ReMoMask: Retrieval-Augmented Masked Motion Generation

Zhengdao Li[1*], Siheng Wang[2*], Zeyu Zhang[1*†], and Hao Tang[1✉]

[1] School of Computer Science, Peking University
[2] Westlake University

https://aigeeksgroup.github.io/ReMoMask/

Abstract. Retrieval-Augmented Text-to-Motion (RAG-T2M) models have demonstrated superior performance over conventional T2M approaches, particularly in handling uncommon and complex textual descriptions by leveraging external motion knowledge. Despite these gains, existing RAG-T2M models remain limited by two closely related factors: coarse-grained text-motion retrieval that overlooks the hierarchical structure of human motion, and underexplored mechanisms for effectively fusing retrieved information into the generative process. In this work, we present ReMoMask, a structure-aware RAG framework for text-to-motion generation that addresses these limitations. To improve retrieval, we propose **H**ierarchical **B**idirectional **M**omentum (HBM) contrastive learning, which employs dual objectives to jointly align global motion semantics and fine-grained part-level features with text. To bridge the gap between structured retrieval and generation, we introduce **T**opology **S**tructured **M**asking (TSM), a training strategy that adaptively masks motion tokens based on semantic relevance, forcing the model to learn robust part-level grounding. Furthermore, we design **S**emantic **S**patial-**T**emporal **A**ttention (SSTA), a topology-aware fusion module that integrates retrieved knowledge via an asymmetric attention mechanism. Extensive experiments on HumanML3D, KIT-ML, and SnapMoGen demonstrate that ReMoMask consistently outperforms prior methods on both text-motion retrieval and text-to-motion generation benchmarks.

Keywords: Text-to-Motion generation · Retrieval-Augmented Generation · Text-motion retrieval

1 Introduction

Human motion generation has attracted increasing attention due to its wide applicability in gaming, film production [31], virtual reality, and robotics. By synthesizing realistic and diverse human motions, these methods aim to significantly reduce the cost of manual animation while improving the efficiency and flexibility of content creation. Among various paradigms, text-to-motion (T2M)

*Equal contribution.
†Project lead.
✉Corresponding author: bjdxtanghao@gmail.com

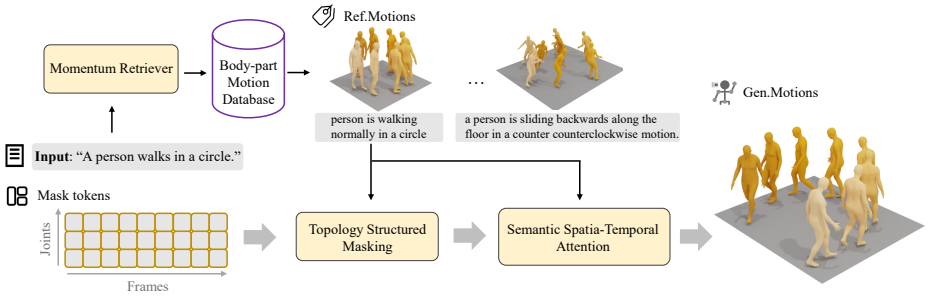


Fig. 1: Overview of ReMoMask.

generation has emerged as a particularly intuitive setting, where a natural language description is directly mapped to a sequence of human joint movements.

Existing T2M methods can be broadly categorized into two lines of research. The first line consists of *conventional T2M models*, which focus on strengthening the generative backbone itself. Representative approaches include generative adversarial networks (GANs) [1–5], variational autoencoders (VAEs) [13], diffusion models [10], visual-language models [14, 30], motion language models [11], and masked generative models [6, 7, 20, 27]. In particular, masked generative models such as MMM [20] and MoMask [7] discretize motion sequences into tokenized representations and perform masked token prediction, achieving high-fidelity and temporally coherent motion synthesis.

The second line of work, known as *retrieval-augmented T2M (RAG-T2M)*, enhances generation by retrieving relevant motion–text pairs from an external database and injecting the retrieved evidence into the generative process. By conditioning generation on exemplar motions, RAG-based approaches improve robustness to uncommon or complex textual inputs. Representative methods include ReMoDiffuse [29], which performs retrieval via text–text similarity using CLIP [22], and ReMoGPT [26], which adopts a cross-modal text–motion retriever.

Despite their promising performance, existing RAG-T2M methods implicitly treat retrieval and fusion as independent modules and largely ignore the *structural consistency* between retrieval alignment and motion representation. Through careful analysis, we identify two fundamental design axes in retrieval-augmented motion generation:

(1) Structural granularity of alignment. As shown in Tab. 1, most text–motion retrieval methods rely on contrastive learning to align global motion embeddings with text [18, 19]. However, human motion is inherently hierarchical, organized by skeletal topology and body-part dependencies. Purely global alignment overlooks fine-grained part-level semantics (e.g., left/right limbs or asymmetric actions), limiting retrieval discriminability. Although some works [26, 32] encode part-level motion features, these representations are typically aggregated without part-level cross-modal alignment, weakening structured correspondence.

Table 1: Comparison of different architecture designs.

Method	Retrieval			Generation	
	Global Align	Part Align	Momentum	Fusion	Latent
MDM [24]	-	-	-	concat	1D
T2M-GPT [28]	-	-	-	concat	1D
MoMask [7]	-	-	-	concat	1D
MARDM [17]	-	-	-	concat	1D
LaMP [15]	-	-	-	cross-attn	1D
TMR [19]	✓	×	×	concat	1D
MoRAG [12]	×	×	×	cross-attn	1D
ReMoGPT [26]	✓	×	×	concat	1D
ReMoDiffuse [29]	✓	×	×	cross-attn	1D
Ours	✓	✓	✓	SSTA	2D

(2) Structural compatibility of fusion. As depicted in Tab. 1, existing RAG-T2M methods often adopt simple concatenation or vanilla cross-attention, without systematically examining how motion latent structure (e.g., 1D vs. 2D spatial–temporal tokens) interacts with fusion mechanisms. This mismatch between retrieved information and motion representation can limit generation quality.

These observations suggest that performance improvements in RAG-T2M are not merely determined by stronger individual modules, but by whether *retrieval alignment and fusion design are structurally consistent with motion topology.*

In this work, we propose **ReMoMask**, a retrieval-augmented masked generative framework for text-to-motion generation. To address alignment granularity, we introduce **H**ierarchical **B**idirectional **M**omentum (HBM) alignment, a structured contrastive learning framework that jointly supervises instance-level (global) and part-level bidirectional text–motion correspondence under a momentum-based paradigm.

To ensure structural compatibility during semantic conditioning, we conduct a systematic study (provided in Sec. 3) on motion latent representations and fusion strategies. Our analysis reveals that preserving motion as 2D spatial–temporal tokens, rather than flattening them into 1D sequences, significantly improves semantic injection and generation stability. Motivated by this observation, we design **S**emantic **S**patial-**T**emporal **A**ttention (SSTA), an attention mechanism tailored to inject retrieved motion semantics into the generative backbone. Meanwhile, to further strengthen structural modeling within this 2D formulation, we adopt a **T**opology **S**tructured **M**asking (TSM) strategy, allowing the network to learn richer topological dependencies across body parts and time.

Our contributions are summarized as follows:

- We identify two key structural design axes—alignment granularity and fusion compatibility—in text-to-motion generation, and demonstrate that enforcing structural consistency across alignment, representation, and fusion is critical for high-quality motion synthesis.

- We propose **HBM**, a hierarchical bidirectional momentum alignment framework that explicitly models both global and part-level text–motion correspondence, enhancing fine-grained semantic grounding.
- We design **SSTA**, a topology-aware spatial–temporal attention mechanism built upon structured 2D motion tokens, together with a **TSM** masking strategy that strengthens topological dependencies across body parts and time.
- Extensive experiments on HumanML3D, KIT-ML, and SnapMoGen show that ReMoMask achieves state-of-the-art performance in both motion generation and cross-modal retrieval, validating the effectiveness of structure-aware masked modeling.

2 Related Work

2.1 Text-to-Motion Generation

Text-to-motion (T2M) generation aims to synthesize realistic human motion sequences conditioned on natural language descriptions. Early approaches explored adversarial learning to establish text–motion correspondence. With the introduction of vector quantization, TM2T [9] and T2M-GPT [28] discretized motion sequences and leveraged autoregressive transformers for semantic control. However, autoregressive decoding often suffers from error accumulation.

Recent advances focus on improving motion representation and generation paradigms. MoMask [7] introduces hierarchical residual quantization and masked bidirectional transformers for parallel decoding, achieving strong performance on HumanML3D. Diffusion-based methods [11,23] further enhance generation quality via non-autoregressive denoising processes. MotionGPT [11] unifies multiple generation tasks under a discrete autoregressive framework.

Beyond backbone improvements, several works investigate fine-grained motion structure modeling. ParCo [32] discretizes whole-body motion into part-level components (limbs, backbone, root) to establish structured priors.

2.2 Retrieval-Augmented Text-to-Motion

Retrieval-augmented generation (RAG) has been extended beyond NLP to multimodal domains, including motion generation [12,16,26,29]. In retrieval-augmented T2M (RAG-T2M), relevant motion–text pairs are retrieved from an external database and injected into the generative backbone to improve robustness under complex or rare textual conditions.

Existing approaches typically rely on global contrastive alignment between text and motion embeddings [18, 19]. For instance, ReMoDiffuse [29] performs retrieval based on text–text similarity to indirectly guide generation, while ReMoGPT [26] introduces part-aware encoders for cross-modal retrieval. However, these methods often aggregate part features without explicit bidirectional cross-modal supervision, leaving fine-grained structural correspondence between text

descriptions and specific body parts underexplored. Furthermore, regarding the integration of retrieved information, prior works commonly adopt feature concatenation or standard cross-attention [26, 29] without systematically considering the compatibility between the retrieved evidence and the underlying motion latent topology. The impact of motion representation structures on fusion effectiveness remains largely uninvestigated.

3 Design Principles of Retrieval-Augmented Motion Generation

While retrieval-augmented T2M methods show promise, the interplay between motion representation and fusion strategy remains underexplored. We investigate two key design axes: (1) *motion latent topology* (1D sequence vs. 2D spatial–temporal grid) and (2) *fusion mechanism* (concatenation vs. cross-attention).

Table 2: Comparison of different latent structures and fusion strategies.

Generator	Latent	Fusion	FID↓	Top1↑
MoMask	1D	concat	$0.057^{\pm.013}$	$0.511^{\pm.003}$
		crossAttn	$\underline{0.043}^{\pm.004}$	$\underline{0.525}^{\pm.003}$
	2D	concat	$0.049^{\pm.007}$	$0.518^{\pm.003}$
		crossAttn	$\mathbf{0.036}^{\pm.005}$	$\mathbf{0.536}^{\pm.002}$

Using MoMask [7] as a backbone generator, we systematically vary these axes while fixing other components. We construct both 1D (z_{1d}) and 2D (z_{2d}) latents, integrating retrieved embeddings (R_m, R_t from ReMoDiffuse [29]) via concatenation or cross-attention. All evaluations are performed on the HumanML3D benchmark.

As summarized in Tab. 2, shifting to 2D latents and adopting cross-attention yields consistent improvements. The optimal configuration combines 2D representation with cross-attention, revealing that effective retrieval injection *hinges on preserving the spatial–temporal topology while enabling structural interaction.* These dual findings directly motivate our **SSTA** module, which is explicitly designed to fuse retrieved semantics with motion tokens via structure-aware cross-attention over a 2D latent grid.

4 Methodology

4.1 Overview

As illustrated in Fig. 1, our framework follows a retrieval-augmented generation paradigm. We first employ **Hierarchical Bidirectional Momentum (HBM)** learning to align text and motion at both instance and part levels, constructing a hierarchically organized embedding space for accurate retrieval. During training, we utilize **Topology Structured Masking (TSM)**, a structure-aware masked modeling objective that adaptively modulates masking probabilities based on hierarchical relevance to strengthen part-level grounding. Given an input text prompt x, relevant motion-text pairs are retrieved from this structured space to obtain reference embeddings (R_m, R_t), which are then injected into the generator

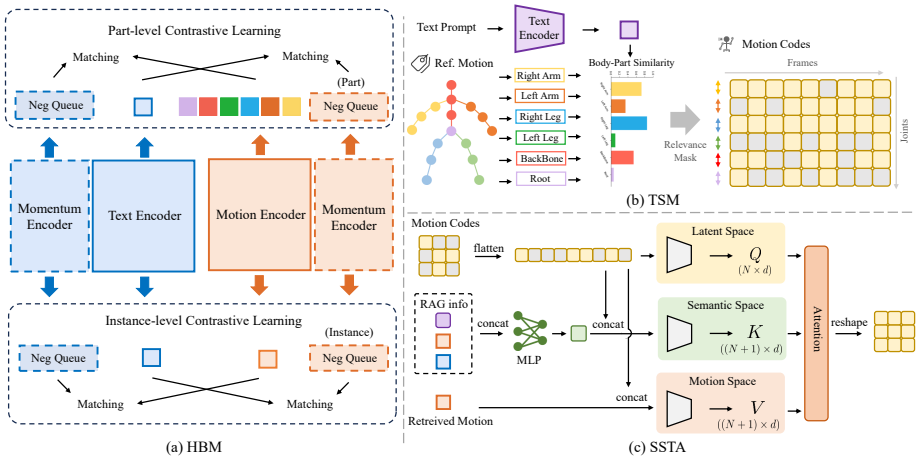


Fig. 2: Framework of ReMoMask. (a) **H**ierarchical **B**idirectional **M**omentum (HBM): Constructs a structured embedding space via instance and part level contrastive alignment with momentum encoders. (b) **T**opology **S**tructured **M**asking (TSM): Generates adaptive masks based on body-part similarity to prioritize reconstructing semantically deficient regions. (c) **S**emantic **S**patial–**T**emporal **A**ttention (SSTA): Injects retrieved semantics via an asymmetric Q-K-V design, where semantics modulate attention (Key) while motion topology is preserved in synthesis (Query/Value).

via our proposed **Semantic Spatial–Temporal Attention (SSTA)** to enable topology-aware semantic conditioning over 2D motion tokens.

4.2 Hierarchical Bidirectional Momentum

Hierarchical Motion Decomposition. As shown in Fig. 2a, we decompose motion m into $K = 6$ semantic parts (e.g., arms, legs, backbone, root), denoted as m^k. Given a batch $\{(x_i, m_i)\}_{i=1}^{B}$, we encode text, global motion, and part motions into a shared space:

$$t_i = f_T(x_i), \quad g_i = f_G(m_i), \quad p_{i,k} = f_P^{(k)}(m_i^k), \tag{1}$$

where f_T, f_G, and $f_P^{(k)}$ are the respective encoders.

Momentum-Based Structured Contrast. To stabilize training and enlarge the negative set, we maintain momentum encoders with parameters updated via exponential moving average:

$$\theta^- \leftarrow \mu\,\theta^- + (1-\mu)\,\theta, \tag{2}$$

where θ and θ^- denote online and momentum parameters. The resulting momentum embeddings $(t_i^-, g_i^-, p_{i,k}^-)$ are stored in a queue $\mathcal{Q}$ serving as negative keys.

Contrastive Objective. Using cosine similarity $\mathrm{sim}(\cdot,\cdot)$ and temperature τ, we define the positive term $\mathcal{P}(q,k) = \exp(\mathrm{sim}(q,k)/\tau)$ and negative sum $\mathcal{N}(q) = \sum_{k^- \in \mathcal{Q}} \exp(\mathrm{sim}(q,k^-)/\tau)$. The InfoNCE loss is:

$$\mathrm{InfoNCE}(q,k) = -\log \frac{\mathcal{P}(q,k)}{\mathcal{P}(q,k) + \mathcal{N}(q)}. \tag{3}$$

Bidirectional Alignment. We enforce bidirectional alignment at the instance level:

$$\mathcal{L}_{\mathrm{Inst}} = \frac{1}{B}\sum_{i=1}^{B}\Big(\mathrm{InfoNCE}(t_i, g_i) + \mathrm{InfoNCE}(g_i, t_i)\Big), \tag{4}$$

and further align text with each part representation to preserve fine-grained semantics:

$$\mathcal{L}_{\mathrm{Part}} = \frac{1}{BK}\sum_{i=1}^{B}\sum_{k=1}^{K}\Big(\mathrm{InfoNCE}(t_i, p_{i,k}) + \mathrm{InfoNCE}(p_{i,k}, t_i)\Big). \tag{5}$$

Final Objective. The total HBM loss combines global and part-level supervision:

$$\mathcal{L}_{\mathrm{HBM}} = \mathcal{L}_{\mathrm{Inst}} + \lambda_P \mathcal{L}_{\mathrm{Part}}. \tag{6}$$

This hierarchical supervision yields a structurally consistent embedding space for precise retrieval.

4.3 Topology Structured Masking

To fully exploit the hierarchical structure learned by HBM during generation, we introduce **Topology Structured Masking (TSM)** as a structure-aware training objective. Unlike conventional uniform random masking that ignores semantic relevance, TSM adaptively modulates masking probabilities based on the hierarchical text–motion alignment established by HBM, ensuring the generator focuses on semantically critical body parts.

Hierarchical Semantic Relevance. As shown in Fig. 2b, given a text embedding t_i and part-level motion embeddings $\{p_{i,k}\}_{k=1}^{K}$ from HBM, we compute part-level relevance weights:

$$\alpha_{i,k} = \mathrm{SoftMax}_k\big(\mathrm{sim}(t_i, p_{i,k})\big), \tag{7}$$

where $\alpha_{i,k}$ reflects the semantic alignment strength between the input text and the k-th body part.

Topology-Aware Mask Distribution. We define the masking probability for each body part as:

$$\pi_{i,k} = \pi_{\mathrm{base}} \cdot (1 - \alpha_{i,k}), \tag{8}$$

where π_{base} is a global masking ratio. This strategy ensures that semantically aligned parts are preserved as structural anchors, while less relevant parts are masked aggressively. By protecting these core semantic features from corruption,

we force the generator to learn how to coordinate global motion conditioned on the key body parts specified by the text, rather than attempting to hallucinate critical semantics from context. These part-level probabilities are then propagated to the 2D spatial–temporal token grid ($J \times T$) according to the skeletal partition, such that all joints belonging to part k share the probability $\pi_{i,k}$.

Structured Masked Modeling Objective. During training, tokens are masked according to $\pi_{i,k}$ and replaced by a learnable mask token. The generator is optimized to reconstruct the original latent codes z:

$$\mathcal{L}_{\text{TSM}} = \mathbb{E}_z \left[\| z_{\text{pred}} - z \|_{\text{masked}} \right]. \tag{9}$$

By aligning the masking distribution with semantic relevance, TSM forces the model to focus on reconstructing critical body parts, thereby enhancing part-level grounding and structural consistency.

4.4 Semantic Spatial-Temporal Attention

Unlike conventional cross-attention that treats retrieved features as uniform tokens, we design **Semantic Spatial–Temporal Attention (SSTA)** to explicitly respect the 2D spatial–temporal organization of motion latents through an asymmetric injection mechanism.

We represent motion as a 2D joint–time token grid, flattened into latent sequences $z \in \mathbb{R}^{B \times N \times d}$ (where $N = T \cdot J$). As illustrated in Fig. 2c, SSTA decouples the roles of semantic conditioning and motion synthesis across the Key and Value pathways.

Query: Motion-Centric Focus. To preserve the structural integrity of the generation process, the Query is derived solely from the current motion latents:

$$Q = W_q z, \tag{10}$$

ensuring that attention computation remains anchored in the spatial–temporal topology of the motion being generated.

Semantic Reference Construction. We compress the multi-modal retrieval context (prompt t, retrieved text R_t, and retrieved motion R_m) into a compact semantic token:

$$h_{\text{sem}} = \text{MLP}(\text{concat}[t;\, R_t;\, R_m]), \tag{11}$$

where $h_{\text{sem}} \in \mathbb{R}^{B \times 1 \times d}$ serves as a global modulation signal rather than a sequence of tokens.

Key: Topology-Aware Gating. The Key pathway integrates both motion states and semantic context to determine *where* to attend:

$$K = W_k \cdot \text{concat}(z,\ h_{\text{sem}}). \tag{12}$$

By injecting h_{sem} only into the Key, semantic information modulates the attention weights without directly overwriting the motion representation.

Table 3: Results of text-to-motion and motion-to-text retrieval benchmark on HumanML3D, KIT-ML, and SnapMoGen datasets. † indicates Results are reproduced by us following the descriptions in the original paper, as no official implementation is publicly available. Best results are highlighted in **bold**, and second-best results are underlined.

Benchmark	Methods	Text-to-motion retrieval						Motion-to-text retrieval					
		R@1↑	R@2↑	R@3↑	R@5↑	R@10↑	MedR↓	R@1↑	R@2↑	R@3↑	R@5↑	R@10↑	MedR↓
HumanML3D	TEMOS [18]	2.12	4.09	5.87	8.26	13.52	173.00	3.86	4.54	6.94	9.38	14.00	183.25
	TMR [19]	5.68	10.59	14.04	20.34	30.94	28.00	9.95	12.44	17.95	23.56	32.69	28.50
	MotionPatches† [25]	10.80	14.98	20.00	26.72	38.02	19.00	11.25	13.86	19.98	26.86	37.40	20.50
	ReMoGPT† [26]	11.00	17.02	22.18	29.48	43.43	14.00	12.25	14.95	21.45	28.34	39.11	19.00
	HBM (Ours)	**18.49**	**21.20**	**25.63**	**32.40**	**46.27**	**12.00**	**14.83**	**17.63**	**25.60**	**30.75**	**44.61**	**16.00**
KIT-ML	TEMOS [18]	7.11	13.25	17.59	24.10	35.66	24.00	11.69	15.30	20.12	26.63	36.39	26.50
	TMR [19]	7.23	13.98	20.36	28.31	40.12	17.00	11.20	13.86	20.12	28.07	38.55	18.00
	MotionPatches† [25]	14.02	21.08	28.91	34.10	50.00	10.50	**13.61**	**17.26**	**27.54**	33.33	44.77	**13.00**
	ReMoGPT† [26]	13.42	17.65	22.24	32.30	44.68	12.50	11.93	14.12	21.59	30.54	39.68	16.50
	HBM (Ours)	**16.75**	**23.10**	**30.50**	**37.56**	**52.20**	**8.50**	12.14	16.36	25.64	**35.58**	**46.15**	14.00
SnapMoGen	TEMOS [18]	5.03	8.51	10.79	14.48	20.36	42.00	4.72	6.67	7.95	10.11	16.74	64.00
	TMR [19]	7.29	11.80	14.43	18.82	27.39	28.50	8.03	10.22	13.49	20.91	29.06	33.00
	MotionPatches† [25]	8.92	14.75	18.13	25.46	34.95	19.00	9.08	12.24	14.73	22.42	33.52	25.50
	ReMoGPT† [26]	8.15	13.27	16.05	24.72	33.48	20.50	9.94	13.55	16.41	23.87	35.53	23.00
	HBM (Ours)	**12.43**	**17.84**	**22.54**	**28.22**	**42.83**	**16.50**	**11.62**	**15.35**	**19.32**	**26.18**	**39.41**	**19.00**

Value: Motion-Domain Synthesis. The Value pathway focuses on *what* information to synthesize, combining current latents with retrieved dynamic priors:

$$V = W_v \cdot \text{concat}(z,\ R_m)\,. \tag{13}$$

Here, R_m provides concrete motion details from the database, while z ensures continuity with the current generation state. Notably, textual semantics (t, R_t) are excluded from Value to prevent domain mismatch.

Output. The final output is computed via standard scaled dot-product attention:

$$\text{Output} = \text{SoftMax}\left(\frac{QK^\top}{\sqrt{d}}\right)V, \tag{14}$$

which is then reshaped back to the 2D grid structure. This asymmetric design ensures that retrieved semantics guide the attention focus (via Key) and enrich the motion content (via R_m in Value), while strictly preserving the spatial–temporal topology of the generative backbone.

5 Experiment

5.1 Dataset and Evaluation Metrics

We evaluate our model on HumanML3D [8], KIT-ML [21], and SnapMoGen [6] datasets. The HumanML3D consists of 14,616 motions and 44,970 texts, while KIT-ML consists of 3,911 motions and 6,278 texts. The SnapMoGen is the largest available dataset, comprising 20,450 motions and 122,565 texts. The motion feature dimensions are 263, 251, and 296 for HumanML3D, KIT-ML, and SnapMoGen, respectively. HumanML3D and KIT-ML are split into training, validation,

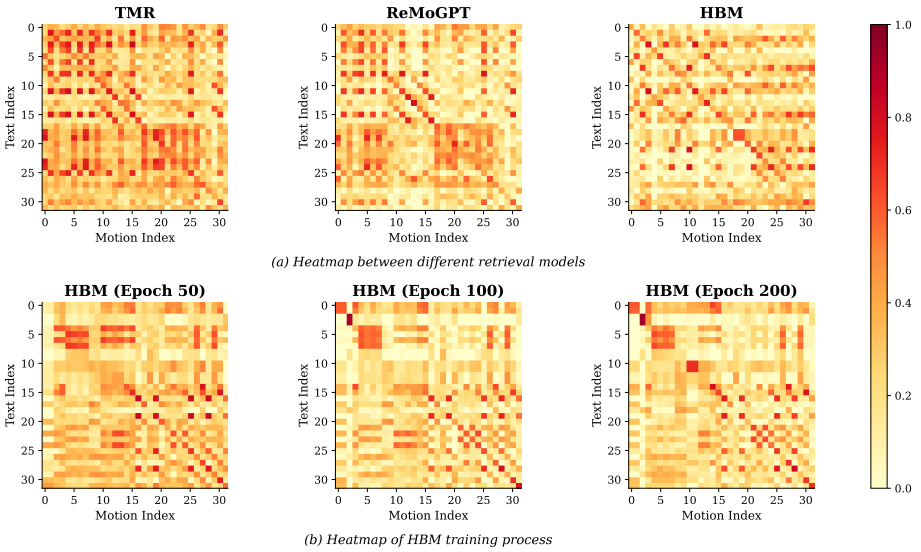


Fig. 3: Heatmap of similarity matrix. The diagonal represents positive pairs, with darker colors meaning higher semantic similarity, conducted on HumanML3D.

and test sets with a ratio of **0.80:0.05:0.15**, while SnapMoGen uses a split of **0.85:0.05:0.10**.

Evaluation Metrics. For text–motion retrieval, we report Recall at different ranks (R@1, R@2, R@3, R@5, and R@10), which measures the proportion of queries whose ground-truth match appears within the top-k retrieved results. We also report the median rank (MedR), where lower values indicate better retrieval performance. For motion generation, following prior work [7, 27], we evaluate text-to-motion (T2M) performance from three perspectives: (1) motion quality, measured by the Frechet Inception Distance (FID); (2) generation diversity, evaluated using Diversity and Multi-Modality (MModality); and (3) text–motion alignment, assessed by R-Precision at top 1/2/3 and the Multi-Modal Distance (MM Dist).

5.2 Implementation Details

For the motion representation, we use a pretrained 2D-RVQ-VAE [27] with 6 quantization layers, each contraining a codebook of 512 codes of 512-dimensios. For the topology structured masking, we set π_{base} to 0.5. For the text–motion retrieval model, the motion encoder employs a 4-layer transformer for each body part, with both part-level and instance-level motion embeddings set to 512 dimensions. The text encoder is a frozen pretrained CLIP [22] ViT-B/32 model with one additional trainable transformer layer, shared across all models. For the motion masked model, the SSTA is set to have 6 layers, 8 heads, and 512 latent dimensions. The retrieval model is trained for 200 epochs on a Tesla A800 GPU, with a λ_P of 1, a batch size of 128, a queue size of 65,536, and a momentum μ

Table 4: Quantitative evaluation on HumanML3D, KIT-ML, and SnapMoGen datasets. We repeat the evaluation 20 times and report the average with 95% confidence interval. Results marked with † are reproduced following the original paper, as no official implementation is publicly available. **Bold** and underline indicate the best and the second best results.

	Methods	Framework	R-Precision↑			FID↓	MMDIST↓	Diversity→	MModality↑
			Top 1	Top 2	Top 3				
HumanML3D	Real	-	$0.511^{\pm.003}$	$0.703^{\pm.003}$	$0.797^{\pm.002}$	$0.002^{\pm0.000}$	$2.974^{\pm0.008}$	$9.503^{\pm0.065}$	–
	MDM [24]	t2m	–	–	$0.611^{\pm.007}$	$0.544^{\pm0.44}$	$5.566^{\pm.027}$	$9.559^{\pm.086}$	$2.799^{\pm.072}$
	T2M-GPT [28]		$0.492^{\pm.003}$	$0.679^{\pm.002}$	$0.775^{\pm.002}$	$0.141^{\pm.005}$	$3.121^{\pm.009}$	$9.761^{\pm.073}$	$1.856^{\pm.003}$
	MoMask [7]		$0.521^{\pm.002}$	$0.713^{\pm.002}$	$0.807^{\pm.002}$	$0.045^{\pm.002}$	$2.958^{\pm.008}$	$9.589^{\pm.035}$	$1.241^{\pm.040}$
	MoGenTS [27]		$0.529^{\pm.003}$	$0.719^{\pm.002}$	$0.812^{\pm.002}$	$0.033^{\pm.001}$	$2.867^{\pm.006}$	$9.570^{\pm.077}$	$1.205^{\pm.043}$
	MoMask++ [6]		$0.517^{\pm.002}$	$0.709^{\pm.002}$	$0.803^{\pm.002}$	$0.069^{\pm.003}$	$2.948^{\pm.007}$	$9.611^{\pm.032}$	$1.192^{\pm.053}$
	MARDM [17]		$0.500^{\pm.004}$	$0.695^{\pm.003}$	$0.795^{\pm.003}$	$0.114^{\pm.007}$	$2.945^{\pm.004}$	$9.718^{\pm.031}$	$2.231^{\pm.071}$
	LaMP [15]		$0.557^{\pm.003}$	$0.751^{\pm.002}$	$0.843^{\pm.001}$	$0.032^{\pm.002}$	$\mathbf{2.759}^{\pm.007}$	$9.571^{\pm.069}$	$2.794^{\pm.041}$
	ReMoDiffuse [29]	RAG-t2m	$0.510^{\pm.005}$	$0.698^{\pm.006}$	$0.795^{\pm.004}$	$0.103^{\pm.004}$	$2.974^{\pm.016}$	$9.018^{\pm.075}$	$1.795^{\pm.043}$
	ReMoGPT† [26]		$0.501^{\pm.003}$	$0.688^{\pm.003}$	$0.792^{\pm.005}$	$0.205^{\pm.023}$	$2.929^{\pm.020}$	$9.763^{\pm.056}$	$2.816^{\pm.003}$
	RMD† [16]		$0.524^{\pm.002}$	$0.715^{\pm.002}$	$0.811^{\pm.001}$	$0.111^{\pm.005}$	$2.879^{\pm.006}$	$\mathbf{9.527}^{\pm.090}$	$2.604^{\pm.084}$
	MoRAG-Diffuse† [12]		$0.511^{\pm.003}$	$0.699^{\pm.003}$	$0.792^{\pm.002}$	$0.270^{\pm.010}$	$2.950^{\pm.001}$	$9.536^{\pm.104}$	$2.773^{\pm.011}$
	ReMoMask(Ours)		$\mathbf{0.566}^{\pm.004}$	$\mathbf{0.763}^{\pm.005}$	$\mathbf{0.850}^{\pm.004}$	$\mathbf{0.026}^{\pm.002}$	$2.867^{\pm.010}$	$9.585^{\pm.053}$	$\mathbf{2.835}^{\pm.005}$
KIT-ML	Real	-	$0.424^{\pm.005}$	$0.649^{\pm.006}$	$0.779^{\pm.006}$	$0.031^{\pm0.004}$	$2.788^{\pm0.012}$	$11.080^{\pm0.97}$	–
	MDM [24]	t2m	–	–	$0.396^{\pm.004}$	$0.497^{\pm.021}$	$9.191^{\pm.022}$	$10.847^{\pm.109}$	$1.907^{\pm.214}$
	T2M-GPT [28]		$0.416^{\pm.006}$	$0.627^{\pm.006}$	$0.745^{\pm.006}$	$0.514^{\pm.029}$	$3.007^{\pm.023}$	$10.860^{\pm.049}$	$1.798^{\pm.157}$
	MoMask [7]		$0.433^{\pm.007}$	$0.656^{\pm.005}$	$0.781^{\pm.005}$	$0.204^{\pm.011}$	$2.779^{\pm.022}$	$10.203^{\pm.038}$	$1.131^{\pm.043}$
	MoGenTS [27]		$0.445^{\pm.006}$	$0.671^{\pm.006}$	$0.797^{\pm.005}$	$0.143^{\pm.004}$	$2.711^{\pm.024}$	$10.918^{\pm.090}$	$1.493^{\pm.024}$
	MoMask++ [6]		$0.428^{\pm.008}$	$0.628^{\pm.006}$	$0.755^{\pm.007}$	$0.194^{\pm.016}$	$2.814^{\pm.006}$	$10.927^{\pm.101}$	$1.247^{\pm.040}$
	MARDM [17]		$0.387^{\pm.006}$	$0.610^{\pm.006}$	$0.749^{\pm.006}$	$0.242^{\pm.014}$	$2.970^{\pm.032}$	$10.891^{\pm.045}$	$1.312^{\pm.053}$
	LaMP [15]		$0.479^{\pm.006}$	$0.691^{\pm.005}$	$0.826^{\pm.005}$	$0.141^{\pm.013}$	$\mathbf{2.704}^{\pm.018}$	$\mathbf{10.929}^{\pm.101}$	$\mathbf{1.973}^{\pm.071}$
	ReMoDiffuse [29]	RAG-t2m	$0.427^{\pm.014}$	$0.641^{\pm.004}$	$0.765^{\pm.055}$	$0.155^{\pm.006}$	$2.814^{\pm.012}$	$10.800^{\pm.105}$	$1.239^{\pm.028}$
	ReMoGPT† [26]		$0.376^{\pm.006}$	$0.585^{\pm.004}$	$0.734^{\pm.005}$	$0.425^{\pm.022}$	$2.935^{\pm.009}$	$10.712^{\pm.092}$	$1.362^{\pm.020}$
	RMD† [16]		$0.433^{\pm.007}$	$0.652^{\pm.005}$	$0.776^{\pm.004}$	$0.320^{\pm.019}$	$2.863^{\pm.015}$	$10.879^{\pm.110}$	$1.364^{\pm.100}$
	MoRAG-Diffuse† [12]		$0.330^{\pm.008}$	$0.542^{\pm.007}$	$0.675^{\pm.004}$	$0.614^{\pm.043}$	$3.251^{\pm.022}$	$10.244^{\pm.092}$	$1.268^{\pm.045}$
	ReMoMask(Ours)		$\mathbf{0.483}^{\pm.006}$	$\mathbf{0.698}^{\pm.005}$	$\mathbf{0.835}^{\pm.005}$	$\mathbf{0.131}^{\pm.004}$	$2.882^{\pm.010}$	$10.922^{\pm.043}$	$\mathbf{1.978}^{\pm.005}$
SnapMoGen	Real	-	$0.940^{\pm.001}$	$0.976^{\pm.001}$	$0.985^{\pm.001}$	$0.001^{\pm.000}$	$2.835^{\pm.001}$	$19.849^{\pm.063}$	–
	MDM [24]	t2m	$0.503^{\pm.002}$	$0.653^{\pm.002}$	$0.727^{\pm.002}$	$57.783^{\pm.092}$	$8.627^{\pm.012}$	$18.701^{\pm.043}$	$\mathbf{13.412}^{\pm.231}$
	T2M-GPT [28]		$0.618^{\pm.002}$	$0.773^{\pm.002}$	$0.812^{.002}$	$32.629^{\pm.087}$	$6.855^{\pm.032}$	$18.513^{\pm.035}$	$9.172^{\pm.181}$
	MoMask [7]		$0.777^{\pm.002}$	$0.888^{\pm.002}$	$0.927^{\pm.002}$	$17.404^{\pm.051}$	$3.812^{\pm.008}$	$19.583^{\pm.074}$	$8.183^{\pm.184}$
	MoGenTS [27]		$0.742^{\pm.004}$	$0.875^{\pm.003}$	$0.912^{\pm.004}$	$23.927^{\pm.045}$	$3.923^{\pm.006}$	$19.457^{\pm.061}$	$8.841^{\pm.197}$
	MoMask++ [6]		$0.802^{\pm.001}$	$0.905^{\pm.002}$	$0.938^{\pm.001}$	$15.060^{\pm.065}$	$3.784^{\pm.006}$	$\mathbf{19.764}^{\pm.027}$	$7.259^{\pm.180}$
	MARDM [17]		$0.659^{\pm.002}$	$0.812^{\pm.002}$	$0.860^{\pm.002}$	$26.878^{\pm.131}$	$3.947^{\pm.009}$	$19.598^{\pm.019}$	$9.812^{\pm.287}$
	LaMP [15]		$0.711^{\pm.005}$	$0.803^{\pm.006}$	$0.900^{\pm.003}$	$19.114^{\pm.085}$	$3.697^{\pm.014}$	$19.618^{\pm.037}$	$9.325^{\pm.239}$
	ReMoDiffuse [29]	RAG-t2m	$0.491^{\pm.003}$	$0.673^{\pm.007}$	$0.785^{\pm.002}$	$68.460^{\pm.091}$	$5.930^{\pm.029}$	$18.753^{\pm.046}$	$8.260^{\pm.215}$
	ReMoGPT† [26]		$0.647^{\pm.002}$	$0.793^{\pm.004}$	$0.839^{\pm.002}$	$44.121^{\pm.0.16}$	$4.883^{\pm.008}$	$19.435^{\pm.051}$	$9.347^{\pm.459}$
	RMD† [16]		$0.534^{\pm.002}$	$0.658^{\pm.002}$	$0.713^{\pm.003}$	$49.084^{\pm.023}$	$5.916^{\pm.017}$	$18.566^{\pm.021}$	$7.559^{\pm.273}$
	MoRAG-Diffuse† [12]		$0.597^{\pm.003}$	$0.792^{\pm.002}$	$0.835^{\pm.002}$	$41.417^{\pm.030}$	$5.924^{\pm.036}$	$18.927^{\pm.032}$	$8.278^{\pm.247}$
	ReMoMask(Ours)		$\mathbf{0.811}^{\pm.004}$	$\mathbf{0.917}^{\pm.003}$	$\mathbf{0.949}^{\pm.002}$	$\mathbf{13.509}^{\pm.041}$	$\mathbf{3.459}^{\pm.007}$	$19.629^{\pm.055}$	$8.820^{\pm.186}$

of 0.999. The masked models are trained on 8 Tesla A800 GPUs for up to 2,000 epochs with a batch size of 64. All models are implemented in PyTorch.

5.3 Main Results

Evaluation of Motion Retrieval. We evaluate ReMoMask on text-motion and motion-text retrieval on HumanML3D, KIT-ML, and SnapMoGen (Tab. 3), achieving state-of-the-art performance across all benchmarks. ReMoMask consistently outperforms prior methods on HumanML3D and SnapMoGen, improving text-to-motion R@1 on HumanML3D by 68.09% over ReMoGPT (18.49

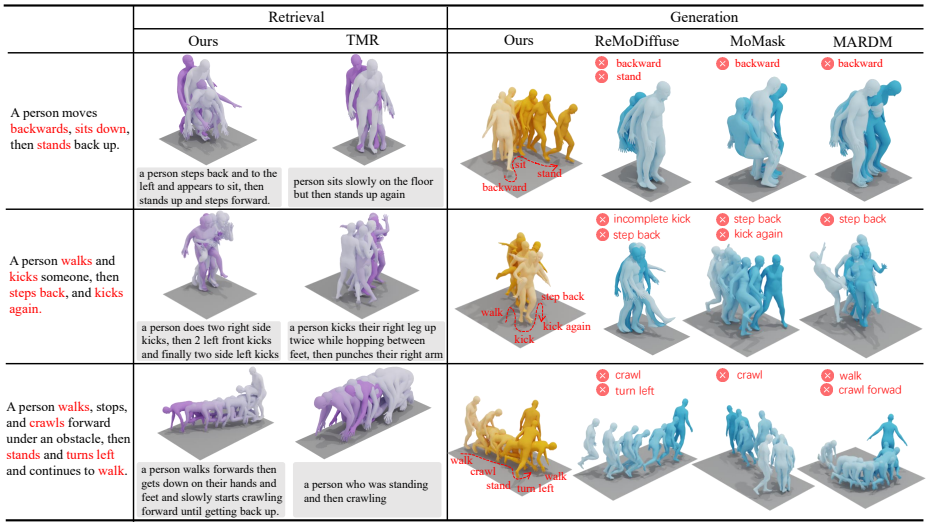


Fig. 4: Qualitative comparisons. The darker colors indicate the later in time. The motions generated by our method closely align with the descriptions, outperforming others that exhibit degraded motions or improper semantics.

vs. 11.00) and increasing R@1 on SnapMoGen from 8.92 to 12.43. On KIT-ML, ReMoMask achieves the best performance at higher recall levels (R@5 and R@10), indicating robust retrieval. Fig. 3 further shows that HBM yields clearer diagonal structures with suppressed off-diagonal responses, confirming more accurate and discriminative text–motion alignment.

Evaluation of Motion Generation. We evaluate motion generation performance on the HumanML3D, KIT-ML, and SnapMoGen datasets, as illustrated in Tab. 4. ReMoMask consistently achieves strong motion generation performance, as reflected by consistently lower FID scores compared to prior state-of-the-art methods across all three benchmarks. On KIT-ML and SnapMoGen, some baselines achieve slightly higher Diversity or MModality, which is expected for retrieval-augmented generation methods that prioritize semantic alignment and motion coherence.

Qualitative comparisons. We compare our method with others in Fig. 4. The visualizations demonstrate that our model achieves superior motion realism and better text-motion alignment.

5.4 Ablation Study

We conduct ablation studies on the HumanML3D dataset to investigate the impact of each component. Quantitative results reported in Tab. 5 show that all the proposed components contribute greatly.

Table 5: Ablation study on HumanML3D dataset.

	FID↓	Top1↑	MMDIST↓
Baseline	$0.054^{\pm.013}$	$0.536^{\pm.003}$	$2.946^{\pm.004}$
+ HBM	$0.037^{\pm.021}$	$0.547^{\pm.003}$	$2.862^{\pm.009}$
+ SSTA	$0.031^{\pm.006}$	$0.558^{\pm.004}$	$2.879^{\pm.004}$
+ TSM	$0.026^{\pm.007}$	$0.565^{\pm.006}$	$2.837^{\pm.008}$

Hierarchical Retrieval and Momentum Encoder. Tab. 6 presents a complete retrieval ablation study on HumanML3D. Compared with single-level supervision, hierarchical retrieval consistently achieves better performance, improving text-to-motion R@1 from 7.32/11.50 (part/instance only) to 14.87 while reducing MedR from 27.0/18.0 to 16.0, demonstrating the complementary benefits of part-level and instance-level alignment. Furthermore, introducing the momentum encoder consistently enhances retrieval quality, with the hierarchical setting achieving the best overall performance, increasing text-to-motion R@1 from 14.87 to 18.34 and R@5 from 30.65 to 32.62. These results verify that hierarchical supervision and momentum learning are complementary in learning more robust cross-modal representations.

Table 6: Retrieval ablation analysis on HumanML3D dataset.

Momentum	Hierarchical Level		Text-motion retrieval			Motion-text retrieval		
	Part	Instance	R@1 ↑	R@5 ↑	MedR ↓	R@1 ↑	R@5 ↑	MedR ↓
×	✓	×	7.32	21.25	27.00	7.95	23.68	28.00
×	×	✓	11.50	28.45	18.00	11.41	27.19	19.00
×	✓	✓	<u>14.87</u>	30.65	<u>16.00</u>	<u>13.78</u>	28.68	18.50
✓	✓	×	7.98	21.44	26.00	8.12	23.26	27.50
✓	×	✓	12.67	<u>30.95</u>	16.00	13.27	<u>29.46</u>	<u>17.50</u>
✓	✓	✓	**18.34**	**32.62**	**12.50**	**14.59**	**30.58**	**16.00**

Generalization Across Backbones. We further evaluate the generalizability of HBM by replacing the original retrieval encoder in ReMoDiffuse and ReMoGPT while keeping all other components unchanged. As shown in Tab. 7, HBM consistently improves both backbones, reducing FID from 0.103 to 0.091 on ReMoDiffuse and from 0.205 to 0.165 on ReMoGPT, while also yielding consistent gains in Top-k accuracy. These results demonstrate the good transferability of the proposed hierarchical retrieval representation across different generation architectures.

Table 7: Generalization across backbones on HumanML3D.

BaseModel	Encoder	FID↓	Top1↑	Top2↑	Top3↑
ReMoDiffuse	Clip	0.103	0.510	0.698	0.795
	HBM	0.091(-0.012)	0.515	0.708	0.799
ReMoGPT	Transformer	0.205	0.501	0.688	0.792
	HBM	0.165(-0.040)	0.505	0.693	0.794

Masking Strategy. We compare the proposed Topology Structured Masking (TSM) with the commonly used cosine scheduler masking in Tab. 8. TSM consistently achieves better performance, reducing FID from 0.034 to 0.025 while improving Top-1 accuracy from 0.556 to 0.567, demonstrating the effectiveness of topology-aware masking for learning more coherent motion representations.

Table 8: Comparison of masking strategies on HumanML3D.

Masking strategies	FID↓	Top1↑	Top2↑	Top3↑
cosine scheduler masking	0.034	0.556	0.751	0.842
topology structured masking	0.025(-0.009)	0.567	0.764	0.848

Information Source. We investigate the roles of different information sources within the SSTA key and value matrices on the HumanML3D dataset. Under identical settings (Tab. 9), our results reveal a clear functional asymmetry: enriching the key matrix with retrieved information consistently improves performance, indicating that the key primarily acts as a selector that benefits from diverse reference signals. However, constraining the value matrix leads to better results, as injecting excessive or misaligned information into values can degrade generation quality, suggesting that effective SSTA design requires rich keys for attention selection and controlled values for stable information aggregation.

Table 9: The impacts of different information source selections for key and value matrices in SSTA on t2m task, conducted on the HumanML3D dataset.

K Matrix			V Matrix			FID↓	R-Precision↑	MMDIST↓
t	R_t	R_m	t	R_t	R_m		Top1	
✓	✓	✓	✓	✓	✓	$\underline{0.043}^{\pm.007}$	$\underline{0.548}^{\pm.008}$	$2.896^{\pm.007}$
✓	✓	✓		✓	✓	$0.104^{\pm.005}$	$0.506^{\pm.004}$	$\mathbf{2.865}^{\pm.005}$
✓	✓	✓			✓	$\mathbf{0.027}^{\pm.005}$	$\mathbf{0.562}^{\pm.004}$	$2.877^{\pm.008}$
✓		✓			✓	$0.073^{\pm.003}$	$0.536^{\pm.003}$	$\underline{2.872}^{\pm.005}$
✓	✓				✓	$0.096^{\pm.007}$	$0.513^{\pm.005}$	$2.973^{\pm.005}$
✓					✓	$0.194^{\pm.005}$	$0.476^{\pm.012}$	$3.242^{\pm.013}$

Hyperparameters. We analyze the sensitivity of HBM contrastive training to key hyperparameters on HumanML3D in Fig. 5. Larger queue sizes improve retrieval by providing more negatives, while optimal performance is achieved with moderate values of λ_P, temperature, and momentum, indicating a balanced and stable training regime.

Database Coverage. We analyze the impact of database coverage on retrieval-augmented motion generation by varying the percentage of the retrieval database. As shown in Fig. 6a, increasing the coverage from 10% to 100% reduces FID from 0.073 to 0.026 while improving R-Precision from 0.502 to 0.566, demonstrating that richer retrieval databases substantially enhance generation quality.

Efficiency Following [7], we evaluate the efficiency and quality of motion generation using various methods. The inference cost is quantified as the mean inference time over 100 samples on a single Nvidia 2080Ti device. The result in Fig. 6b shows that ReMoMask offers better quality-efficiency trade-offs than baseline methods.

6 Limitations

The performance of our framework depends on the quality and coverage of the retrieval database. Building richer motion databases remains an important direction for future work.

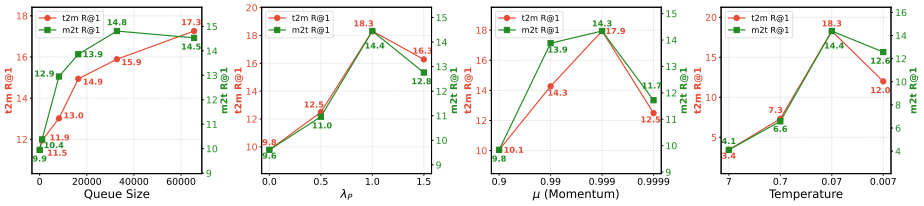
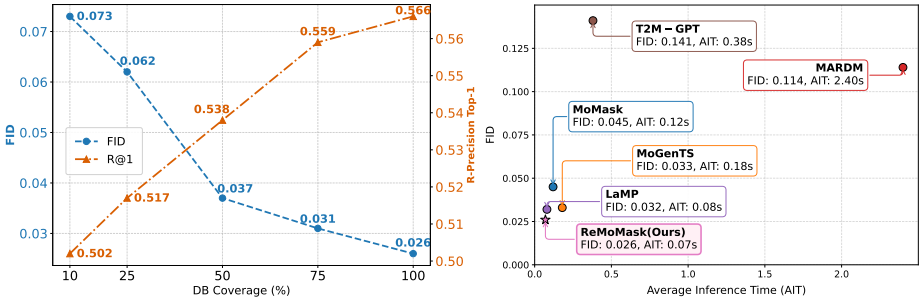


Fig. 5: Sensitivity of contrastive training to key hyperparameters, conducted on HumanML3D.



(a) Impact of database coverage. **(b)** Comparisons on FID and Inference Cost.

Fig. 6: (a) Impact of database coverage. A richer retrieval database substantially enhances generation quality. **(b) Comparisons on FID and Inference Cost.** All the tests are conducted on a Nvidia2080Ti GPU.

7 Conclusion

In this work, we present ReMoMask, a retrieval-augmented masked model for text-to-motion generation. To overcome the limitations of coarse-grained retrieval and ineffective fusion, we introduce **H**ierarchical **B**idirectional **M**omentum (HBM) for precise text-motion alignment and **T**opology **S**tructured **M**asking (TSM) to enforce structural consistency during training. Complemented by **S**emantic **S**patial **T**emporal **A**ttention (SSTA) for knowledge integration, ReMoMask effectively bridges structured retrieval with high-quality generation. Extensive experiments on HumanML3D, KIT-ML, and SnapMoGen confirm that our method consistently outperforms state-of-the-art approaches in both retrieval accuracy and motion fidelity.

Acknowledgements

This work was supported by the Fundamental Research Funds for the Central Universities, Peking University.

References

1. Dong, J., Koniusz, P., Qu, X., Ong, Y.S.: Stabilizing modality gap & lowering gradient norms improve zero-shot adversarial robustness of vlms. In: Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1. pp. 236–247 (2025)
2. Dong, J., Koniusz, P., Zhang, Y., Zhu, H., Liu, W., Qu, X., Ong, Y.S.: Improving zero-shot adversarial robustness in vision-language models by closed-form alignment of adversarial path simplices. In: Forty-second International Conference on Machine Learning (2025)
3. Dong, J., Liu, J., Qu, X., Ong, Y.S.: Confound from all sides, distill with resilience: Multi-objective adversarial paths to zero-shot robustness. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 624–634 (2025)
4. Dong, J., Zhang, C., Qu, X., Ma, Z., Koniusz, P., Ong, Y.S.: Robust superalignment: Weak-to-strong robustness generalization for vision-language models. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025)
5. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Auto-encoding variational bayes. arXiv preprint arXiv:1406.2661 (2014)
6. guo, c., Hwang, I., Wang, J., Zhou, B.: Snapmogen: Human motion generation from expressive texts. In: Belgrave, D., Zhang, C., Lin, H., Pascanu, R., Koniusz, P., Ghassemi, M., Chen, N. (eds.) Advances in Neural Information Processing Systems. vol. 38, pp. 99939–99955. Curran Associates, Inc. (2025)
7. 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)
8. 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 (CVPR). pp. 5152–5161 (June 2022)
9. 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)
10. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
11. 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)
12. Kalakonda, S.S., Maheshwari, S., Sarvadevabhatla, R.K.: Morag - multi-fusion retrieval augmented generation for human motion. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) (2025)
13. Kingma, D.P., Welling, M.: Generative adversarial networks. arXiv preprint arXiv:1312.6114 (2022)
14. Li, Y., Yang, J., Yang, Z., Li, B., He, H., Yao, Z., Han, L., Chen, Y.V., Fei, S., Liu, D., et al.: Cama: Enhancing multimodal in-context learning with context-aware modulated attention. arXiv preprint arXiv:2505.17097 (2025)
15. Li, Z., Yuan, W., He, Y., Qiu, L., Zhu, S., Gu, X., Shen, W., Dong, Y., Dong, Z., Yang, L.: Lamp: Language-motion pretraining for motion generation, retrieval, and captioning. In: International Conference on Learning Representations. vol. 2025, pp. 84238–84250 (2025)

16. Liao, Z., Zhang, M., Wang, W., Yang, L., Komura, T.: Rmd: A simple baseline for more general human motion generation via training-free retrieval-augmented motion diffuse. arXiv preprint arXiv:2412.04343 (2024)
17. 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)
18. Petrovich, M., Black, M.J., Varol, G.: TEMOS: Generating diverse human motions from textual descriptions. In: European Conference on Computer Vision (ECCV) (2022)
19. 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)
20. 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)
21. Plappert, M., Mandery, C., Asfour, T.: The KIT motion-language dataset. Big Data **4**(4), 236–252 (dec 2016)
22. 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)
23. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2022)
24. Tevet, G., Raab, S., Gordon, B., Shafir, Y., Cohen-Or, D., Bermano, A.H.: Human motion diffusion model. arXiv preprint arXiv:2209.14916 (2022)
25. Yu, Q., Tanaka, M., Fujiwara, K.: Exploring vision transformers for 3d human motion-language models with motion patches. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2024)
26. Yu, Q., Tanaka, M., Fujiwara, K.: Remogpt: Part-level retrieval-augmented motion-language models. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 9635–9643 (2025)
27. Yuan, W., Shen, W., HE, Y., Dong, Y., Gu, X., Dong, Z., Bo, L., Huang, Q.: Mogents: Motion generation based on spatial-temporal joint modeling. In: Neural Information Processing Systems (NeurIPS) (2024)
28. Zhang, J., Zhang, Y., Cun, X., Huang, S., Zhang, Y., Zhao, H., Lu, H., Shen, X.: T2m-gpt: Generating human motion from textual descriptions with discrete representations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
29. 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)
30. Zhang, Y., He, Y., Shao, Y., Yao, Z., Xu, H., Dong, J., Yao, Z., Dong, Z.: Chromouvqa: Benchmarking vision-language models under chromatic camouflaged images. arXiv preprint arXiv:2512.05137 (2025)
31. Zhang, Z., Wang, Y., Wu, B., Chen, S., Zhang, Z., Huang, S., Zhang, W., Fang, M., Chen, L., Zhao, Y.: Motion avatar: Generate human and animal avatars with arbitrary motion. arXiv preprint arXiv:2405.11286 (2024)
32. 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)