Reconstruction-Anchored Diffusion Model for Text-to-Motion Generation

Yifei Liu[1,2], Changxing Ding[1] *, Ling Guo[1], Huaiguang Jiang[1], and Qiong Cao[2]

[1] South China University of Technology
{ft_lyf, 202510182670}@mail.scut.edu.cn,
{chxding, hihuagong2021}@scut.edu.cn
[2] Joy Future Academy
mathqiong2012@gmail.com

Abstract. Diffusion models have seen widespread adoption for text-driven human motion generation and related tasks due to their impressive generative capabilities and flexibility. However, current motion diffusion models face two major limitations: a representational gap caused by pre-trained text encoders that lack motion-specific information, and error propagation during the iterative denoising process. This paper introduces **R**econstruction-**A**nchored Diffusion **M**odel (**RAM**) to address these challenges. First, RAM leverages a motion latent space as intermediate supervision for text-to-motion generation. To this end, RAM co-trains a motion reconstruction branch with two key objective functions: self-regularization to enhance the discrimination of the motion space and motion-centric latent alignment to enable accurate mapping from text to the motion latent space. Second, we propose Reconstructive Error Guidance (REG), a testing-stage guidance mechanism that exploits the motion diffusion model's inherent self-correction ability to mitigate error propagation. At each denoising step, REG uses the motion reconstruction branch to reconstruct the previous estimate, reproducing the prior error patterns. By amplifying the residual between the current prediction and the reconstructed estimate, REG highlights the improvements in the current prediction. Extensive experiments demonstrate that RAM achieves significant improvements and state-of-the-art performance. Our code will be released on `https://feifeifeiliu.github.io/RAM` .

Keywords: Human Motion Generation · Text-to-Motion Generation · Diffusion Models · Representation Learning · Diffusion Guidance

1 Introduction

Imagine giving a textual description and immediately witnessing a lifelike avatar execute it, with physically plausible, faithful body movements in the correct sequence. This vision drives human motion generation with applications in virtual

* Corresponding author.

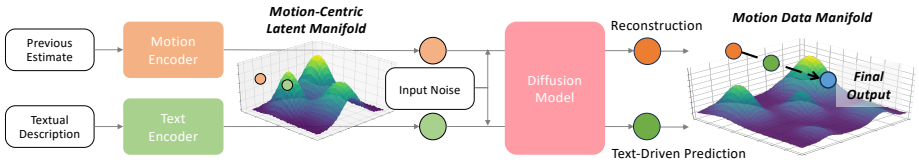

Fig. 1: At inference time, RAM first maps a textual description onto a motion-centric latent manifold and then predicts using a diffusion model. Meanwhile, it reconstructs previous estimates that contain error patterns. By contrasting these predictions, RAM uses the reconstruction as a negative reference to drive the output away from poor estimates and towards the real data manifold. Best viewed in color.

reality [10], game content creation [24], and embodied robotics [47]. The task is inherently challenging because the language is abstract. At the same time, motion is continuous, high-dimensional, and kinematically constrained—demanding both fine-grained semantic understanding and robust many-to-many mappings between natural language and human motion dynamics.

This challenge has sparked extensive research interest, which can be broadly categorized into two main approaches: VQ-VAE-based and diffusion-based methods. Among these, diffusion-based methods have gained widespread adoption across downstream tasks, including motion in-betweening [5], human-object interaction [21], and multi-human motion modeling [8, 25, 40, 45], owing to their exceptional flexibility and controllability. Existing diffusion-based methods typically leverage pre-trained text encoders to obtain robust textual embeddings, such as T5 [29], CLIP [36], and DistilBERT [37]. Conditioned on these textual embeddings, motion diffusion models learn to recover motion data from noise via iterative denoising. Recent advances have incorporated various techniques, including latent diffusion [4], preference optimization [38], hierarchical semantic graphs [18], and retrieval-augmented generation [51], achieving notable improvements in inference speed, motion realism, and semantic-motion alignment.

Nevertheless, motion diffusion models still face severe limitations in both text models and the denoising process. First, pre-trained text models typically lack cues about motion dynamics. Although CLIP captures visual concepts that correlate with actions, it fails to encode essential temporal dynamics and kinematic constraints, as it was trained exclusively on static image-text pairs. This absence forces models to bridge an unnecessarily large representational gap, hindering the learning of accurate semantic-to-dynamic mappings. Second, diffusion models suffer from error propagation [22]. More specifically, early denoising steps, which must recover motion from nearly pure noise, are particularly prone to generating error patterns. Once such artifacts emerge, they can propagate across subsequent denoising steps, leading to a degradation in sample quality.

Herein, we introduce Reconstruction-Anchored Diffusion Model (**RAM**), a novel diffusion-based framework to address these challenges. For the first problem, we leverage the latent space learned from motion reconstruction as an intermediate supervisory signal for text-to-motion generation. Specifically, RAM

employs a two-stream pipeline [2]: motion reconstruction—where the diffusion model reconstructs motion sequences conditioned on motion-encoder latents; and text-to-motion generation—where the same diffusion model generates motion from text-encoder latents. Based on this pipeline, RAM incorporates two objectives: (a) **self-regularization**, which computes a cross-entropy loss in the motion latent space to enhance discrimination between motion latents, helping to learn a compact yet expressive motion representation; and (b) **motion-centric latent alignment**, aligning the text latent space with the motion latent space, with carefully designed gradients to ensure stable end-to-end training. Together, these designs enable RAM to map text embeddings into a latent space sensitive to motion, inherently embedding the dynamic features required for realistic motion synthesis and bridging the representation gap.

To address the second problem, we introduce the Reconstructive Error Guidance (**REG**), which harnesses the self-correction capabilities of text-to-motion diffusion models to mitigate error propagation. Our core insight is that diffusion models can inherently self-correct, as they do when restoring clean data from noise. To maximize this property, at each denoising step during the testing stage, the motion reconstruction branch reconstructs the previous estimate, thereby capturing earlier error patterns. We then calculate the residual between the current text-driven prediction and the reconstruction, and integrate it into the prediction to generate the final output. This residual highlights the improvements in the current prediction. By amplifying this term, REG directs the sampling process away from error-prone regions, thereby reducing error propagation and enhancing the quality of generated motions throughout denoising.

By integrating these core innovations, RAM enables the generation of more realistic and semantically aligned motions from text. Experiments show that RAM achieves state-of-the-art performance: on the HumanML3D dataset [13], RAM achieves an R-Precision@1 of 56.1% and an FID of 0.032 with only 20 inference steps—surpassing previous diffusion-based methods and outperforming most VQ-VAE-based models, which have historically been stronger than diffusion-based ones on this metric. Consistent performance gains are also observed on the KIT-ML dataset [35]. Comprehensive ablation studies further confirm that each component makes a meaningful contribution to overall performance improvements.

2 Related Work

Text-Driven Human Motion Generation. Current research on text-to-motion generation has consolidated mainly around two principal families: diffusion models and vector-quantized variational autoencoders (VQ-VAE). Early diffusion-based approaches such as Motion Diffusion Model [44] and MotionDiffuse [50] trained denoising networks directly in the raw motion space, followed by a series of extensions that target finer semantic alignment [52], open-vocabulary coverage [24], retrieval-enhanced consistency [51], keyframe-centric stability [3]; recent advances further include step-aware fine-tuning [41], part-level motion

composition [20], and reward-guided alignment [46]. In parallel, latent diffusion methods first encode motions into a continuous latent space and denoise them there, aiming to improve efficiency and quality, *e.g.*, MLD [4], MotionLCM [7], Salad [15]. In contrast, VQ-VAE-based pipelines—pioneered by T2M-GPT [49] and advanced through MMM [34], MoMask [12], BAMM [33], MoGenTS [48], BAD [16], KinMo [53], and LaMP [23]—have empirically exhibited higher motion fidelity, typically reflected in lower FID scores than their diffusion counterparts. In this paper, RAM demonstrates that *diffusion-based approaches can achieve FID performance comparable to that of VQ-based approaches.*

Pre-trained Text Models and Two-Stream Methods. Since text-to-motion datasets are significantly smaller than typical text or text–image datasets, most methods utilize pre-trained text models to extract robust text embeddings. CLIP is widely used for its visual-textual embedding space [42], but recent research suggests that it may not be optimal for aligning text and motion. Instead, these studies suggest fine-tuning text encoders to explicitly learn a joint language–motion embedding space [28, 53]. This approach can be traced back to early two-stream methods [2], which use dual branches—motion reconstruction and text-to-motion generation—and share a decoder to learn a joint language-motion space implicitly. Subsequent works further constrain this space using latent alignment, KL divergence, or contrastive learning [11, 31, 32]. These approaches inspire our method but differ fundamentally: we center the alignment on a carefully designed motion latent space, with the text space aligning to it. We demonstrate that, when employing a diffusion model as the decoder, focusing on modeling detailed motion dynamics yields better motion synthesis results than forcing the learning of a joint language-motion space.

Diffusion Guidance. Guidance in diffusion sampling typically combines multiple score estimateps to enrich the effective target distribution or to impose auxiliary conditioning [9, 14, 19]. Common estimates include conditional score estimates $\nabla_{x_t} \log p(x_t|t, c)$, unconditional score estimates $\nabla_{x_t} \log p(x_t|t)$, classifier gradients $\nabla_{x_t} \log p(y|x_t)$, and CLIP-derived similarity gradients [9, 14, 30]. Recent works further introduce deliberately weakened scores by degrading the predictor—*e.g.*, applying dropout [19], skipping layers [39], or perturbing attention [1]. These weak scores function as contrastive references: amplifying samples favored by stronger scores while suppressing those aligned with weaker ones improves fidelity and semantic alignment. In the same spirit, we derive a weakened score by conditioning the predictor on a motion latent that carries previously introduced error patterns, and use it as a contrastive reference within our guidance mechanism.

3 Method

Overview. RAM generates a sequence of realistic human motion from a given text description. This process starts by extracting text embeddings with pre-trained text models. These embeddings are mapped onto a motion latent man-

ifold and decoded into a motion sequence using a diffusion model. RAM also reconstructs the past motion estimate as a negative reference in the inference stage, as illustrated in Fig. 1. By guiding predictions away from this reference, RAM achieves improved sampling quality.

For a thorough understanding of RAM, we begin by presenting the overall architecture, which features two main branches: motion reconstruction and text-to-motion generation (Sec. 3.1). Next, we detail the training objectives, introducing self-regularization and motion-centric latent alignment, which facilitate learning an expressive motion latent space and enable effective mapping from text to motion latents (Sec. 3.2). Lastly, we provide an in-depth explanation of Reconstructive Error Guidance and inference sampling (Sec. 3.3). Fig. 2 provides an overview.

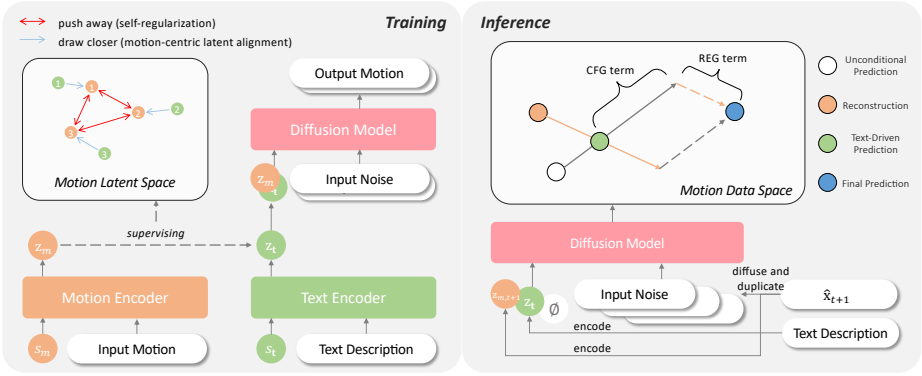

Fig. 2: Overview of RAM. During training, RAM learns a motion latent space through motion reconstruction, with self-regularization to encourage better separability between motion latents, resulting in improved semantic resolution. The text latents from the text encoder are drawn closer to corresponding motion latents through motion-centric latent alignment. At each inference step, given the last step prediction $\hat{\mathbf{x}}_{t+1,s}$ and text description, RAM first encodes them into latents $\mathbf{z}_{m,t+1}$ and $\mathbf{z_t}$. Then, these latents, together with a zero vector $\varnothing$ and input noise, are fed into the diffusion model to separately obtain reconstruction, text-driven prediction, and unconditional prediction. These outputs are combined to produce the final output. Best viewed in color.

3.1 RAM Architecture

Given a motion sequence $\mathbf{x}_0 \in \mathbb{R}^{T\times d}$ or a text description $\mathbf{t}$, RAM either reconstructs the input motion or generates a motion sequence to match a given description. This is achieved through two branches: motion reconstruction and text-to-motion generation, both of which share the Motion Diffusion Model (MDM) [44] as decoder.

Motion Diffusion Model. MDM is modeled as a Markov noising chain $\{\mathbf{x}_t\}_{t=0}^{T}$ with $\mathbf{x}_0$ drawn from the data distribution. The forward diffusion process incrementally adds Gaussian noise: $q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}\big(\mathbf{x}_t; \sqrt{1-\beta_t}\,\mathbf{x}_{t-1}, \beta_t I\big)$. In the reverse process, a denoiser D learns to recover clean motion from a noisy input $\mathbf{x}_t$: $\hat{\mathbf{x}}_t = D(\mathbf{x}_t, t, c)$, where $\hat{\mathbf{x}}_t$ is the motion estimate at timestep t and c denotes conditions.

Motion Reconstruction. The motion reconstruction branch encodes $\mathbf{x}$ into a latent motion $\mathbf{z}_m$ using a transformer-based motion encoder $E_m(\cdot)$, which takes a special token $\mathbf{s}_m$ and the motion sequence as input. The output $\mathbf{z}_m$ represents the global concept of the sequence. The diffusion decoder D takes $\mathbf{z}_m$, timestep t, and noisy motion $\mathbf{x}_t$ to predict clean motion. This process can be expressed as follows:

$$\mathbf{z}_m = E_m(\mathbf{s}_m, \mathbf{x}_0), \qquad \hat{\mathbf{x}}_t = D(\mathbf{x}_t, t, \mathbf{z}_m). \tag{1}$$

Here t denotes the diffusion timestep, which is sampled uniformly as $t \sim \mathcal{U}\{0, \ldots, T-1\}$, where T is the total number of diffusion steps.

Text-to-Motion Generation. The text-to-motion branch mirrors the motion reconstruction branch, encoding the text embedding $\mathbf{f_t}$ with a text encoder $E_\mathbf{t}$ and then decoding with D:

$$\mathbf{z_t} = E_\mathbf{t}(\mathbf{s_t}, \mathbf{f_t}), \qquad \hat{\mathbf{x}}_t = D(\mathbf{x}_t, t, \mathbf{z_t}). \tag{2}$$

Here, $\mathbf{f_t} \in \mathbb{R}^{L \times d_f}$ is the text embedding at the token-level extracted from $\mathbf{t}$ (where L is the sequence length), $\mathbf{s}_t$ is the special input token, and $\mathbf{z_t}$ is the latent produced by the text encoder $E_\mathbf{t}$.

3.2 Optimization Objectives

The training objectives of RAM consist of four key components: reconstruction, text-driven generation, self-regularization, and motion-centric latent alignment. For clarity, we divide these objectives into two categories: (1) reconstruction and text-driven generation, which follow established two-stream approaches [2, 31], and (2) self-regularization and motion-centric latent alignment, which are our novel contributions aimed at learning a compact yet expressive motion latent space and enabling effective mapping from text to motion latents. In the following, we provide a detailed introduction to the formulation and specific function of each objective.

Reconstruction. Given the latent motion $\mathbf{z}_m$, the timestep t, and the noisy motion $\mathbf{x}_t$, this objective encourages the diffusion model to accurately reconstruct the input motion sequence.

$$L_{\text{rec}} = \mathbb{E}_{\mathbf{x}_0,t}\big[\|D(\mathbf{x}_t, t, \mathbf{z}_m) - \mathbf{x}_0\|_2^2\big] = \mathbb{E}_{\mathbf{x}_0,t}\big[\|D(\mathbf{x}_t, t, E_m(\mathbf{s}_m, \mathbf{x}_0)) - \mathbf{x}_0\|_2^2\big]. \tag{3}$$

This loss jointly trains both the diffusion model and the motion encoder, aiming for a strong motion decoder, an encoder that extracts abstract motion representations, and a compact latent space with essential motion dynamics.

Text-Driven Generation. In this objective, the diffusion model learns to generate motion conditioned on the text latent $\mathbf{z_t}$, timestep t, and noisy motion $\mathbf{x}_t$:

$$L_{\text{gen}} = \mathbb{E}_{\mathbf{x}_0,t,\mathbf{t}}\left[\|D(\mathbf{x}_t, t, \mathbf{z_t}) - \mathbf{x}_0\|_2^2\right] = \mathbb{E}_{\mathbf{x}_0,t}\left[\|D(\mathbf{x}_t, t, E_{\mathbf{t}}(\mathbf{s_t}, \mathbf{f_t})) - \mathbf{x}_0\|_2^2\right]. \quad (4)$$

This objective encourages the diffusion model to adapt to conditioning on the text latent manifold, since there are inherent differences between the text and motion manifolds.

Self-Regularization. This objective can be viewed as a cross-entropy loss operating on the motion latent space. For a batch of size B, let the normalized motion latents be $\tilde{\mathbf{z}}_m^i$, and define the similarity $\text{sim}(\tilde{\mathbf{z}}_m^i, \tilde{\mathbf{z}}_m^j) = (\tilde{\mathbf{z}}_m^i)^\top \tilde{\mathbf{z}}_m^j$, which corresponds to the cosine similarity after normalization. With a temperature parameter $\tau = 1$, and treating only identical indices as positive pairs, the loss is defined as:

$$L_{\text{sr}} = \frac{1}{B}\sum_{i=1}^{B} -\log \frac{\exp(\text{sim}(\tilde{\mathbf{z}}_m^i, \tilde{\mathbf{z}}_m^i)/\tau)}{\sum_{j=1}^{B}\exp(\text{sim}(\tilde{\mathbf{z}}_m^i, \tilde{\mathbf{z}}_m^j)/\tau)}. \quad (5)$$

This loss encourages greater separability among motion latents, producing a broader, more expressive manifold with improved semantic resolution. Consequently, the refined latent space enables more precise mapping from text representations to motion latents in the subsequent alignment objective.

Motion-Centric Latent Alignment. This objective aligns the text manifold with the motion manifold. Given a paired text description and motion sequence, this objective minimizes the distance between the corresponding text latent $\mathbf{z_t}$ and motion latent $\mathbf{z}_m$:

$$L_{\text{latent}} = \mathbb{E}_{\mathbf{z}_m,\mathbf{z_t}}\left[\|\mathbf{z_t} - (1-\beta)\,\text{sg}(\mathbf{z}_m) - \beta\mathbf{z}_m\|_2^2\right], \quad (6)$$

where $\text{sg}(\cdot)$ is the stop-gradient operator and β modulates the flow of gradients to the motion encoder E_m. We set $\beta = 0.01$ so that RAM's latent space remains motion-centric while adapting minimally to the text space. This is based on two insights: (1) prioritizing motion space leads to stronger performance than enforcing a fully joint language-motion space, as mapping motion to text sacrifices important motion dynamics, and (2) with end-to-end training, motion latents evolve during alignment. Completely blocking gradients to E_m ($\beta = 0$) makes the alignment harder. Thus, a small β supports convergence while retaining motion information.

Overall Objective. The final training objective is a weighted sum:

$$L_{\text{overall}} = L_{\text{rec}} + L_{\text{gen}} + w_{\text{sr}}L_{\text{sr}} + w_{\text{latent}}L_{\text{latent}}, \quad (7)$$

where w_{sr} and w_{latent} are hyperparameters that determine the significance of the terms L_{sr} and L_{latent}, respectively.

3.3 Inference

Reconstructive Error Guidance. During training, diffusion models operate exclusively on the canonical data manifold, where the noised input follows $x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon$. However, during inference, their predictions often exhibit error patterns and drift away from this manifold. Denoising based on such off-manifold predictions further exacerbates the deviation. In general, diffusion models can cause the sampling path to deviate from the data manifold, leading to degraded sampling quality [22].

We hypothesize that diffusion models possess an inherent capacity to self-correct such error patterns—a capability analogous to their fundamental ability to recover clean data from noise. However, this corrective potential requires explicit activation and guidance. To harness this intrinsic error-correction capability, we propose an intuitive approach that operates at each denoising step.

Specifically, at inference step t, we first reconstruct the prediction from the previous step $t+1$ to capture embedded error patterns explicitly. We then amplify the improvement from the current step's prediction using a residual amplification mechanism. Let $\hat{\mathbf{x}}_{t+1,s}$ denote the final output at step $t+1$. Our method can be formulated as:

$$\hat{\mathbf{x}}_{t,s} = D(\mathbf{x}_t, t, \mathbf{z_t}) + w\left(D(\mathbf{x}_t, t, \mathbf{z_t}) - D(\mathbf{x}_t, t, \mathbf{z}_{m,t+1})\right), \tag{8}$$

where $\mathbf{z}_{m,t+1} = E_m(\mathbf{s}_m, \hat{\mathbf{x}}_{t+1,s})$ represents the reconstructed motion latent from the previous step, and $w \geq 0$ is a weighting coefficient that controls the amplification strength of the residual correction term. We term this inference strategy Reconstructive Error Guidance (REG).

Inference Sampling. Finally, during inference, we combine REG with the commonly used classifier-free guidance (CFG) for sampling. The final output for each denoising step t, denoted as $\hat{\mathbf{x}}_{t,s}$, is computed as:

$$\begin{aligned}\hat{\mathbf{x}}_{t,s} =& D(\mathbf{x}_t, t, \mathbf{z_t}) + w_1 \underbrace{\left(D(\mathbf{x}_t, t, \mathbf{z_t}) - D(\mathbf{x}_t, t, \mathbf{z}_{m,t+1})\right)}_{\text{REG term}} \\ &+ w_2 \underbrace{\left(D(\mathbf{x}_t, t, \mathbf{z_t}) - D(\mathbf{x}_t, t, \varnothing)\right)}_{\text{CFG term}},\end{aligned} \tag{9}$$

where the final term represents the standard CFG residual between conditional and unconditional predictions (with unconditional input denoted by $\varnothing$). Here, w_1 and w_2 respectively control the influence of REG and CFG.

4 Experiment

4.1 Datasets and Metrics

Datasets. HumanML3D [13] is a large-scale text–motion dataset containing 14,616 motion sequences from AMASS [27], each annotated with 44,970

sequence-level textual descriptions. By comparison, the KIT-ML dataset [35] is smaller, offering 3,911 motion sequences and 6,353 textual descriptions. For both datasets, we use the standard redundant motion representation [13], which includes joint velocities, positions, and rotations.

Metrics. We assess the generated motions with five complementary metrics. R-Precision and Multimodal-Dist measure the semantic alignment between generated motions and text descriptions. Fréchet Inception Distance (FID) evaluates the distributional similarity between generated motions and the ground truth in a learned latent space. Diversity (Div) quantifies the variability within the generated motion set, while MultiModality (MM) captures the average variance among motions conditioned on the same description.

Table 1: Quantitative results of text-to-motion generation on the HumanML3D test set. Methods are grouped into VQ-VAE-based and diffusion-based categories. In each group, the best results are highlighted in **bold**, while the second-best results are underlined. Notably, RAM achieves an FID of 0.032, surpassing prior diffusion-based methods and outperforming most VQ-VAE-based models.

	Method	FID↓	R-Precision↑			MM Dist↓	Diversity	MM
			Top 1	Top 2	Top 3			
	Ground Truth	$0.002^{\pm.000}$	$0.511^{\pm.003}$	$0.703^{\pm.003}$	$0.797^{\pm.002}$	$2.974^{\pm.008}$	$9.503^{\pm.065}$	-
VQ-VAE-based	T2M-GPT [49]	$0.116^{\pm.004}$	$0.491^{\pm.003}$	$0.680^{\pm.003}$	$0.775^{\pm.002}$	$3.118^{\pm.011}$	$9.761^{\pm.081}$	$1.856^{\pm.011}$
	MMM [34]	$0.080^{\pm.003}$	$0.504^{\pm.003}$	$0.696^{\pm.003}$	$0.794^{\pm.002}$	$2.998^{\pm.007}$	$9.411^{\pm.058}$	$1.164^{\pm.041}$
	MoMask [12]	$0.045^{\pm.002}$	$0.521^{\pm.002}$	$0.713^{\pm.002}$	$0.807^{\pm.002}$	$2.958^{\pm.008}$	-	$1.241^{\pm.040}$
	BAMM [33]	$0.055^{\pm.002}$	$0.525^{\pm.002}$	$0.720^{\pm.003}$	$0.814^{\pm.003}$	$2.919^{\pm.008}$	$9.717^{\pm.089}$	$1.687^{\pm.051}$
	MoGenTS [48]	$0.033^{\pm.001}$	$0.529^{\pm.003}$	$0.719^{\pm.002}$	$0.812^{\pm.002}$	$2.867^{\pm.006}$	$9.570^{\pm.077}$	-
	BAD [16]	$0.065^{\pm.003}$	$0.517^{\pm.002}$	$0.713^{\pm.003}$	$0.808^{\pm.003}$	$2.901^{\pm.008}$	$9.694^{\pm.068}$	$1.194^{\pm.044}$
	KinMo [53]	$0.039^{\pm.003}$	$0.532^{\pm.002}$	$0.724^{\pm.003}$	$0.821^{\pm.003}$	$2.901^{\pm.010}$	$9.674^{\pm.058}$	$1.321^{\pm.039}$
	LaMP [23]	$\mathbf{0.032}^{\pm.002}$	$\mathbf{0.557}^{\pm.003}$	$\mathbf{0.751}^{\pm.002}$	$\mathbf{0.843}^{\pm.001}$	$\mathbf{2.759}^{\pm.007}$	$9.571^{\pm.069}$	-
Diffusion-based	MDM [44]	$0.489^{\pm.025}$	$0.418^{\pm.005}$	$0.604^{\pm.001}$	$0.707^{\pm.004}$	$3.360^{\pm.023}$	$9.450^{\pm.066}$	$2.860^{\pm1.11}$
	MotionDiffuse [50]	$0.630^{\pm.001}$	$0.491^{\pm.001}$	$0.681^{\pm.001}$	$0.782^{\pm.001}$	$3.113^{\pm.001}$	$9.410^{\pm.049}$	$1.553^{\pm.040}$
	MLD [4]	$0.473^{\pm.013}$	$0.481^{\pm.003}$	$0.673^{\pm.003}$	$0.772^{\pm.002}$	$3.196^{\pm.010}$	$9.724^{\pm.082}$	$2.413^{\pm.079}$
	ReMoDiffuse [51]	$0.103^{\pm.004}$	$0.510^{\pm.005}$	$0.698^{\pm.006}$	$0.795^{\pm.004}$	$2.974^{\pm.016}$	$9.018^{\pm.075}$	$1.795^{\pm.043}$
	FineMoGen [52]	$0.151^{\pm.008}$	$0.504^{\pm.002}$	$0.690^{\pm.002}$	$0.784^{\pm.002}$	$2.998^{\pm.008}$	$9.263^{\pm.094}$	$2.696^{\pm.079}$
	MotionLCM [7]	$0.304^{\pm.012}$	$0.502^{\pm.003}$	$0.698^{\pm.002}$	$0.798^{\pm.002}$	$3.012^{\pm.007}$	$9.607^{\pm.066}$	$2.259^{\pm.092}$
	MotionLCM v2 [6]	$0.056^{\pm.003}$	$0.553^{\pm.003}$	$0.746^{\pm.002}$	$0.837^{\pm.002}$	$2.773^{\pm.009}$	$9.598^{\pm.067}$	$1.758^{\pm.056}$
	StableMoFusion [17]	$0.098^{\pm.003}$	$0.553^{\pm.003}$	$0.748^{\pm.002}$	$0.841^{\pm.002}$	-	$9.748^{\pm.092}$	$1.774^{\pm.051}$
	CLoSD [43]	$0.283^{\pm.000}$	$0.464^{\pm.000}$	$0.668^{\pm.000}$	$0.777^{\pm.000}$	$3.150^{\pm.000}$	$9.210^{\pm.000}$	-
	Salad [15]	$0.076^{\pm.002}$	$\mathbf{0.581}^{\pm.003}$	$\mathbf{0.769}^{\pm.003}$	$\mathbf{0.857}^{\pm.002}$	$\mathbf{2.649}^{\pm.009}$	$9.696^{\pm.096}$	$1.751^{\pm.062}$
	sMDM [3]	$0.130^{\pm.000}$	$0.494^{\pm.000}$	$0.682^{\pm.000}$	$0.776^{\pm.000}$	$3.051^{\pm.000}$	$9.663^{\pm.000}$	-
	COME [26]	$0.041^{\pm.002}$	$0.526^{\pm.003}$	$0.723^{\pm.002}$	$0.816^{\pm.002}$	$2.898^{\pm.006}$	$9.532^{\pm.062}$	$1.704^{\pm.059}$
	RAM (Ours)	$\mathbf{0.032}^{\pm.002}$	$0.561^{\pm.003}$	$0.751^{\pm.002}$	$0.839^{\pm.002}$	$2.716^{\pm.007}$	$9.487^{\pm.084}$	$1.142^{\pm.038}$

4.2 Implementation Details

We adopt exactly the same text and motion encoders as those in TEMOS [31]. Both of them are implemented as 6-layer, encoder-only transformers. The text encoder takes the text embeddings extracted by DistilBERT [37] as input. The latent dimensionality is set to 256 for HumanML3D and 192 for KIT-ML. For

Table 2: Quantitative results of text-to-motion generation on the KIT-ML test set. Methods are grouped into VQ-VAE-based and diffusion-based categories. In each group, the best results are highlighted in **bold**, while the second-best results are underlined. Notably, on this challenging, smaller dataset, RAM achieves the most balanced performance among diffusion models, ranking second in both FID and R-Precision. It substantially outperforms the FID leader (ReMoDiffuse) in R-Precision and the R-Precision leader (Salad) in FID.

	Method	FID↓	R-Precision↑			MM Dist↓	Diversity	MM
			Top 1	Top 2	Top 3			
	Ground Truth	$0.031^{\pm.004}$	$0.424^{\pm.005}$	$0.649^{\pm.006}$	$0.779^{\pm.006}$	$2.788^{\pm.012}$	$11.08^{\pm.097}$	-
VQ-VAE-based	T2M-GPT [49]	$0.512^{\pm.029}$	$0.416^{\pm.006}$	$0.627^{\pm.006}$	$0.745^{\pm.006}$	$3.007^{\pm.023}$	$10.92^{\pm.108}$	$1.856^{\pm.011}$
	MMM [34]	$0.316^{\pm.028}$	$0.404^{\pm.005}$	$0.621^{\pm.005}$	$0.744^{\pm.004}$	$2.977^{\pm.019}$	$10.91^{\pm.101}$	$1.232^{\pm.039}$
	MoMask [12]	$0.204^{\pm.011}$	$0.433^{\pm.007}$	$0.656^{\pm.005}$	$0.781^{\pm.005}$	$2.779^{\pm.022}$	-	$1.131^{\pm.043}$
	BAMM [33]	$0.183^{\pm.013}$	$0.438^{\pm.009}$	$0.661^{\pm.009}$	$0.788^{\pm.005}$	$2.723^{\pm.026}$	$11.01^{\pm.094}$	$1.609^{\pm.065}$
	MoGenTS [48]	$0.143^{\pm.004}$	$0.445^{\pm.006}$	$0.671^{\pm.006}$	$0.797^{\pm.005}$	$2.711^{\pm.024}$	$10.92^{\pm.090}$	-
	BAD [16]	$0.221^{\pm.012}$	$0.417^{\pm.006}$	$0.631^{\pm.006}$	$0.750^{\pm.006}$	$2.941^{\pm.025}$	$11.00^{\pm.100}$	$1.170^{\pm.047}$
	LaMP [23]	$\mathbf{0.141}^{\pm.013}$	$\mathbf{0.479}^{\pm.006}$	$\mathbf{0.691}^{\pm.005}$	$\mathbf{0.826}^{\pm.005}$	$\mathbf{2.704}^{\pm.018}$	$10.93^{\pm.101}$	-
Diffusion-based	MDM [44]	$0.547^{\pm.070}$	$0.404^{\pm.002}$	$0.616^{\pm.013}$	$0.737^{\pm.005}$	$3.074^{\pm.018}$	$10.75^{\pm.203}$	$1.806^{\pm.180}$
	MLD [4]	$0.404^{\pm.027}$	$0.390^{\pm.008}$	$0.609^{\pm.008}$	$0.734^{\pm.007}$	$3.204^{\pm.027}$	$10.80^{\pm.117}$	$2.192^{\pm.071}$
	ReMoDiffuse [51]	$\mathbf{0.155}^{\pm.006}$	$0.427^{\pm.014}$	$0.641^{\pm.004}$	$0.765^{\pm.055}$	$2.814^{\pm.012}$	$10.80^{\pm.105}$	$1.239^{\pm.028}$
	FineMoGen [52]	$0.178^{\pm.007}$	$0.432^{\pm.006}$	$0.649^{\pm.005}$	$0.772^{\pm.006}$	$2.869^{\pm.014}$	$10.85^{\pm.115}$	$1.877^{\pm.093}$
	StableMoFusion [17]	$0.258^{\pm.029}$	$0.445^{\pm.006}$	$0.660^{\pm.005}$	$0.782^{\pm.004}$	-	$10.94^{\pm.077}$	$1.362^{\pm.062}$
	Salad [15]	$0.296^{\pm.012}$	$\mathbf{0.477}^{\pm.006}$	$\mathbf{0.711}^{\pm.005}$	$\mathbf{0.828}^{\pm.005}$	$\mathbf{2.585}^{\pm.016}$	$11.10^{\pm.095}$	$1.004^{\pm.040}$
	RAM (Ours)	$0.172^{\pm.010}$	$0.464^{\pm.006}$	$0.684^{\pm.006}$	$0.803^{\pm.005}$	$2.653^{\pm.024}$	$11.15^{\pm.101}$	$1.010^{\pm.043}$

the diffusion model that generates motion sequences from the latents, we use the MDM architecture [44], consisting of an 8-layer, encoder-only transformer backbone with a latent size of 512. Training is performed with a batch size of 64, a learning rate of 0.0001, and the AdamW optimizer. Models are trained for 450K steps on HumanML3D and 400K steps on KIT-ML. We set the loss weights w_{sr} to 1 and w_{latent} to 0.5. During training, we set the diffusion timestep T to 50, with 10% of conditional latents replaced by zero vectors for classifier-free guidance. For inference on HumanML3D, we use 20 denoising steps, spaced linearly in $[0, \ldots, T-1]$, resulting in 20 inference steps. The weights for REG and classifier-free guidance are set to 5.0 and 1.5, respectively.

4.3 Comparison with State-of-the-Art Methods

We quantitatively compare RAM with state-of-the-art (SOTA) methods on HumanML3D and KIT-ML. The results are shown in Tab. 1 and Tab. 2, respectively. As demonstrated in Tab. 1, RAM achieves SOTA performance on the most widely used HumanML3D benchmark. Compared with diffusion-based methods, RAM shows a substantial improvement in FID and achieves near-SOTA performance in semantic accuracy as measured by R-Precision, ranking just behind Salad. When compared to VQ-VAE-based approaches, RAM surpasses them in semantic accuracy and achieves highly competitive FID—*a feat not previously attained by diffusion-based methods*. RAM thus demonstrates that diffusion-based motion generation models can reach state-of-the-art FID levels.

Table 3: Incremental ablation experiments on the key components of RAM. We evaluate the impact of the motion reconstruction branch, Self-Regularization (L_{sr}), Motion-Centric Latent Alignment (L_{latent}), and Reconstructive Error Guidance (REG). The results show that each component makes an essential contribution, with the full model achieving the best performance on both FID and R-Precision. In each column, the best result is highlighted in **bold** and the second-best is underlined.

Components				FID↓	R-Precision↑			MM D↓	Diversity
E_m	L_{latent}	L_{sr}	REG		Top 1	Top 2	Top 3		
				$0.786^{\pm.016}$	$0.417^{\pm.002}$	$0.613^{\pm.002}$	$0.729^{\pm.003}$	$3.433^{\pm.012}$	$10.063^{\pm.076}$
✓				$0.624^{\pm.013}$	$0.493^{\pm.004}$	$0.695^{\pm.002}$	$0.800^{\pm.002}$	$3.045^{\pm.013}$	$10.188^{\pm.096}$
✓	✓			$0.187^{\pm.005}$	$0.530^{\pm.002}$	$0.720^{\pm.002}$	$0.814^{\pm.002}$	$2.885^{\pm.008}$	$9.703^{\pm.077}$
✓	✓	✓		$\underline{0.132}^{\pm.005}$	$\mathbf{0.561}^{\pm.002}$	$\mathbf{0.752}^{\pm.002}$	$\underline{0.838}^{\pm.002}$	$\underline{2.744}^{\pm.006}$	$9.754^{\pm.067}$
✓	✓	✓	✓	$\mathbf{0.032}^{\pm.002}$	$\mathbf{0.561}^{\pm.003}$	$\underline{0.751}^{\pm.002}$	$\mathbf{0.839}^{\pm.002}$	$\mathbf{2.716}^{\pm.007}$	$9.487^{\pm.084}$

On the KIT-ML dataset, the limited training scale poses significant challenges for generative modeling, particularly for diffusion-based architectures. Consequently, RAM, similar to the leading diffusion-based method Salad [15], exhibits a performance variance compared to larger benchmarks. Nevertheless, RAM remains highly competitive, achieving strong performance in both motion realism and semantic alignment.

Inference Efficiency. Although REG adds one reconstruction forward per step, RAM stays efficient: with full REG at 20 steps it runs at 0.398 s AITS—faster than MDM at 50 steps (0.490 s)—yet reaches a far lower FID, and restricting REG to the first 6 steps adds only 26% time for a 71% FID drop (0.132→0.038); see Appendix B of the supplementary material for more details.

4.4 Ablation Studies

To assess the impact of key design choices within RAM, we conduct comprehensive ablation studies on HumanML3D. Specifically, these studies include: (1) *Incremental Experiments*—starting from a baseline model, we progressively introduce key design components, culminating in the complete RAM; and (2) *Guidance Evaluation*—we examine the effectiveness of our proposed Reconstructive Error Guidance and the additional benefits achieved when it is combined with classifier-free guidance (CFG). In addition, we provide further experiments on the influence of internal loss parameters β and τ, loss weights w_{st} and w_{latent}, and encoder latent dimensionality in Appendix C of the supplementary material.

Incremental Experiments. Tab. 3 presents the results of our incremental studies. To rigorously assess the contribution of each component, we begin with a clean baseline consisting of RAM's text encoder and diffusion model only. We keep these modules exactly the same as those in RAM and progressively add key components. The baseline demonstrates limited performance, partly due to the challenging inference setting of only 20 steps. Introducing the motion encoder E_m—which forms a dual-branch architecture, similar to a direct application of

Language2pose [2] to diffusion models—provides a modest improvement, suggesting that implicitly learning a joint language-motion space is suboptimal. Incorporating L_{latent} delivers substantial further gains, though still falls short of state-of-the-art performance. Adding L_{sr} leads to results that surpass most diffusion-based approaches reported in Tab. 1. Finally, enabling REG elevates RAM to state-of-the-art performance. Collectively, these findings demonstrate the necessity and effectiveness of each design choice.

Table 4: Effect of Reconstructive Error Guidance (REG) and classifier-free guidance (CFG). w_1 and w_2 respectively control the influence of REG and CFG. The results demonstrate that CFG substantially improves semantic accuracy (higher R-Precision), while REG notably enhances motion realism (lower FID). Furthermore, combining both strategies enables RAM to achieve state-of-the-art performance. In each column, the best result is highlighted in **bold** and the second-best is underlined.

w_1	w_2	FID↓	R-Precision↑ Top 1	Top 2	Top 3	MM D↓	Diversity
0.0	0.0	$0.297^{\pm.010}$	$0.529^{\pm.002}$	$0.719^{\pm.002}$	$0.812^{\pm.002}$	$2.912^{\pm.007}$	$9.707^{\pm.072}$
3.0	0.0	$\underline{0.088}^{\pm.005}$	$0.542^{\pm.002}$	$0.732^{\pm.002}$	$0.825^{\pm.002}$	$2.808^{\pm.007}$	$9.418^{\pm.075}$
4.0	0.0	$0.097^{\pm.005}$	$0.539^{\pm.002}$	$0.729^{\pm.001}$	$0.823^{\pm.001}$	$2.817^{\pm.007}$	$9.328^{\pm.071}$
5.0	0.0	$0.128^{\pm.006}$	$0.537^{\pm.002}$	$0.726^{\pm.001}$	$0.820^{\pm.001}$	$2.837^{\pm.008}$	$9.242^{\pm.068}$
0.0	1.5	$0.132^{\pm.005}$	$0.561^{\pm.002}$	$0.752^{\pm.002}$	$0.838^{\pm.002}$	$2.744^{\pm.006}$	$9.754^{\pm.067}$
0.0	2.5	$0.107^{\pm.004}$	$\mathbf{0.563}^{\pm.002}$	$\underline{0.753}^{\pm.002}$	$\mathbf{0.840}^{\pm.002}$	$2.729^{\pm.006}$	$9.711^{\pm.066}$
0.0	3.5	$0.097^{\pm.004}$	$\underline{0.562}^{\pm.002}$	$\mathbf{0.754}^{\pm.002}$	$\mathbf{0.840}^{\pm.002}$	$\underline{2.727}^{\pm.006}$	$9.668^{\pm.067}$
0.0	4.5	$0.095^{\pm.003}$	$\underline{0.562}^{\pm.002}$	$0.752^{\pm.002}$	$\underline{0.839}^{\pm.002}$	$2.734^{\pm.007}$	$9.633^{\pm.068}$
5.0	1.5	$\mathbf{0.032}^{\pm.002}$	$0.561^{\pm.003}$	$0.751^{\pm.002}$	$\underline{0.839}^{\pm.002}$	$\mathbf{2.716}^{\pm.007}$	$9.487^{\pm.084}$

Guidance Evaluation. Tab. 4 presents the results of our guidance evaluation. The findings show that classifier-free guidance (CFG) substantially improves semantic accuracy, as measured by R-Precision. In contrast, our proposed reconstructive error guidance (REG) notably enhances the overall realism of the generated motion, reflected by better FID scores. Furthermore, combining both strategies enables RAM to achieve state-of-the-art performance.

4.5 Comprehensive Comparison of Latent Space Training Strategies

In Sec. 4.4, we conducted a preliminary comparison between directly applying dual-branch architecture to diffusion models and our reconstruction-anchored approach to demonstrate the superiority of RAM over traditional two-stream methods. Our reconstruction-anchored approach is characterized by two key components: (1) self-regularization performed in the motion latent space, and (2) motion-centric latent alignment that aligns text representations with the motion latent space, where this motion latent space is learned through motion reconstruction. In this section, we conduct a more comprehensive comparison across multiple variants to provide stronger empirical evidence supporting our approach.

First, we introduce several variants for comparison:

- Two-stream baseline (Variant A): This variant represents the most basic approach where reconstruction and generation branches share a latent space without directly imposing any loss on the latent space, only implicitly learning a joint language–motion embedding space. A representative method is Language2Pose [2].
- Bidirectional latent alignment (Variant B and C): This variant directly applies alignment loss between motion and text latent variables in the latent space. A representative method is TEMOS [31]. We provide two weight configurations: a loss weight of 1.0 (Variant B) and a loss weight of 10^{-5} (Variant C), where the latter reproduces the parameter settings from the original TEMOS paper.
- Cross-modal contrastive learning (Variant D): This variant applies cross-modal contrastive learning in the latent space, which constitutes one key objective of TMR [32] and KinMo [53].
- Cross-modal contrastive learning & bidirectional latent alignment (Variant E): This variant simultaneously applies both cross-modal contrastive learning and bidirectional latent alignment in the latent space. TMR is a representative method using this approach. To reproduce TMR's parameter settings, we set the contrastive learning temperature $\tau = 0.1$ with a weight of 0.1, and the bidirectional latent alignment weight to 10^{-5}.
- Ours: This variant employs self-regularization and motion-centric alignment.

Implementation details of these variants, including network architectures, hyperparameters, and inference configurations, are provided in Appendix D of the supplementary material.

Table 5: Comparison of latent space training strategies. **Variant A**: Two-stream baseline. **Variant B**: Bidirectional latent alignment (weight=1.0). **Variant C**: Bidirectional latent alignment (weight=10^{-5}). **Variant D**: Cross-modal contrastive learning. **Variant E**: Cross-modal contrastive learning & bidirectional latent alignment. The results indicate that while Variant E achieves semantic accuracy (R-Precision) comparable to ours, Our method significantly outperforms it in terms of FID. This demonstrates that our motion-centric alignment better preserves the intrinsic structure of the motion manifold compared to bidirectional or contrastive approaches.

Variant	FID↓	R-Precision↑ Top 1	Top 2	Top 3	MM Dist↓	Diversity
Variant A	$0.624^{\pm.013}$	$0.493^{\pm.004}$	$0.695^{\pm.002}$	$0.800^{\pm.002}$	$3.045^{\pm.013}$	$10.188^{\pm.096}$
Variant B	$0.776^{\pm.015}$	$0.415^{\pm.002}$	$0.610^{\pm.003}$	$0.724^{\pm.002}$	$3.430^{\pm.012}$	$10.029^{\pm.084}$
Variant C	$0.540^{\pm.008}$	$0.494^{\pm.002}$	$0.696^{\pm.002}$	$0.799^{\pm.002}$	$3.020^{\pm.008}$	$10.029^{\pm.064}$
Variant D	$0.708^{\pm.015}$	$0.419^{\pm.004}$	$0.626^{\pm.003}$	$0.740^{\pm.003}$	$3.349^{\pm.011}$	$10.107^{\pm.077}$
Variant E	$0.218^{\pm.006}$	$0.558^{\pm.003}$	$0.750^{\pm.002}$	$0.839^{\pm.002}$	$2.767^{\pm.010}$	$9.820^{\pm.074}$
Ours	$0.132^{\pm.005}$	$0.561^{\pm.002}$	$0.752^{\pm.002}$	$0.838^{\pm.002}$	$2.744^{\pm.006}$	$9.754^{\pm.067}$

As shown in Tab. 5, directly applying bidirectional latent alignment (Variant B) and cross-modal contrastive learning (Variant D) to the diffusion-based two-stream baseline fails to yield performance improvements and actually results in

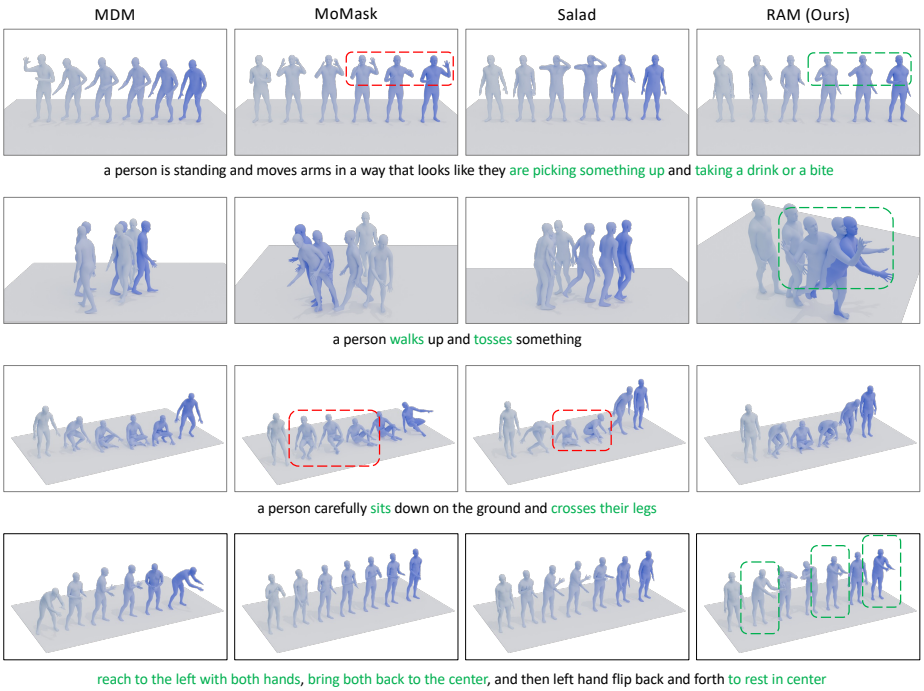

Fig. 3: Qualitative evaluation on the HumanML3D Dataset. Actions corresponding to the text are highlighted with green dashed lines, while unnatural artifacts are indicated with red dashed lines. It can be observed that baseline methods often fail to faithfully execute the entire set of actions described in the text. For example, in the second row ("a person walks up and tosses something"), most methods only execute the walking motion. Some outputs also exhibit distortions, such as unnatural drifting (in the third row, MoMask during sitting down and Salad during standing up) and error patterns (in the first row, MoMask's hands move erratically up and down after completing the "drink" action). The fourth row presents the greatest challenge, involving four actions. Our method completes at least three, whereas others accomplish only one or two. Further comparisons can be found in the **supplementary video**.

degradation. This indicates that incorporating these techniques into two-stream methods requires careful hyperparameter tuning. When employing hyperparameters empirically determined from prior work (Variant C [31] and Variant E [32]), performance gains are observed. Notably, Variant E achieves R-Precision comparable to our approach. However, a substantial gap remains between Variant E and our method in terms of FID. This suggests that our motion-centric approach better preserves the intrinsic structure of the motion space, demonstrating superior motion fidelity.

4.6 Qualitative Evaluation

We compare the qualitative results of RAM with those generated by three representative or leading text-to-motion methods: MDM [44], MoMask [12], and Salad [15]. Figure 3 illustrates four groups of comparisons, with the corresponding text descriptions shown below each group. The text prompts are sampled from the test set. Since each prompt consists of multiple actions, this setting poses a significant challenge for accurate motion generation. The results show that our method achieves a high degree of semantic accuracy and realism. Additional qualitative examples are provided in the **supplementary video**, and we further validate the effectiveness of our method from a perceptual perspective through a **user study** in Appendix A of the supplementary material.

5 Conclusion

In this work, we present **RAM**, a novel framework that leverages motion reconstruction to regularize text-driven motion diffusion models. Our approach focuses on learning a motion-centric latent space via motion reconstruction, specifically designed to capture essential motion dynamics while achieving high semantic resolution. This latent space serves as intermediate supervision for text-to-motion generation, bridging the representational gap between abstract language and high-dimensional, kinematically constrained human motion. We further present Reconstructive Error Guidance, a technique that mitigates error propagation during sampling by exploiting the motion diffusion model's self-correcting ability. Experiments show that RAM achieves state-of-the-art performance on standard benchmarks. One possible limitation of RAM is that we adopt a single token to represent the entire textual description. In the future, we plan to try finer-grained conditions that contain multiple tokens for long-horizon motion generation.

Acknowledgments. This work was supported by the National Natural Science Foundation of China under Grants 62476099 and 62076101, Guangdong Basic and Applied Basic Research Foundation under Grants 2024B1515020082 and 2023A1515010007, and the TCL Young Scholars Program.

References

1. Ahn, D., Cho, H., Min, J., Jang, W., Kim, J., Kim, S., Park, H.H., Jin, K.H., Kim, S.: Self-rectifying diffusion sampling with perturbed-attention guidance. In: European Conference on Computer Vision (ECCV). pp. 1–17. Springer (2024)
2. Ahuja, C., Morency, L.P.: Language2pose: Natural language grounded pose forecasting. In: International Conference on 3D Vision (3DV). pp. 719–728. IEEE (2019)
3. Bae, J., Hwang, I., Lee, Y.Y., Guo, Z., Liu, J., Ben-Shabat, Y., Kim, Y.M., Kapadia, M.: Less is more: Improving motion diffusion models with sparse keyframes. arXiv preprint arXiv:2503.13859 (2025)

4. Chen, X., Jiang, B., Liu, W., Huang, Z., Fu, B., Chen, T., Yu, G.: Executing your commands via motion diffusion in latent space. In: Computer Vision and Pattern Recognition (CVPR). pp. 18000–18010 (2023)
5. Cohan, S., Tevet, G., Reda, D., Peng, X.B., van de Panne, M.: Flexible motion in-betweening with diffusion models. In: International Conference on Computer Graphics and Interactive Techniques (SIGGRAPH). pp. 1–9 (2024)
6. Dai, W., Chen, L.H., Huo, Y., Wang, J., Liu, J., Dai, B., Tang, Y.: Real-time controllable motion generation via latent consistency model. Hugging Face Blog (2024), https://huggingface.co/blog/wxDai/motionlcm-v2
7. Dai, W., Chen, L.H., Wang, J., Liu, J., Dai, B., Tang, Y.: Motionlcm: Real-time controllable motion generation via latent consistency model. In: European Conference on Computer Vision (ECCV). pp. 390–408. Springer (2024)
8. Dai, Y., Zhu, W., Li, R., et al.: Tcdiff++: An end-to-end trajectory-controllable diffusion model for harmonious music-driven group choreography. International Journal of Computer Vision (IJCV) **134**(1), 61 (2026). https://doi.org/10.1007/s11263-025-02611-3, https://doi.org/10.1007/s11263-025-02611-3
9. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Conference on Neural Information Processing Systems (NeurIPS) **34**, 8780–8794 (2021)
10. Du, Y., Kips, R., Pumarola, A., Starke, S., Thabet, A., Sanakoyeu, A.: Avatars grow legs: Generating smooth human motion from sparse tracking inputs with diffusion model. In: Computer Vision and Pattern Recognition (CVPR). pp. 481–490 (2023)
11. Ghosh, A., Cheema, N., Oguz, C., Theobalt, C., Slusallek, P.: Synthesis of compositional animations from textual descriptions. In: International Conference on Computer Vision (ICCV). pp. 1396–1406 (2021)
12. Guo, C., Mu, Y., Javed, M.G., Wang, S., Cheng, L.: Momask: Generative masked modeling of 3d human motions. In: Computer Vision and Pattern Recognition (CVPR). pp. 1900–1910 (2024)
13. Guo, C., Zou, S., Zuo, X., Wang, S., Ji, W., Li, X., Cheng, L.: Generating diverse and natural 3d human motions from text. In: Computer Vision and Pattern Recognition (CVPR). pp. 5152–5161 (2022)
14. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022)
15. Hong, S., Kim, C., Yoon, S., Nam, J., Cha, S., Noh, J.: Salad: Skeleton-aware latent diffusion for text-driven motion generation and editing. In: Computer Vision and Pattern Recognition (CVPR). pp. 7158–7168 (2025)
16. Hosseyni, S.R., Rahmani, A.A., Seyedmohammadi, S.J., Seyedin, S., Mohammadi, A.: Bad: Bidirectional auto-regressive diffusion for text-to-motion generation. In: International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5. IEEE (2025)
17. Huang, Y., Yang, H., Luo, C., Wang, Y., Xu, S., Zhang, Z., Zhang, M., Peng, J.: Stablemofusion: Towards robust and efficient diffusion-based motion generation framework. International Conference on Multimedia (MM) (2024)
18. Jin, P., Wu, Y., Fan, Y., Sun, Z., Yang, W., Yuan, L.: Act as you wish: Fine-grained control of motion diffusion model with hierarchical semantic graphs. Advances in Neural Information Processing Systems (NeurIPS) **36**, 15497–15518 (2023)
19. Karras, T., Aittala, M., Kynkäänniemi, T., Lehtinen, J., Aila, T., Laine, S.: Guiding a diffusion model with a bad version of itself. Advances in Neural Information Processing Systems (NeurIPS) **37**, 52996–53021 (2024)
20. Li, C., Xie, X., Cao, Y., Geiger, A., Pons-Moll, G.: Frankenmotion: Part-level human motion generation and composition. arXiv preprint arXiv:2601.10909 (2026)

21. Li, J., Clegg, A., Mottaghi, R., Wu, J., Puig, X., Liu, C.K.: Controllable human-object interaction synthesis. In: European Conference on Computer Vision (ECCV). pp. 54–72. Springer (2024)
22. Li, Y., van der Schaar, M.: On error propagation of diffusion models. In: International Conference on Learning Representations (ICLR) (2024)
23. Li, Z., Yuan, W., He, Y., Qiu, L., Zhu, S., Gu, X., Shen, W., Dong, Y., Dong, Z., Yang, L.T.: Lamp: Language-motion pretraining for motion generation, retrieval, and captioning. International Conference on Learning Representations (ICLR) (2025)
24. Liang, H., Bao, J., Zhang, R., Ren, S., Xu, Y., Yang, S., Chen, X., Yu, J., Xu, L.: Omg: Towards open-vocabulary motion generation via mixture of controllers. In: Computer Vision and Pattern Recognition (CVPR). pp. 482–493 (2024)
25. Liang, H., Zhang, W., Li, W., Yu, J., Xu, L.: Intergen: Diffusion-based multi-human motion generation under complex interactions. International Journal of Computer Vision (IJCV) pp. 1–21 (2024)
26. Lyu, G., Cheng, X., Liu, Q., Xu, C., Yan, J., Yang, M., Fang, F., Deng, C.: Come: Advancing representation learning and generative modeling for high-quality text-to-motion generation. OpenReview (2025), https://openreview.net/forum?id=D8fW5tAHq8
27. Mahmood, N., Ghorbani, N., Troje, N.F., Pons-Moll, G., Black, M.J.: AMASS: Archive of motion capture as surface shapes. In: International Conference on Computer Vision (ICCV). pp. 5442–5451 (2019)
28. Maldonado, G., Pazho, A.D., Noghre, G.A., Katariya, V., Tabkhi, H.: Moclip motion-aware fine-tuning and distillation of clip for human motion generation. In: Computer Vision and Pattern Recognition (CVPR). pp. 2931–2941 (2025)
29. Ni, J., Abrego, G.H., Constant, N., Ma, J., Hall, K.B., Cer, D., Yang, Y.: Sentence-t5: Scalable sentence encoders from pre-trained text-to-text models. arXiv preprint arXiv:2108.08877 (2021)
30. Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., Chen, M.: Glide: Towards photorealistic image generation and editing with text-guided diffusion models. arXiv preprint arXiv:2112.10741 (2021)
31. Petrovich, M., Black, M.J., Varol, G.: Temos: Generating diverse human motions from textual descriptions. In: European Conference on Computer Vision (ECCV). pp. 480–497. Springer (2022)
32. Petrovich, M., Black, M.J., Varol, G.: Tmr: Text-to-motion retrieval using contrastive 3d human motion synthesis. In: International Conference on Computer Vision (ICCV). pp. 9488–9497 (2023)
33. Pinyoanuntapong, E., Saleem, M.U., Wang, P., Lee, M., Das, S., Chen, C.: Bamm: Bidirectional autoregressive motion model. In: European Conference on Computer Vision (ECCV). pp. 172–190. Springer (2024)
34. Pinyoanuntapong, E., Wang, P., Lee, M., Chen, C.: Mmm: Generative masked motion model. In: Computer Vision and Pattern Recognition (CVPR). pp. 1546–1555 (2024)
35. Plappert, M., Mandery, C., Asfour, T.: The kit motion-language dataset. Big data **4**(4), 236–252 (2016)
36. 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 (ICML). pp. 8748–8763. PmLR (2021)
37. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108 (2019)

38. Sheng, J., Lin, M., Zhao, A., Pruvost, K., Wen, Y.H., Li, Y., Huang, G., Liu, Y.J.: Exploring text-to-motion generation with human preference. In: Computer Vision and Pattern Recognition (CVPR). pp. 1888–1899 (2024)
39. Stability AI: Stable diffusion 3.5. `https://github.com/Stability-AI/sd3.5` (2024)
40. Sun, M., Xu, C., Jiang, X., Liu, Y., Sun, B., Huang, R.: Beyond talking–generating holistic 3d human dyadic motion for communication. International Journal of Computer Vision (IJCV) **133**(5), 2910–2926 (2025)
41. Tan, X., Weng, W., Lei, H., Wang, H.: Easytune: Efficient step-aware fine-tuning for diffusion-based motion generation. In: International Conference on Learning Representations (ICLR) (2026)
42. 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 (ECCV). pp. 358–374. Springer (2022)
43. Tevet, G., Raab, S., Cohan, S., Reda, D., Luo, Z., Peng, X.B., Bermano, A.H., van de Panne, M.: Closd: Closing the loop between simulation and diffusion for multi-task character control. In: International Conference on Learning Representations (ICLR) (2025)
44. Tevet, G., Raab, S., Gordon, B., Shafir, Y., Cohen-Or, D., Bermano, A.H.: Human motion diffusion model. arXiv preprint arXiv:2209.14916 (2022)
45. Wang, Y., Wang, S., Zhang, J., Fan, K., Wu, J., Xue, Z., Liu, Y.: Timotion: Temporal and interactive framework for efficient human-human motion generation. In: Computer Vision and Pattern Recognition (CVPR). pp. 7169–7178 (2025)
46. Weng, W., Tan, X., Wang, J., Xie, G.S., Zhou, P., Wang, H.: Realign: Text-to-motion generation via step-aware reward-guided alignment. In: AAAI Conference on Artificial Intelligence (2026)
47. Xia, F., Li, C., Martín-Martín, R., Litany, O., Toshev, A., Savarese, S.: Relmogen: Integrating motion generation in reinforcement learning for mobile manipulation. In: International Conference on Robotics and Automation (ICRA). pp. 4583–4590. IEEE (2021)
48. 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 (NeurIPS) **37**, 130739–130763 (2024)
49. 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: Computer Vision and Pattern Recognition (CVPR). pp. 14730–14740 (2023)
50. 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)
51. Zhang, M., Guo, X., Pan, L., Cai, Z., Hong, F., Li, H., Yang, L., Liu, Z.: Remodiffuse: Retrieval-augmented motion diffusion model. In: International Conference on Computer Vision (ICCV). pp. 364–373 (2023)
52. Zhang, M., Li, H., Cai, Z., Ren, J., Yang, L., Liu, Z.: Finemogen: Fine-grained spatio-temporal motion generation and editing. Advances in Neural Information Processing Systems (NeurIPS) **36**, 13981–13992 (2023)
53. Zhang, P., Liu, P., Kim, H., Garrido, P., Chaudhuri, B.: Kinmo: Kinematic-aware human motion understanding and generation. International Conference on Computer Vision (ICCV) (2025)