representation and training scheme can improve generation fidelity while maintaining (and in several cases improving) text-motion alignment.

Table 2: Matched HumanML3D localized editing evaluation. *FID* is the standard HumanML3D test-set generation metric included for context. *TSR*, *R@3*, *Spill.*, *Drift*, and *Skate* are measured on the 200-case editing evaluation set using the same source motions, edit windows, prompts, seeds, and scoring code. Each method is evaluated through its available editing mechanism.

Method	FID ↓	TSR ↑	R@3 ↑	Spill. ↓	Drift ↓	Skate ↓
MoMask [16]	.045	20.0	26.6	.164	.74	.009
BAMM [35]	.055	7.0	21.9	.136	1.33	.009
ParCo* [62]	.109	11.5	27.6	.190	1.15	.016
MoGenTS [54]	**.028**	19.0	25.5	.168	.92	.018
K-Mask	.041	**22.5**	**28.4**	**.108**	**.41**	**.002**

*ParCo is evaluated as editing-as-regeneration because its released implementation does not expose matched localized editing.

Quantitative Localized Editing. Standard text-to-motion metrics capture global realism and text alignment, yet they do not isolate the core editing trade-off: applying a requested local change while preserving the surrounding motion. Table 2 reports results on our 200-case HumanML3D editing evaluation set. Under this protocol, K-Mask obtains the highest *TSR* and *R@3*, suggesting that the requested edits are reflected in both the geometric success criterion and the retrieval-based semantic measure, rather than arising from overly conservative or no-op outputs. The absolute *TSR* values remain modest, which reflects the difficulty of localized semantic motion editing; nevertheless, K-Mask achieves strong target-edit performance under the same cases and scoring protocol. For preservation, K-Mask obtains the lowest *Spillover*, *Drift*, and *Skate*. These gains are consistent with its anatomically addressable token interface: in the strict editing setting, only the target kinematic group set is resampled, while non-target groups are kept fixed. Importantly, the preservation improvement does not come from simply avoiding the edit, since K-Mask also leads *TSR* and *R@3*. MoGenTS achieves the lowest standard HumanML3D *FID*, so we interpret these results as showing that K-Mask remains competitive for standard generation while providing a stronger editability–preservation trade-off for localized editing.

Qualitative Evaluation for Text-to-Motion and Limb Edits. Figure 3 provides qualitative comparisons on multi-action prompts that require long-horizon coherence. For "walks forward, bends down, picks up something, then keeps walking," K-Mask produces a single coherent pick-up transition while maintaining the surrounding walk, whereas MoMask [16] and BAMM [35] often repeat or misplace the interaction. For "walks in a circle, then sits down," K-Mask typically performs the correct action ordering, while MoMask may drift to a different action (e.g., lying) and BAMM may insert the sit prematurely.

For limb edits (Figure 4), K-Mask leverages kinematic group-level tokenization to resample the target group while keeping other groups fixed. In the "walking in a circle" → "raises their right hand" edit, K-Mask primarily changes the right arm while preserving the original trajectory and overall motion, whereas MoMask

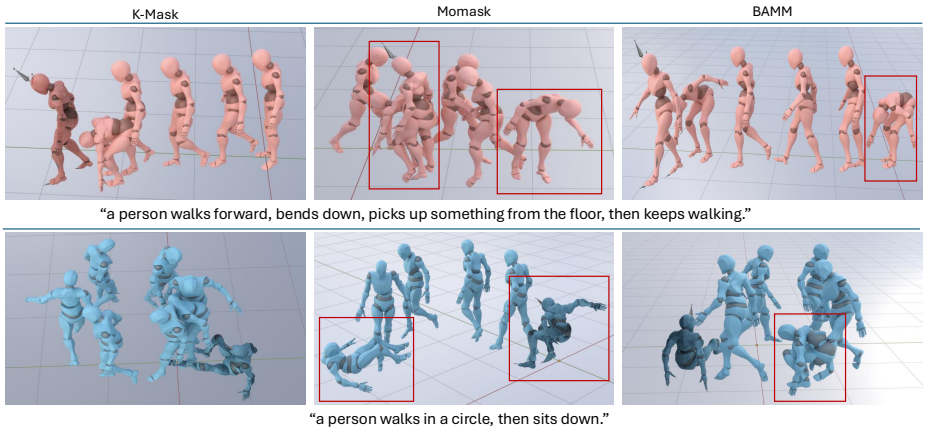

Fig. 3: Qualitative comparison of text-to-motion generation. Only keyframes are shown. Compared to previous methods, K-Mask produces higher-quality motions and demonstrates a superior understanding of complex prompts involving multiple actions. Red bounding boxes highlight inconsistencies in the generated motion sequences.

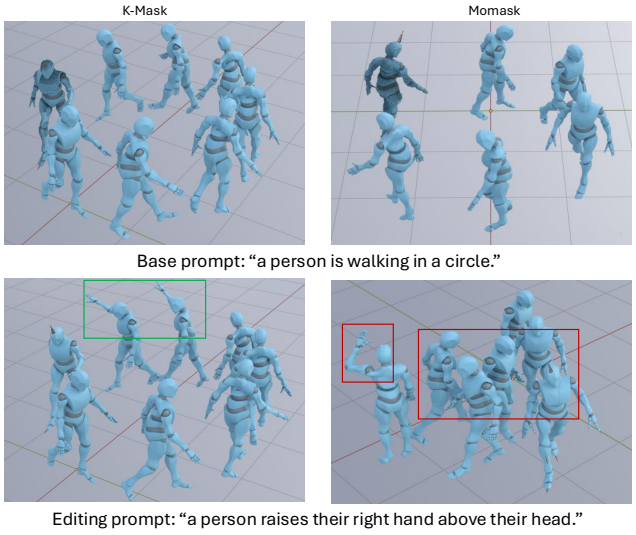

Fig. 4: K-Mask vs. Momask on limb-level editing. Both start with a base motion from "a person is walking in a circle" and are then edited with "a person raises their right hand above their head." The red boxes highlight inconsistencies (wrong arm) and trajectory shifts in the Momask result, while K-Mask (green box) applies the edit without any side-effect (green bounding box).

Table 3: Grouping and LAKD sensitivity on HumanML3D. All variants use $d = 384$, $Q = 6$, and a matched 30-epoch short-run budget; the table is intended for relative sensitivity ablation. *Leak.* denotes the cross-group leakage ratio.

Setting	G	Alloc.	λ	$\text{FID}_{\text{rec}} \downarrow$	MPJPE↓	Leak.↓
Whole-body	1	–	–	.033	30.6	–
Coarse	3	prop.	2.0	.010	17.9	.071
No LAKD	6	prop.	0.0	.008	16.4	2.484
Low λ	6	prop.	0.5	.009	18.7	.184
High λ	6	prop.	4.0	.015	19.5	.104
Uniform	6	unif.	2.0	.010	18.5	.106
Joint-level	22	prop.	2.0	.013	20.3	1.489
K-Mask	6	prop.	2.0	.010	17.2	.088

more frequently exhibits leakage (e.g., affecting other limbs or shifting the path). These qualitative results align with the localized editing metrics in Section 4.2.

4.3 Ablation Studies

We conduct extensive ablation studies on the HumanML3D dataset to analyze the impact of the key components of K-Mask: the KG-RVQ design and LAKD loss, the generation architecture, and the masking strategy.

Impact of Grouping and LAKD. To analyze the tokenizer design, we vary group granularity, latent-capacity allocation, and the LAKD weight under a matched 30-epoch short-run budget. This study is intended to measure relative sensitivity in reconstruction and cross-group leakage; the main results use the full training schedule. Table 3 shows that the six-group anatomical factorization provides a practical balance between addressability and capacity. The whole-body variant ($G = 1$) removes part-level addressability and gives poor reconstruction, while the coarse split ($G = 3$) is too broad for side-specific limb edits despite its low leakage. On the other hand, joint-level grouping ($G = 22$) fragments capacity and increases leakage, showing that finer granularity is not automatically better. Proportional allocation also improves *MPJPE* relative to a uniform split. The LAKD rows isolate its role: without LAKD, the $G = 6$ tokenizer obtains the best reconstruction but increases leakage from .088 to 2.484, indicating that groups encode information about one another. The chosen weight, $\lambda_{\text{LAKD}} = 2.0$, substantially reduces leakage while preserving reconstruction quality, whereas a larger weight over-regularizes reconstruction without improving the trade-off. Overall, these results support using six anatomical groups with proportional capacity allocation, and characterize LAKD as an anti-leakage regularizer for editability rather than a reconstruction-improving loss.

Impact of Kinematic Masking Strategy. We analyze the impact of our hybrid masking strategy by varying the kinematic masking probability, p_{kin}, as shown in Table 4. A baseline using only random token masking ($p_{kin} = 0.0$) yields the weakest performance, with an FID of 0.108. Introducing a portion of structured

Table 4: Impact of the kinematic masking probability p_{kin} on HumanML3D. *R@1* and *R@3* denote R-Precision at Top-1 and Top-3.

Strategy	R@1 ↑	R@3 ↑	FID ↓
$p_{\text{kin}} = 0.00$ (random only)	0.511	0.802	0.108
$p_{\text{kin}} = 0.10$	0.519	0.815	0.083
$p_{\text{kin}} = 0.25$ (ours)	**0.531**	**0.824**	**0.041**
$p_{\text{kin}} = 0.50$	0.522	0.808	0.078

Table 5: Impact of transformer factorization on HumanML3D. *R@1* and *R@3* denote R-Precision at Top-1 and Top-3.

Attention type	R@1 ↑	R@3 ↑	FID ↓
Temporal–spatial	**0.531**	**0.824**	**0.041**
Spatial–temporal	0.512	0.809	0.077
Full self-attention	0.529	0.822	0.049

kinematic masking significantly improves results, as it explicitly trains the model to reason about inter-group coordination when entire limbs are occluded. We find the optimal balance at $p_{kin} = 0.25$, which achieves the best generative fidelity (FID 0.041) and text alignment (R-Prec Top 3 0.824). Increasing the probability further (e.g., $p_{kin} = 0.50$) begins to degrade performance, confirming that a hybrid approach is superior to using either strategy in isolation.

Impact of Model Architecture. We ablate the architecture of our transformer blocks in Table 5. Our factorized temporal-spatial attention achieves the best performance across all metrics. We observe that the order of attention has some impact on generation quality; reversing the factorization to spatial-temporal attention causes a performance drop. While standard full self-attention performs competitively on R-Precision, our factorized approach is more efficient and achieves slightly better FID.

5 Conclusion

In this work, we introduced K-Mask, a generative framework that bridges the gap between whole-body synthesis and fine-grained, part-level control. We moved beyond part-agnostic tokenization by proposing a kinematic-group RVQ (KG-RVQ) that factorizes motion into anatomically-grounded latents. A novel LAKD loss ensures these latents remain disentangled by actively penalizing information leakage between groups. We then introduced a hybrid kinematic masking strategy to train our spatiotemporal transformers to explicitly reason about inter-group coordination. K-Mask achieves competitive generative fidelity and, crucially, enables intuitive kinematic editing, allowing users to modify specific limbs without corrupting the surrounding motion.

References

1. Cmu graphics lab motion capture database. http://mocap.cs.cmu.edu/, accessed: 2025-11-12
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. 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)
4. Aliakbarian, S., Saleh, F., Collier, D., Cameron, P., Cosker, D.: Hmd-nemo: Online 3d avatar motion generation from sparse observations. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9622–9631 (2023)
5. Athanasiou, N., Cseke, A., Diomataris, M., Black, M.J., Varol, G.: Motionfix: Text-driven 3d human motion editing. In: SIGGRAPH Asia 2024 Conference Papers. pp. 1–11 (2024)
6. Athanasiou, N., Petrovich, M., Black, M.J., Varol, G.: Teach: Temporal action composition for 3d humans. In: 2022 International Conference on 3D Vision (3DV). pp. 414–423. IEEE (2022)
7. 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)
8. Bertasius, G., Wang, H., Torresani, L.: Is space-time attention all you need for video understanding? In: Icml. vol. 2, p. 4 (2021)
9. Cai, H., Bai, C., Tai, Y.W., Tang, C.K.: Deep video generation, prediction and completion of human action sequences. In: Proceedings of the European conference on computer vision (ECCV). pp. 366–382 (2018)
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, 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)
12. Chi, S., Chi, H.g., Ma, H., Agarwal, N., Siddiqui, F., Ramani, K., Lee, K.: M2d2m: Multi-motion generation from text with discrete diffusion models. In: European conference on computer vision. pp. 18–36. Springer (2024)
13. Dabral, R., Mughal, M.H., Golyanik, V., Theobalt, C.: Mofusion: A framework for denoising-diffusion-based motion synthesis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9760–9770 (2023)
14. 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)
15. 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)
16. 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)
17. 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)

18. 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)
19. 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)
20. 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)
21. Herre, J., Faller, C., Ertel, C., Hilpert, J., Hoelzer, A., Spenger, C.: Mp3 surround: Efficient and compatible coding of multi-channel audio. In: Audio Engineering Society Convention 116. Audio Engineering Society (2004)
22. Holden, D., Saito, J., Komura, T.: A deep learning framework for character motion synthesis and editing. ACM Transactions on Graphics (ToG) **35**(4), 1–11 (2016)
23. 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. pp. 7158–7168 (2025)
24. 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)
25. Jin, P., Li, H., Cheng, Z., Li, K., Yu, R., Liu, C., Ji, X., Yuan, L., Chen, J.: Local action-guided motion diffusion model for text-to-motion generation. In: European Conference on Computer Vision. pp. 392–409. Springer (2024)
26. 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)
27. Lin, A.S., Wu, L., Corona, R., Tai, K., Huang, Q., Mooney, R.J.: Generating animated videos of human activities from natural language descriptions. Learning **1**(2018), 1 (2018)
28. Lin, J., Chang, J., Liu, L., Li, G., Lin, L., Tian, Q., Chen, C.w.: Being comes from not-being: Open-vocabulary text-to-motion generation with wordless training. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 23222–23231 (2023)
29. Liu, C.M., Lee, W.C., Chien, C.T., et al.: Bit allocation for advanced audio coding using bandwidth-proportional noise-shaping criterion. In: Proc. 6th International Conference on Digital Audio Effects (DAFx'03) (2003)
30. 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)
31. Mandery, C., Terlemez, Ö., Do, M., Vahrenkamp, N., Asfour, T.: The kit whole-body human motion database. In: 2015 International Conference on Advanced Robotics (ICAR). pp. 329–336. IEEE (2015)
32. 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)
33. 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)

34. 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)
35. 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)
36. 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)
37. Plappert, M., Mandery, C., Asfour, T.: The kit motion-language dataset. Big data **4**(4), 236–252 (2016)
38. Plappert, M., Mandery, C., Asfour, T.: Learning a bidirectional mapping between human whole-body motion and natural language using deep recurrent neural networks. Robotics and Autonomous Systems **109**, 13–26 (2018)
39. Pulkki, V., Karjalainen, M.: Communication acoustics: an introduction to speech, audio and psychoacoustics. John Wiley & Sons (2015)
40. 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)
41. 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)
42. Ribeiro-Gomes, J., Cai, T., Milacski, Z.A., Wu, C., Prakash, A., Takagi, S., Aubel, A., Kim, D., Bernardino, A., De La Torre, F.: Motiongpt: Human motion synthesis with improved diversity and realism via gpt-3 prompting. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 5070–5080 (2024)
43. Shafir, Y., Tevet, G., Kapon, R., Bermano, A.H.: Human motion diffusion as a generative prior. In: International Conference on Learning Representations (ICLR) (2024)
44. 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)
45. Tevet, G., Raab, S., Gordon, B., Shafir, Y., Cohen-Or, D., Bermano, A.H.: Mdm: Human motion diffusion model. In: International Conference on Learning Representations (ICLR) (2023)
46. Tulyakov, S., Liu, M.Y., Yang, X., Kautz, J.: Mocogan: Decomposing motion and content for video generation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1526–1535 (2018)
47. Van Den Oord, A., Vinyals, O., et al.: Neural discrete representation learning. Advances in neural information processing systems **30** (2017)
48. Wan, W., Dou, Z., Komura, T., Wang, W., Jayaraman, D., Liu, L.: Tlcontrol: Trajectory and language control for human motion synthesis. In: European Conference on Computer Vision. pp. 37–54. Springer (2024)
49. Wang, H., Zhu, Y., Green, B., Adam, H., Yuille, A., Chen, L.C.: Axial-deeplab: Stand-alone axial-attention for panoptic segmentation. In: European conference on computer vision. pp. 108–126. Springer (2020)
50. Wang, J., Dabral, R., Luvizon, D., Cao, Z., Liu, L., Beeler, T., Theobalt, C.: Ego4o: Egocentric human motion capture and understanding from multi-modal input. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 22668–22679 (2025)

51. Wang, Z., Yu, P., Zhao, Y., Zhang, R., Zhou, Y., Yuan, J., Chen, C.: Learning diverse stochastic human-action generators by learning smooth latent transitions. In: Proceedings of the AAAI conference on artificial intelligence. vol. 34, pp. 12281–12288 (2020)
52. Yin, T., Hoyet, L., Christie, M., Cani, M.P., Pettré, J.: The one-man-crowd: Single user generation of crowd motions using virtual reality. IEEE Transactions on Visualization and Computer Graphics **28**(5), 2245–2255 (2022)
53. Yu, L., Cheng, Y., Sohn, K., Lezama, J., Zhang, H., Chang, H., Hauptmann, A.G., Yang, M.H., Hao, Y., Essa, I., et al.: Magvit: Masked generative video transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10459–10469 (2023)
54. 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)
55. 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)
56. 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)
57. Zhang, M., Cai, Z., Pan, L., Hong, F., Guo, X., Yang, L., Liu, Z.: Motiondiffuse: Text-driven human motion generation with diffusion model. IEEE transactions on pattern analysis and machine intelligence **46**(6), 4115–4128 (2024)
58. 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)
59. Zhang, P., Liu, P., Garrido, P., Kim, H., Chaudhuri, B.: Kinmo: Kinematic-aware human motion understanding and generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11187–11197 (2025)
60. Zhang, R., Yu, B., Min, J., Xin, Y., Wei, Z., Shi, J.N., Huang, M., Kong, X., Xin, N.L., Jiang, S., et al.: Generative ai for film creation: A survey of recent advances. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 6267–6279 (2025)
61. Zhang, Y., Huang, D., Liu, B., Tang, S., Lu, Y., Chen, L., Bai, L., Chu, Q., Yu, N., Ouyang, W.: Motiongpt: Finetuned llms are general-purpose motion generators. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 7368–7376 (2024)
62. Zou, Q., Yuan, S., Du, S., Wang, Y., Liu, C., Xu, Y., Chen, J., Ji, X.: Parco: Part-coordinating text-to-motion synthesis. In: European Conference on Computer Vision. pp. 126–143. Springer (2024)