ANT: Adaptive Neural Temporal-Aware Text-to-Motion Model

Wenshuo Chen*
wenshuochen@hkust-gz.edu.cn
The Hong Kong University of Science and Technology (Guangzhou)
Guangzhou, China

Kuimou Yu*
kyu745@connect.hkust-gz.edu.cn
The Hong Kong University of Science and Technology (Guangzhou)
Guangzhou, China

Jia Haozhe*
haozhejia@hkust-gz.edu.cn
The Hong Kong University of Science and Technology (Guangzhou)
Guangzhou, China

Kaishen Yuan
kaishenyuan@hkust-gz.edu.cn
The Hong Kong University of Science and Technology (Guangzhou)
Guangzhou, China

Zexu Huang
Zexu.Huang@student.uts.edu.au
University of Technology Sydney
School of Electrical and Data Engineering (SEDE)
Sydney, NSW, Australia

Bowen Tian
bowentian@hkust-gz.edu.cn
The Hong Kong University of Science and Technology (Guangzhou)
Guangzhou, China

Songning Lai
songninglai@hkust-gz.edu.cn
The Hong Kong University of Science and Technology (Guangzhou)
Guangzhou, China

Hongru Xiao
hongru_xiao@tongji.edu.cn
Tongji University
College of Civil Engineering
Shanghai, China

Erhang Zhang
seve19861@gmail.com
Shandong University
Chongxin
Qingdao, China

Lei Wang
l.wang4@griffith.edu.au
Griffith University
Brisbane, Queensland
Australia
Data61/CSIRO
Canberra, ACT, Australia

Yutao Yue†
yutaoyue@hkust-gz.edu.cn
The Hong Kong University of Science and Technology (Guangzhou)
Thrust of Artificial Intelligence and Thrust of Intelligent Transportation
Guangzhou, China
Institute of Deep Perception Technology
JITRI
Wuxi, China

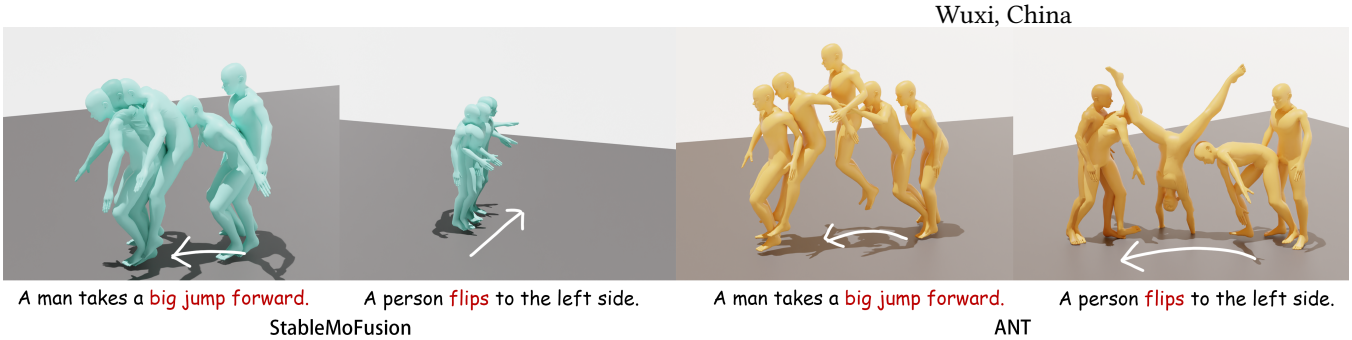

Figure 1: Our ANT can be seamlessly plugged into diffusion-based text-to-motion models to generate semantically rich, fine-grained, and naturally smooth motions with high precision.

Abstract

While diffusion models advance text-to-motion generation, their static semantic conditioning ignores temporal-frequency demands:

*Represents equal contribution.
†Correspondence to Yutao Yue <yutaoyue@hkust-gz.edu.cn>

MM '25, Dublin, Ireland

ACM ISBN 979-8-4007-2035-2/2025/10
https://doi.org/10.1145/3746027.3755168

early denoising requires structural semantics for motion foundations while later stages need localized details for text alignment. This mismatch mirrors biological morphogenesis where developmental phases demand distinct genetic programs. Inspired by epigenetic regulation governing morphological specialization, we propose **(ANT)**, an **A**daptive **N**eural **T**emporal-Aware architecture. ANT orchestrates semantic granularity through: **(i) Semantic Temporally Adaptive (STA) Module:** Automatically partitions denoising into low-frequency structural planning and high-frequency refinement via spectral analysis. **(ii) Dynamic Classifier-Free Guidance scheduling (DCFG):** Adaptively adjusts conditional to unconditional ratio enhancing efficiency while maintaining fidelity. Extensive experiments show that ANT can be applied to various baselines, significantly improving model performance, and achieving state-of-the-art semantic alignment on StableMoFusion. Code can be found on https://github.com/CCSCovenant/ANT.

"Details make perfection, and perfection is not a detail."
— Leonardo da Vinci

CCS Concepts

• **Computing methodologies** → **Computer vision**.

Keywords

Temporal-Aware Semantic Denoising Process, Text-to-Motion

ACM Reference Format:
Wenshuo Chen, Kuimou Yu, Jia Haozhe, Kaishen Yuan, Zexu Huang, Bowen Tian, Songning Lai, Hongru Xiao, Erhang Zhang, Lei Wang, and Yutao Yue. 2025. ANT: Adaptive Neural Temporal-Aware Text-to-Motion Model. In *Proceedings of the 33rd ACM International Conference on Multimedia (MM '25), October 27–31, 2025, Dublin, Ireland.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3746027.3755168

1 Introduction

Text-driven human motion generation has recently attracted significant attention due to the semantic richness and intuitive nature of natural language descriptions, with broad applications in animation, film production, virtual/augmented reality (VR/AR), and robotics [3, 4, 30–32]. While textual prompts provide valuable semantic guidance for motion synthesis, they often suffer from incomplete or inaccurate semantic representations, leading to suboptimal generation quality [4, 39]. Ensuring faithful alignment between generated motions and textual descriptions remains a critical challenge.

Current research in text-to-motion generation primarily focuses on two paradigms: VAE-based models that encode motions into discrete tokens for prediction using autoregressive (AR) [22, 45] or non-autoregressive (NAR) [8, 31, 32] frameworks, and diffusion-based models that gradually transform Gaussian noise into motion sequences through iterative denoising under text conditioning [5, 6, 20, 40, 48]. Among these approaches, vector quantization (VQ)-based discrete generation methods have become the dominant paradigm in human motion synthesis [8, 31]. However, despite their effectiveness, these methods suffer from inherent drawbacks, including information loss and reduced motion diversity [11, 45]. Conversely, diffusion-based approaches offer unique advantages, such as fine-grained detail generation, diverse motion outputs, and physically plausible movement synthesis, making them a promising alternative [3, 40, 42, 46]. Nevertheless, diffusion-based methods still lag behind VQ-based models in terms of overall performance [20, 40].

Recent efforts have sought to bridge this performance gap by enhancing motion representations within diffusion models [3, 17]. One notable approach involves projecting motion data into a compact and fine-grained latent space using a 1D ResNet-based autoencoder, thereby improving motion structure modeling and prediction accuracy [17]. While such methods mitigate some of the limitations of VAE-based approaches, they remain significantly inferior to state-of-the-art techniques. Moreover, recent studies [4] reveal that diffusion-based text-to-motion models exhibit limitations in aligning generated motions faithfully with input text descriptions. To better understand the generative process of diffusion-based models, prior work [3] has analyzed the denoising mechanism and proposed a two-phase generation framework: a semantic planning stage for low-frequency feature modeling and a fine-grained refinement stage for high-frequency generating. This decomposition raises a crucial research question: **What distinct roles does textual information play in these two phases? How can we enhance the semantic alignment of diffusion-based approaches to achieve more accurate and expressive motion synthesis?**

Drawing inspiration from biological processes, Hinton likened the metamorphosis of insects to different stages of computational learning: larval stages prioritize energy absorption, while adult stages focus on locomotion and reproduction [15]. Analogously, the diffusion process in motion generation should follow a phase-wise prioritization, capturing low-frequency motion structures in early denoising steps and refining high-frequency motion details in later stages. However, existing methods [3, 20, 40] apply CLIP-encoded [35] text embeddings uniformly across all denoising steps, failing to distinguish between these two phases. This uniform application can lead to incomplete motion structures in early stages and insufficient detail refinement in later stages. However, explicitly decomposing semantic information in the frequency domain remains a significant challenge.

To address these issues, we propose ANT (Adaptive Neural Temporal-Aware Text-to-Motion Model). Unlike conventional diffusion based approaches, ANT incorporates a plug-and-play adaptive Semantic Temporal-Aware Module (STA) and a Dynamic schedule method (DCFG) based CFG [16]. STA model dynamically adjusts its response at different timesteps in the denoising process. This module adapts autonomously without requiring explicit supervision, enabling progressive semantic modulation. Specifically, STA prioritizes global motion structures (low-frequency) during the early denoising steps and refines detailed motion variations (high-frequency) in the later stages. For the CFG Schedule, based on STA's ability to distinguish between early and late stages in text attention, we dynamically adjust the ratio between conditional and unconditional results during sampling [24, 37]. In the later stages, we switch to the more efficient unconditional generation. Through these temporally-aware method, our method improves semantic alignment, resulting in more accurate, diverse, and physically plausible motion synthesis. We conducted experiments on MDM [40] and StableMoFusion [20]. The results demonstrate significant improvements in FID and R-Precision. Additionally, by taking

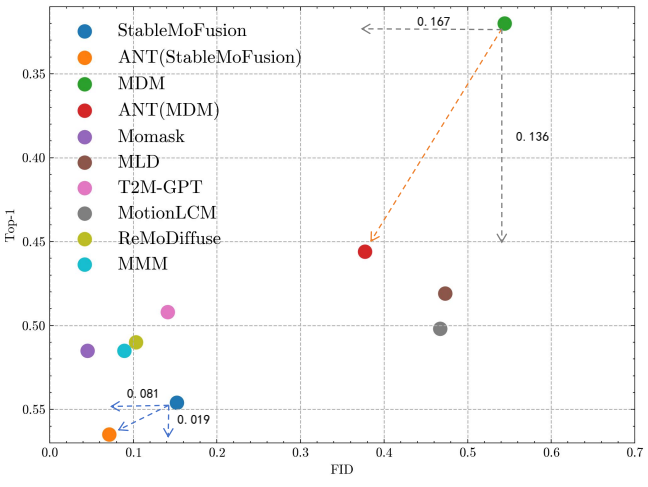

Figure 2: Comparison with SOTA models. The figure presents a comparison with the best-performing text-to-motion models to date, where a closer distance to the origin indicates better overall performance. ANT achieves significant improvements when applied to both MDM and StableMoFusion. Notably, ANT on StableMoFusion outperforms all other models in terms of R-Precision, highlighting its effectiveness and superiority among state-of-the-art methods.

StableMoFusion as its baseline, ANT surpasses representative VAE-based models (Figure 2), such as MMM [32] and T2M-GPT [45], across all metrics, further demonstrating the potential of diffusion-based methods. Our method is simple and efficient, introducing minimal additional computational overhead while being effective across different architectures, including DDIM [21] and DPM-Solver [25]. This provides a novel and interpretable research direction for diffusion-based text-to-motion methods.

We summarize our contributions as follows:

(1) We propose the first dynamic text encoding modulation framework for text-to-motion generation by analyzing the denoising mechanism. The integration of our Semantic Temporal Awareness Module significantly improves alignment between generated motions and textual descriptions.
(2) We conduct an in-depth analysis of the denoising mechanism in text-to-motion generation and provide a new DCFG schedule Method to improve sampling efficiency.
(3) Extensive experiments demonstrate that our method achieves state-of-the-art performance in text-motion alignment, validated by quantitative metrics and user studies.

2 Related Work

Text-to-Motion Generation. Recent advancements in text conditioned human motion generation have been driven by two primary methodologies: diffusion-based models and VAE-based models. Diffusion models [5, 6, 38, 39, 47, 48] have shown remarkable potential in modeling the complex relationships between textual inputs and motion sequences. Prominent works include MotionDiffuse [47], which leverages cross-attention for text integration; MDM [40], which explores diverse denoising networks such as Transformer and GRU; PhysDiff [43], which incorporates physical constraints to enhance realism. While ReMoDiffuse [48] improves performance through retrieval mechanisms, MotionLCM [6] achieves real-time, controllable generation via a latent consistency model.

On the other hand, VAE-based models have also demonstrated strong performance in multi-modal motion generation. ACTOR [28] proposes a Transformer-based VAE for generating motion from predefined action categories, while TEMOS [29] extends ACTOR with an additional text encoder to support diverse motion sequences, primarily focusing on short sentences. Guo et al. [9] introduce an autoregressive conditional VAE that conditions on generated frames and text features, predicting motions based on text length. TEACH [2] builds upon TEMOS to enable the generation of longer, temporally coherent motion compositions from sequential natural language descriptions. TM2T [12] jointly trains both text-to-motion and motion-to-text tasks, improving bidirectional generation quality. T2M-GPT [45] quantizes motion clips into discrete markers and utilizes a transformer-based model to generate subsequent markers.

Despite the advancements in both diffusion-based and VAE-based approaches, a common limitation persists: reliance on the CLIP encoder [35]. Many existing methods, including diffusion-based models [20, 40, 47] and VAE-based models like MotionCLIP [38], process textual descriptions through CLIP to obtain fixed text feature representations. However, this static encoding fails to provide rich, dynamic semantic information throughout the motion generation process. As a result, models struggle to adaptively interpret textual nuances over time, leading to inconsistencies in generated motions and limiting expressiveness.

Senabtic embedding of Text-to-Motion In the field of text-to-motion synthesis, a prevalent strategy involves leveraging the CLIP text encoder [35] to derive semantic embeddings. This approach is adopted by models such as MoMask [8] and StableMoFusion [20]. Alternative methodologies incorporate pretrained word embeddings in conjunction with sequence processing layers, typically Transformers or Gated Recurrent Units (GRUs) [12], as demonstrated in the TM2T framework. More recent developments have seen the integration of novel encoders: MotionGPT [22] employs the T5 model, while MDM [40], initially utilizing CLIP, has subsequently experimented with BERT as an alternative text representation module.

However, each of these encoding strategies exhibits distinct limitations. Specifically, the GRU-based encoder architecture, as implemented in TM2T [12], encounters difficulties with processing extended sequences, capturing long-range dependencies effectively, and maintaining computational efficiency. These shortcomings can potentially impair the nuanced understanding of complex textual prompts and consequently reduce the fidelity of the generated motion sequences. Although encoders predicated on CLIP, T5 [36] and BERT [7] generally yield more robust textual representations, a significant constraint stems from the static nature of their output embeddings throughout the iterative diffusion denoising process. As these embeddings do not dynamically evolve in relation to the diffusion time step, they may not sufficiently address the global conditioning requirements intrinsic to sophisticated generative modeling paradigms. Contrasting existing techniques our method dynamically modulates text embeddings per timestep via a Semantic Temporal-Aware Module (STA). This module aligns adaptive semantic evolution with the denoising process boosting motion-text alignment and generation fidelity.

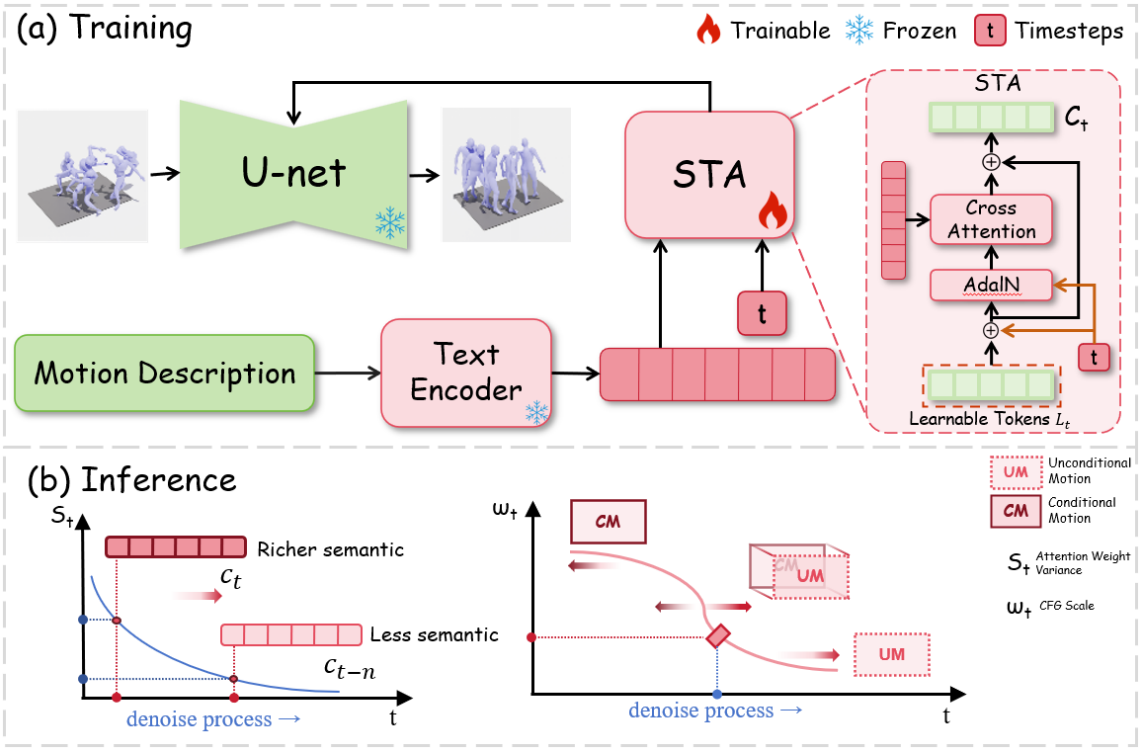

Figure 3: Overall Architecture of ANT. In part (a) Training, we introduce the Semantic Temporal Awareness (STA) module. STA is inserted between the text encoder and the U-Net, dynamically modulating the text features by attending to the diffusion timestep t. In part (b) Inference, we observe that STA enables the model's attention to textual semantics to gradually decrease as sampling progresses. Based on this, we propose a CFG process that aligns with dynamic semantic adjustment: the CFG scale is progressively reduced during sampling, and a more efficient unconditional generation is applied in the later stages.

3 Method

3.1 Preliminaries

3.1.1 ***Text-to-Motion Process.*** We follow the diffusion framework of StableMoFusion [20] for text-conditioned human motion generation. Let $c \in \mathbb{R}^{d_c}$ be a textual description encoded by a pretrained language model, where d_c denotes the text embedding dimension. The target motion sequence $\mathbf{x}_0 \in \mathbb{R}^{N \times d_m}$ consists of N frames, where each frame contains d_m-dimensional motion parameters. The forward diffusion process progressively adds Gaussian noise to $\mathbf{x}_0$ over T timesteps:

$$\mathbf{x}_t = \sqrt{\alpha_t}\mathbf{x}_0 + \sqrt{1-\alpha_t}\boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{1}$$

where $\{\alpha_t\}_{t=1}^T \in (0,1)^T$ is the noise schedule with $\alpha_t = \prod_{s=1}^t (1-\beta_s)$ and β_s as the variance schedule.

In the reverse process, a motion-denoising network G_θ parameterized by θ is trained to predict the original motion $\mathbf{x}_0$ from the noisy input $\mathbf{x}_t$, conditioned on the timestep t and text embedding $\mathbf{c}$:

$$\min_\theta \mathbb{E}_{t,\mathbf{x}_0,\boldsymbol{\epsilon},\mathbf{c}}\left[\|G_\theta(\mathbf{x}_t, t, \mathbf{c}) - \mathbf{x}_0\|_2^2\right]. \tag{2}$$

During inference, $\mathbf{x}_0$ is generated by iteratively denoising from $\mathbf{x}_T \sim \mathcal{N}(0, \mathbf{I})$ using a sampler such as DPM-Solver++ [25], which accelerates generation by reducing the number of reverse steps from $T = 1000$ to as few as 10 steps. Given a textual prompt $\mathcal{T}$, the embedded text vector $\mathbf{c} = \text{CLIP}(\mathcal{T})$ is cached to avoid redundant computations. Classifier-Free Guidance (CFG) is applied to trade off text-motion alignment and motion fidelity:

$$G_s(\mathbf{x}_t, t, \mathbf{c}) = G_\theta(\mathbf{x}_t, t, \emptyset) + \omega \cdot \big(G_\theta(\mathbf{x}_t, t, \mathbf{c}) - G_\theta(\mathbf{x}_t, t, \emptyset)\big), \tag{3}$$

where ω is the guidance scale and $G_\theta(\mathbf{x}_t, t, \emptyset)$ denotes unconditional prediction. Half-precision floating-point computation (FP16) further accelerates inference without compromising quality. This pipeline ensures efficient and high-quality generation of N-frame motion sequences within a fraction of the original computational cost.

3.1.2 ***The Denoising Process.*** In diffusion models, low-frequency components are recovered earlier than high-frequency components during the reverse denoising process [34]. Formally, given a motion signal in the frequency domain $\hat{m}_t(\omega)$, the signal-to-noise ratio (SNR) at frequency ω is defined as:

$$\text{SNR}(\omega) = \frac{|\hat{m}_0(\omega)|^2}{\int_0^t g^2(s)\, ds}, \tag{4}$$

where $\hat{m}_0(\omega)$ is the initial power spectral density of the motion signal and $\int_0^t g^2(s)\, ds$ represents the accumulated noise energy. For higher frequencies ω_H, the SNR decreases more rapidly than for lower frequencies ω_L, i.e.,

$$\text{SNR}(\omega_H) < \text{SNR}(\omega_L), \quad \forall\, \omega_H > \omega_L. \tag{5}$$

As a result, low-frequency components are restored first, providing a semantic foundation for the subsequent recovery of high-frequency details.

3.1.3 ***Low-Frequency Structure Dependence on High-Frequency Motion.*** In the reverse denoising process of diffusion models for motion generation, the accurate restoration of high-frequency motion components is structurally dependent on the consistent reconstruction of low-frequency components.

Let $\mathbf{m}_t$ denote the motion signal representation at timestep t. Let its frequency domain representation be partitioned into low-frequency components $\hat{\mathbf{m}}_{t,L}$ (corresponding to frequencies $\omega \in \Omega_L$) and high-frequency components $\hat{\mathbf{m}}_{t,H}$ (corresponding to frequencies $\omega \in \Omega_H$, where typically $\min(\Omega_H) > \max(\Omega_L)$). The reverse diffusion step estimates the posterior distribution $p(\mathbf{m}_{t-1}|\mathbf{m}_t)$.

The dependency implies that the uncertainty regarding the high-frequency components $\hat{\mathbf{m}}_{t-1,H}$ at step $t-1$, given the noisy observation $\mathbf{m}_t$, is reduced when conditioned on the concurrently estimated low-frequency components $\hat{\mathbf{m}}_{t-1,L}$. Formally, this relationship can be expressed via conditional variance reduction:

$$\mathbb{E}_{\hat{\mathbf{m}}_{t-1,L}\sim p(\cdot|\mathbf{m}_t)}\left[\mathrm{Var}[\hat{\mathbf{m}}_{t-1,H} \mid \mathbf{m}_t, \hat{\mathbf{m}}_{t-1,L}]\right] < \mathrm{Var}[\hat{\mathbf{m}}_{t-1,H} \mid \mathbf{m}_t]. \quad (6)$$

This inequality holds assuming that $\hat{\mathbf{m}}_{t-1,H}$ and $\hat{\mathbf{m}}_{t-1,L}$ are not conditionally independent given $\mathbf{m}_t$. This dependency arises because the denoising network, in estimating $p(\mathbf{m}_{t-1}|\mathbf{m}_t)$, implicitly leverages the structural information inferred from the more robust low-frequency content within $\mathbf{m}_t$ (and consequently $\hat{\mathbf{m}}_{t-1,L}$) to guide the reconstruction of the high-frequency details $\hat{\mathbf{m}}_{t-1,H}$.

Furthermore, the inherent characteristics of signal corruption during the diffusion process amplify this dependency. The signal-to-noise ratio (SNR) typically decreases with increasing frequency, particularly in later diffusion steps (smaller t):

$$\mathrm{SNR}(\omega_H) \ll \mathrm{SNR}(\omega_L), \quad \text{for typical } \omega_H \in \Omega_H, \omega_L \in \Omega_L. \quad (7)$$

Consequently, the high-frequency components within the noisy signal $\mathbf{m}_t$ are significantly more obscured by noise compared to the low-frequency components. Accurate restoration of $\hat{\mathbf{m}}_{t-1,H}$ therefore relies substantially on the contextual foundation provided by the simultaneously restored low-frequency structure $\hat{\mathbf{m}}_{t-1,L}$. This hierarchical reliance is crucial for generating motions that exhibit both semantic consistency (e.g., fine-grained gestures aligning with overall body posture and action) and spatial coherence (e.g., detailed limb movements respecting the constraints imposed by the larger skeletal configuration).

3.2 ANT Architecture

The ANT architecture (Figure 3) optimizes the denoising steps for both semantic information and Classifier-Free Guidance (CFG). To fully explore the distinct roles of textual semantics in the denoising process of diffusion models, we introduce a Semantic Temporal Awareness (STA) module in Section 3.2.1. This module dynamically adjusts the text embeddings by incorporating timestep features, emphasizing low-frequency semantics in the early denoising stages and focusing on high-frequency details in the later stages. In Section 3.2.2, based on observations of the attention to textual features in the ANT architecture, we design a dynamic planning strategy for Classifier-Free Guidance (CFG) that gradually weakens textual guidance during the denoising process.

3.2.1 ***Semantic Temporal Awareness Module***. Unlike previous approaches [8, 20, 40, 45], which directly use the text features c from a text encoder as conditions for predicting the output, we introduce the STA module. Positioned between the text features and the U-Net, this module is inspired by [18] and aims to design a connector that enhances the informativeness of the conditions used for noise prediction. Below is the specific process of the STA: The Learnable Tokens L are combined with the text features $\mathbf{c}$ from the encoder to form a timestep-modulated feature L_t:

$$L_t = L \oplus z_t. \quad (8)$$

Here, z_t is the time feature and $\oplus$ denotes residual connection. This timestep-modulated feature is then processed through the Adaptive Layer Normalization (AdaLN) [19], which is defined by the following equation:

$$\hat{L}_t = \gamma \cdot \frac{L_t + \alpha(z_t)}{\sigma + \varepsilon} + \beta, \quad (9)$$

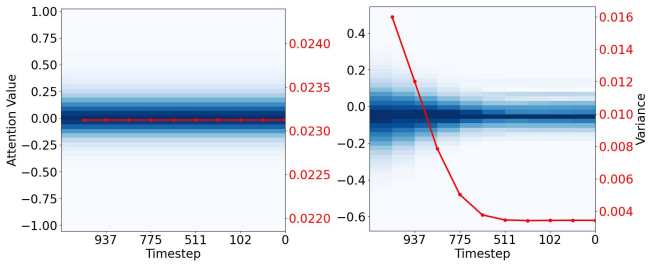

Figure 4: Distribution of attention weights in the UNet cross-attention module over the entire test set. The left panel corresponds to the baseline, and the right panel represents our method. In each panel, the red line denotes the variance of the attention weights, while the blue heatmap illustrates their distribution, with darker colors indicating a higher frequency of occurrence. The heatmap values have been normalized to a range of 0 to 1.

where $\gamma, \alpha(t), \sigma$, and β are learnable parameters, and ε is a small constant to prevent division by zero. The output, after being processed through CrossAttention and a residual connection, is integrated with temporal information to produce the semantic feature with time-dependent text feature $\mathbf{c}_t$:

$$\hat{c}_t = \hat{L}_t \oplus \mathrm{CrossAttention}(c, \hat{L}_t). \quad (10)$$

This process allows the model to dynamically incorporate both the temporal and semantic aspects, and unsupervisedly leverage the denoising prior to learn the semantic features at different time steps.

3.2.2 ***Adaptive Guidance Schedule***. Classifier-Free Guidance (CFG) is a widely used technique in generative models (such as diffusion models) that enhances conditional generation by computing both conditional and unconditional outputs and interpolating between them using a guidance scale. While prior work [20, 40] typically adopts a fixed guidance scale (e.g., $\omega = 4.5$) to improve semantic alignment, our denoising observations reveal that a static scale fails to adapt to the varying needs of different stages in the generation process.

As illustrated in the right of Figure 4, we visualize the distribution of attention values to text features in the ANT U-Net across the entire test set (in blue), and additionally report the variance of the attention values (in red). We observe that, as denoising progresses, the attention values increasingly concentrate around 0. This indicates that in the early stages, the network tends to focus more on capturing semantic information—reflected in a broader and more dispersed distribution of attention values across different tokens—while in the later stages, semantic attention diminishes, and the model gradually shifts toward unconditional generation. These insights suggest that stronger guidance should be applied during the early denoising steps, followed by a progressive reduction in guidance intensity, ultimately favoring unconditioned generation as the process converges.

Moreover, CFG increases computational overhead by requiring both conditional and unconditional generations. From an efficiency perspective, we further analyze that when the number of denoising steps is sufficiently large, the model has already captured sufficient semantic information. At this stage, the primary objective shifts to refining fine-grained, high-frequency details. Compared to conditional generation, unconditional generation is better at capturing the natural and fluid distribution of motion. Furthermore, based on Theorem 3.1.3, high-frequency information generation relies heavily on low-frequency components. This insight motivates us to forgo CFG in the later stages of denoising and instead perform more efficient unconditional generation.

To implement this adaptive guidance mechanism, we define a time-dependent guidance scale ω_t as follows:

Method	Venue	FID ↓	R-Precision ↑			Diversity →	MM-Dist↓	Multimodality↑
			Top1	Top2	Top3			
HumanML3D								
Ground Truth	-	$0.002^{\pm0.000}$	$0.511^{\pm0.003}$	$0.703^{\pm0.003}$	$0.797^{\pm0.002}$	$9.503^{\pm0.065}$	-	-
MLD [5]	CVPR 2023	$0.473^{\pm0.013}$	$0.481^{\pm0.003}$	$0.673^{\pm0.003}$	$0.772^{\pm0.002}$	$9.724^{\pm0.082}$	$3.196^{\pm0.010}$	$2.413^{\pm0.079}$
ReMoDiffuse [48]	ICCV 2023	$0.103^{\pm0.004}$	$0.510^{\pm0.005}$	$0.698^{\pm0.006}$	$0.795^{\pm0.004}$	$9.018^{\pm0.075}$	$3.025^{\pm0.008}$	$1.795^{\pm0.043}$
MotionDiffuse [47]	TPAMI 2024	$0.630^{\pm0.001}$	$0.491^{\pm0.001}$	$0.681^{\pm0.001}$	$0.782^{\pm0.001}$	$9.410^{\pm0.049}$	$3.113^{\pm0.001}$	$1.553^{\pm0.042}$
MotionLCM [6]	ECCV 2024	$0.467^{\pm0.012}$	$0.502^{\pm0.003}$	$0.701^{\pm0.002}$	$0.803^{\pm0.002}$	$9.361^{\pm0.660}$	$3.012^{\pm0.007}$	$2.172^{\pm0.082}$
T2M-GPT [45]	CVPR 2023	$0.141^{\pm0.005}$	$0.492^{\pm0.003}$	$0.679^{\pm0.002}$	$0.775^{\pm0.002}$	$9.722^{\pm0.082}$	$3.121^{\pm0.009}$	$1.831^{\pm0.048}$
MMM [32]	CVPR 2024	$0.089^{\pm0.002}$	$0.515^{\pm0.002}$	$0.708^{\pm0.002}$	$0.804^{\pm0.002}$	$9.577^{\pm0.050}$	$2.926^{\pm0.007}$	$1.226^{\pm0.035}$
MoMask [8]	CVPR 2024	$0.045^{\pm0.002}$	$0.521^{\pm0.002}$	$0.713^{\pm0.002}$	$0.807^{\pm0.002}$	-	$2.958^{\pm0.008}$	$1.241^{\pm0.040}$
MDM [40]	ICLR 2023	$0.544^{\pm0.044}$	$0.320^{\pm0.005}$	$0.498^{\pm0.004}$	$0.611^{\pm0.007}$	$9.559^{\pm0.086}$	$5.556^{\pm0.027}$	$2.799^{\pm0.072}$
StableMoFusion [20]	ACM MM 2024	$0.152^{\pm0.004}$	$0.546^{\pm0.002}$	$0.742^{\pm0.002}$	$0.835^{\pm0.002}$	$9.466^{\pm0.002}$	$2.781^{\pm0.011}$	$1.362^{\pm0.062}$
StableMoFusion Efficiency	ACM MM 2024	$2.845^{\pm0.027}$	$0.401^{\pm0.003}$	$0.599^{\pm0.003}$	$0.719^{\pm0.003}$	$8.699^{\pm0.098}$	-	$2.276^{\pm0.065}$
ANT (Ours, on MDM)	-	$0.377^{\pm0.038}$	$0.456^{\pm0.006}$	$0.656^{\pm0.007}$	$0.763^{\pm0.006}$	$9.886^{\pm0.068}$	-	$2.595^{\pm0.006}$
ANT (Ours, on StableMoFusion)	-	$0.099^{\pm0.004}$	$0.560^{\pm0.002}$	$0.751^{\pm0.003}$	$0.841^{\pm0.002}$	$9.585^{\pm0.090}$	$2.8748^{\pm0.015}$	$1.874^{\pm0.062}$
ANT (Ours, on StableMoFusion, w/o DCFG)	-	$0.071^{\pm0.004}$	$0.565^{\pm0.006}$	$0.756^{\pm0.003}$	$0.843^{\pm0.002}$	$9.585^{\pm0.090}$	$2.763^{\pm0.016}$	$1.827^{\pm0.110}$
KIT-ML								
Ground Truth	-	$0.031^{\pm0.004}$	$0.424^{\pm0.005}$	$0.649^{\pm0.006}$	$0.779^{\pm0.006}$	$11.080^{\pm0.097}$	-	-
MLD [5]	CVPR 2023	$0.404^{\pm0.027}$	$0.390^{\pm0.008}$	$0.609^{\pm0.008}$	$0.734^{\pm0.007}$	$10.800^{\pm0.117}$	$3.204^{\pm0.027}$	$2.192^{\pm0.071}$
ReMoDiffuse [48]	ICCV 2023	$0.155^{\pm0.006}$	$0.427^{\pm0.014}$	$0.641^{\pm0.004}$	$0.765^{\pm0.055}$	$10.800^{\pm0.105}$	$1.239^{\pm0.028}$	$1.239^{\pm0.028}$
MotionDiffuse [47]	TPAMI 2024	$1.954^{\pm0.062}$	$0.417^{\pm0.004}$	$0.621^{\pm0.004}$	$0.739^{\pm0.004}$	$11.100^{\pm0.143}$	$2.958^{\pm0.005}$	$0.730^{\pm0.013}$
T2M-GPT [45]	CVPR 2023	$0.514^{\pm0.029}$	$0.416^{\pm0.006}$	$0.627^{\pm0.006}$	$0.745^{\pm0.006}$	$10.921^{\pm0.108}$	$3.007^{\pm0.023}$	$1.570^{\pm0.039}$
MotionGPT [22]	NeurIPS 2023	$0.510^{\pm0.016}$	$0.366^{\pm0.005}$	$0.558^{\pm0.004}$	$0.680^{\pm0.005}$	$10.350^{\pm0.084}$	$3.527^{\pm0.021}$	$2.328^{\pm0.117}$
MMM [32]	CVPR 2024	$0.316^{\pm0.028}$	$0.404^{\pm0.005}$	$0.621^{\pm0.005}$	$0.744^{\pm0.004}$	$10.910^{\pm0.101}$	$2.977^{\pm0.019}$	$1.232^{\pm0.039}$
MoMask [8]	CVPR 2024	$0.204^{\pm0.011}$	$0.433^{\pm0.007}$	$0.656^{\pm0.005}$	$0.781^{\pm0.005}$	-	$2.779^{\pm0.022}$	$1.131^{\pm0.043}$
MDM [40]	ICLR 2023	$0.497^{\pm0.021}$	$0.164^{\pm0.004}$	$0.291^{\pm0.004}$	$0.396^{\pm0.004}$	$10.847^{\pm0.119}$	$9.191^{\pm0.022}$	$1.907^{\pm0.214}$
StableMoFusion [20]	ACM MM 2024	$0.258^{\pm0.029}$	$0.445^{\pm0.006}$	$0.660^{\pm0.005}$	$0.782^{\pm0.004}$	$10.936^{\pm0.077}$	$2.800^{\pm0.018}$	$1.362^{\pm0.062}$
ANT (Ours)	-	$0.236^{\pm0.015}$	$0.465^{\pm0.007}$	$0.694^{\pm0.006}$	$0.813^{\pm0.005}$	$11.029^{\pm0.102}$	$2.689^{\pm0.020}$	$1.578^{\pm0.056}$

Table 1: Quantitative comparison on the HumanML3D and KIT-ML datasets. ± indicates a 95% confidence interval. ↓: Lower is better. ↑: Higher is better. →: Closer to the Ground Truth (GT) is better. Red and Blue indicate the best and the second-best results respectively across all methods for each metric. Our method, ANT, demonstrates state-of-the-art or highly competitive performance across multiple key metrics on both datasets.

$$\omega_t = \omega_{\min} + \phi(t)(\omega_{\max} - \omega_{\min}), \tag{11}$$

where $\omega_{\max}$ and $\omega_{\min}$ denote the maximum and minimum guidance scales respectively, and $\phi(t)$ is a monotonically decreasing function that controls the decay of guidance strength over the denoising timestep t. In this paper, We leverage cosine schedule as $\phi(t)$.

$$\omega_t = max\{\omega_{\min} + \frac{1}{2}\left(1 + \cos(\lambda \frac{T-t}{T}\pi)\right)(\omega_{\max} - \omega_{\min}), 0\}. \tag{12}$$

Here, λ denotes the period coefficient, and T represents the total number of time steps. Accordingly, the noise prediction at timestep t with Classifier-Free Guidance is computed as:

$$\hat{\epsilon}_{\text{CFG}} = \hat{\epsilon}_{\text{uncond}} + \omega_t \cdot (\hat{\epsilon}_{\text{cond}} - \hat{\epsilon}_{\text{uncond}}), \tag{13}$$

where $\hat{\epsilon}_{\text{cond}}$ and $\hat{\epsilon}_{\text{uncond}}$ represent the conditional and unconditional denoising predictions, respectively.

To further improve efficiency during the late stages of denoising where semantic conditioning becomes less influential, we omit the conditional branch altogether when t exceeds a certain threshold (e.g., $t > 0.5T$). In such cases, we use:

$$\hat{\epsilon}_t = \hat{\epsilon}_{\text{uncond}}. \tag{14}$$

Thereby reducing computational overhead while promoting the generation of coherent high-frequency details aligned with the learned data distribution.

4 Experiments

We evaluate our approach on two standard motion-language benchmarks: HumanML3D [10] and KIT-ML [33]. HumanML3D con- tains 14,616 motion sequences from AMASS [26] and HumanAct12 [13], each paired with three text descriptions (44,970 total), covering diverse actions like walking, exercising, and dancing. KIT-ML in- cludes 3,911 motions and 6,278 descriptions, serving as a smaller-scale benchmark. We follow StableMoFusion's pose representation and apply mirroring-based augmentation. Data is split into training, validation, and test sets with a ratio of 0.8:0.15:0.05. **Evaluation Metrics.** In addition to the commonly utilized metrics such as Frechet Inception Distance (FID) [14], R-Precision, Multimodal Distance (MM-Dist), and Diversity, which are employed by StableMoFusion [20]. Furthermore, human evaluation is employed to obtain accuracy and human preference results for the outputs generated by the model.

4.1 Experimental Setup

We adopt model architecture settings similar to those of MDM [40] and StableMoFusion [20]. For MDM, we use a batch size of 64 and the AdamW [23] optimizer. Our models are trained with a time step of $T = 50$, following a cosine noise schedule. The total number of training iterations is fixed at 120,000, with a learning rate of 1×10^{-4}. For StableMoFusion, we adhere to its training methodology, running for 200,000 steps. The time step is set to $T = 50$, and we employ DPM-Solver for inference, using 10 actual sampling steps. The λ is set to 1.5 and unconditional generation start time is set to 0.5T. ω_{max} and ω_{min} is set to 3.0 and 1.5, The optimal values for ω_{max}

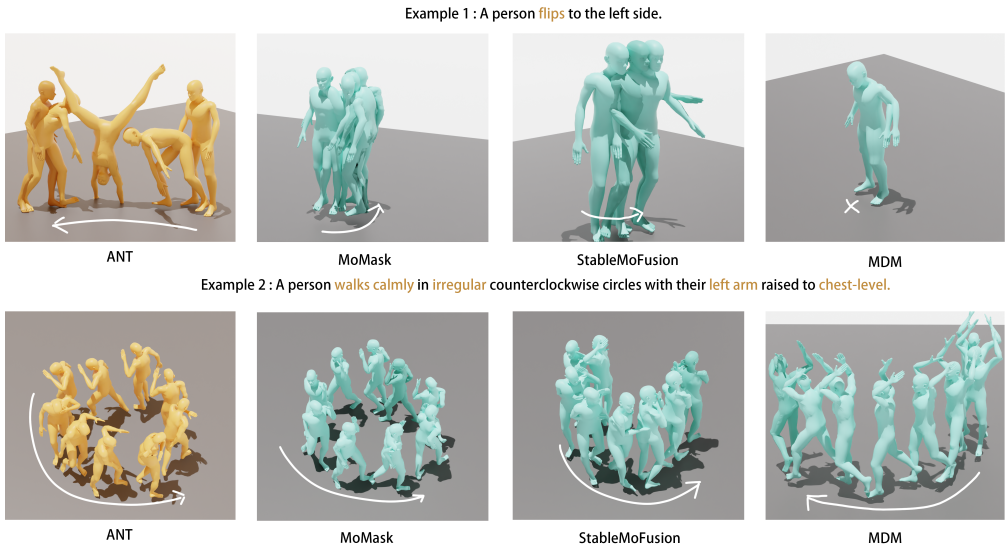

Figure 5: Visualization Comparison. We compare the visual results of ANT with other three state-of-the-art methods. In both examples, ANT consistently demonstrates more accurate, natural, and fine-grained motion generation compared to the others.

Method	Top1	Top2	Top3
BERT (w/o STA)	$0.547^{\pm0.003}$	$0.742^{\pm0.002}$	$0.835^{\pm0.002}$
BERT (w STA)	$0.551^{\pm0.002}$	$0.747^{\pm0.003}$	$0.836^{\pm0.002}$
MoCLIP (w/o STA)	$0.528^{\pm0.002}$	$0.718^{\pm0.002}$	$0.812^{\pm0.002}$
MoCLIP (w STA)	$0.554^{\pm0.003}$	$0.747^{\pm0.003}$	$0.838^{\pm0.002}$
LongCLIP (w/o STA)	$0.528^{\pm0.002}$	$0.718^{\pm0.002}$	$0.812^{\pm0.002}$
LongCLIP (w STA)	$0.513^{\pm0.003}$	$0.702^{\pm0.002}$	$0.797^{\pm0.003}$
CLIP (w/o STA)	$0.538^{\pm0.003}$	$0.730^{\pm0.002}$	$0.822^{\pm0.002}$
CLIP (w STA)	$0.523^{\pm0.003}$	$0.717^{\pm0.002}$	$0.812^{\pm0.002}$
T5 (w/o STA)	$0.549^{\pm0.004}$	$0.747^{\pm0.004}$	$0.838^{\pm0.003}$
T5 (w STA)	$0.565^{\pm0.006}$	$0.756^{\pm0.003}$	$0.843^{\pm0.002}$

Table 2: Experimental results for different methods on text-to-motion generation. The term "w STA" indicates the use of an STA. All values are reported as mean and ± indicates a 95% confidence interval.

and ω_{min} are selected via grid search on the validation set, selecting those that achieved the best Top-1 performance (See detail in the appendix F), respectively. The entire training process can be efficiently executed on a single RTX 4090 GPU with 24 GB of memory.

4.2 Comparison to State-of-the-art Approaches

Quantitative comparisons. Following [8, 45], we report the average over 20 repeated generations with a 95% confidence interval. Table 1 presents evaluations on the HumanML3D [10] and KIT-ML [33] datasets, respectively, in comparison with state-of-the-art (SOTA) approaches.

In terms of improvement over the baseline, ANT demonstrates strong performance, achieving substantial gains on metrics such as FID (HumanML3D: **0.071** vs. 0.152; KIT-ML: **0.236** vs. 0.258) and R-Precision (Top-1 HumanML3D: **0.565** vs. 0.546; KIT-ML: **0.465** vs. 0.445). This indicates that ANT can significantly enhance the performance of the baseline model. When compared with SOTA VQ-based models such as MMM and MoMask, our method still yields competitive results on HumanML3D and KIT-ML. Although our FID is slightly higher than that of MoMask, our R-Precision surpasses all existing SOTA models. It is worth noting that the baseline model we used performs significantly worse than SOTA methods like MMM. However, with our proposed ANT enhancements, we achieve competitive results.

Method	Average Time (s)
ANT (w/o DCFG)	0.949
Baseline method (w/o DCFG)	0.878
ANT (w DCFG)	0.741
Baseline method (w DCFG)	0.717

Table 3: Average Processing Times per 32-Batch (use T5 text encoder).

Qualitative comparison. Figure 5 shows the visual comparison between our model and other state-of-the-art methods under the same prompts. For the prompt "A person flips to the left side," MDM, StableMoFusion, and MoMask all produce incorrect motions, indicating a lack of semantic alignment. In contrast, our model generates natural and smooth motion that accurately reflects the input text. For the more detailed prompt, "A person walks calmly in irregular counterclockwise with their left arm raised to chest-level", StableMoFusion and MDM incorrectly raises the arm, while MoMask not only overlooks the meaning of "irregular" but also raises the left arm higher than chest level. Our model successfully captures all the fine-grained information and produces motion that remains faithful to the prompt. Overall, the visualization results demonstrate that our approach outperforms baselines and other models in terms of semantic accuracy, motion fluency, and attention to detail.

ANT Boosts Performance on Complex Text Descriptions. To evaluate our model's performance on semantically rich text, we select prompts from the HumanML3D test set that are longer than the average length and contain at least two verbs or include adverbs within a single sentence. Table 4 reports the results before and after applying ANT. As shown, ANT significantly improves all metrics on long and fine-grained prompts compared to the baseline. This demonstrates the strong advantages of our model in text comprehension, fine-grained motion generation, and handling long textual inputs.

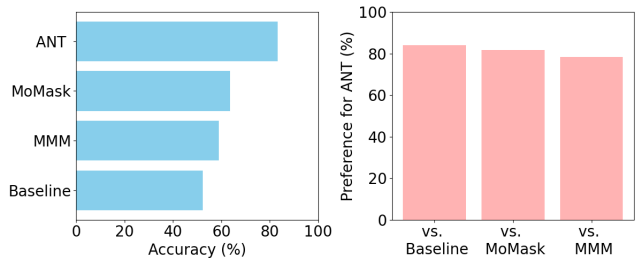

Figure 6: Comparative performance of ANT (StableMoFusion) versus other baseline methods based on human evaluation. The left panel depicts the accuracy achieved during manual assessments. The right panel illustrates the win rate when compared against other baseline methods. vs.Baseline means against original StableMoFusion.

Plug-and-Play Architecture. We validate the effectiveness of ANT not only on the StableMoFusion (DPM-Solver) but also on the MDM (DDIM). The results are shown in Table 1. ANT (MDM) achieves significant improvements over the baseline in both FID and R-Precision (FID: 0.377 vs. 0.544; Top-1: 0.456 vs. 0.320). These results confirm the adaptability of our method to different diffusion architectures and demonstrate its potential as a practical enhancement for diffusion-based text-to-motion models.

Efficiency. In table 3, we compare the efficiency of two our methods and baseline methods under the same text encoder and batch conditions. Our approach (ANT) shows a relatively significant performance improvement over the baseline, given the relatively small overhead of its implementation. Furthermore, based on insights gained from the ANT method, we applied unconditional generation at a later stage to improve efficiency, as indicated by the comparison between ANT and baseline method with efficiency sampling. Under conditions of limited performance degradation, this approach provides a significant enhancement, representing a favorable trade-off between efficiency and performance. In summary, the experimental results confirm that the ANT-based method is effective for achieving notable performance gains with minimal additional overhead, making it a viable solution for applications demanding both high performance and efficiency.

Human Evaluation. We randomly selected 100 text descriptions from the HumanML3D test set and invited 20 participants to subjectively evaluate the motion sequences generated by four models: ANT (StableMoFusion), StableMoFusion, MoMask, and MMM. The participants, unaware of the model names, were asked to rate whether the generated motions were semantically accurate, providing a "yes" or "no" response. We then calculated the average accuracy of motion generation for each model. Additionally, in the Pairwise Preference task, for each text description, we presented the results from ANT and one of the other three models (StableMoFusion, MoMask, or MMM), and asked the participants to choose the more natural and coherent motion from the two options. Each model pair was evaluated 100 times, resulting in a total of 300 binary comparisons. The Figure 6 shows that ANT leads in motion accuracy, achieving 83.2%. In terms of subjective preference, ANT also consistently outperformed the other models, securing an average preference rate of 81.3%. ANT demonstrated significantly stronger semantic understanding and superior motion generation quality.

4.3 Ablation Study

Analysis of Architectural Contributions. As shown in Table 1, we validate the effectiveness of STA. The improvement brought by STA alone over StableMoFusion has been discussed in Section 4.2. In Table 3, the sampling time for a single batch is reduced from 0.949s to 0.741s when using DCFG.

Method	FID	Top1	Top2	Top3	MultiModality
GT	$0.015^{\pm0.008}$	$0.4643^{\pm0.006}$	$0.661^{\pm0.005}$	$0.757^{\pm0.005}$	-
StableMoFusion	$1.316^{\pm0.040}$	$0.394^{\pm0.002}$	$0.576^{\pm0.004}$	$0.680^{\pm0.003}$	$1.873^{\pm0.052}$
ANT	$0.372^{\pm0.017}$	$0.505^{\pm0.009}$	$0.700^{\pm0.006}$	$0.794^{\pm0.007}$	$2.009^{\pm0.066}$

Table 4: Experimental results for fine-grained generation of ANT. All values are reported as mean and ± indicates a 95% confidence interval.

Method	FID	Top1	Top2	Top3
Baseline	$0.152^{\pm0.004}$	$0.546^{\pm0.002}$	$0.742^{\pm0.002}$	$0.835^{\pm0.002}$
Resampler	$0.101^{\pm0.004}$	$0.563^{\pm0.003}$	$\mathbf{0.756^{\pm0.003}}$	$\mathbf{0.844^{\pm0.002}}$
Abstractor	$0.071^{\pm0.004}$	$\mathbf{0.565^{\pm0.006}}$	$\mathbf{0.756^{\pm0.003}}$	$0.843^{\pm0.002}$

Table 5: Experimental results for different architecture of STA. All values are reported as mean and ± indicates a 95% confidence interval.

This results in a 21.9% increase in efficiency, with almost no loss in accuracy. This demonstrates the effectiveness of the DCFG when integrated into the STA architecture.

We also observe that applying DCFG directly to the baseline does not work. The baseline fails to distinguish between the early and late stages of the denoising process (Figure 4, due to its lack of temporal text awareness. This comparison further shows that the ANT architecture can effectively leverage the denoising prior in diffusion to improve prediction quality.

Text Encoder. In Table 2, we investigate the impact of applying ANT to various text encoders. We observe consistent performance improvements when using T5 [36], BERT [7], and MoCLIP. In contrast, models based on CLIP and LongCLIP [44] show performance degradation. This discrepancy can be attributed to the varying capacities of these models to capture fine-grained textual information. Large-scale encoders such as BERT and T5 benefit from rich pretraining, enabling them to generate detailed text representations. These representations facilitate dynamic modulation across time steps and support hierarchical, nuanced semantic understanding. On the other hand, CLIP and LongCLIP tend to produce coarser textual features, which often leads to semantic misalignment during dynamic processing. In contrast, MoCLIP demonstrates stronger alignment with motion semantics, as it has been fine-tuned on motion-specific datasets. This makes it more suitable for tasks that require precise and temporally coherent semantic modulation. Based on these findings, we adopt T5 as the text encoder in our work, as it delivers the strongest overall performance.

STA Architecture. In Table 5, we explore two different architectures for the STA module. We conduct experiments on StableMoFusion and evaluate performance using FID and R-Precision. Both STA architectures demonstrate significant improvements over the baseline across all metrics while achieving comparable results in semantic alignment (R-Precision). Notably, Abstractor [41] outperforms Resampler [1] in terms of FID (**0.071** vs. 0.101). Therefore, we adopt Abstractor in this work as the method for integrating semantic features and temporal steps.

5 Conclusion

In this paper, we design semantic temporal-aware methods for both the training (STA) and inference (DCFG) stages, based on the unique denoising mechanism of diffusion. Our method provides plug-and-play functionality, achieving more precise semantic alignment and more efficient sampling, demonstrating the potential of diffusion-based methods in T2M.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. 2022. Flamingo: a Visual Language Model for Few-Shot Learning. arXiv:2204.14198 [cs.CV] https://arxiv.org/abs/2204.14198

[2] Nikos Athanasiou, Mathis Petrovich, Michael J. Black, and Gül Varol. 2022. TEACH: Temporal Action Composition for 3D Humans. arXiv:2209.04066 [cs.CV] https://arxiv.org/abs/2209.04066

[3] Wenshuo Chen, Haozhe Jia, Songning Lai, Keming Wu, Hongru Xiao, Lijie Hu, and Yutao Yue. 2025. Free-T2M: Frequency Enhanced Text-to-Motion Diffusion Model With Consistency Loss. arXiv:2501.18232 [cs.CV] https://arxiv.org/abs/2501.18232

[4] Wenshuo chen, Hongru Xiao, Erhang Zhang, Lijie Hu, Lei Wang, Mengyuan Liu, and Chen Chen. 2024. SATO: Stable Text-to-Motion Framework. In Proceedings of the 32nd ACM International Conference on Multimedia (Melbourne VIC, Australia) (MM '24). Association for Computing Machinery, New York, NY, USA, 6989–6997. doi:10.1145/3664647.3681034

[5] Xin Chen, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, Jingyi Yu, and Gang Yu. 2023. Executing your Commands via Motion Diffusion in Latent Space. arXiv:2212.04048 [cs.CV] https://arxiv.org/abs/2212.04048

[6] Wenxun Dai, Ling-Hao Chen, Jingbo Wang, Jinpeng Liu, Bo Dai, and Yansong Tang. 2024. MotionLCM: Real-time Controllable Motion Generation via Latent Consistency Model. arXiv:2404.19759 [cs.CV] https://arxiv.org/abs/2404.19759

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805 [cs.CL] https://arxiv.org/abs/1810.04805

[8] Chuan Guo, Yuxuan Mu, Muhammad Gohar Javed, Sen Wang, and Li Cheng. 2023. MoMask: Generative Masked Modeling of 3D Human Motions. (2023). arXiv:2312.00063 [cs.CV]

[9] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. 2022. Generating diverse and natural 3d human motions from text. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 5152–5161.

[10] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. 2022. Generating Diverse and Natural 3D Human Motions From Text. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 5152–5161.

[11] Chuan Guo, Xinxin Zuo, Sen Wang, and Li Cheng. 2022. TM2T: Stochastic and Tokenized Modeling for the Reciprocal Generation of 3D Human Motions and Texts. arXiv:2207.01696 [cs.CV] https://arxiv.org/abs/2207.01696

[12] Chuan Guo, Xinxin Zuo, Sen Wang, and Li Cheng. 2022. Tm2t: Stochastic and tokenized modeling for the reciprocal generation of 3d human motions and texts. In European Conference on Computer Vision. Springer, 580–597.

[13] Chuan Guo, Xinxin Zuo, Sen Wang, Shihao Zou, Qingyao Sun, Annan Deng, Minglun Gong, and Li Cheng. 2020. Action2motion: Conditioned generation of 3d human motions. In Proceedings of the 28th ACM International Conference on Multimedia. 2021–2029.

[14] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In Proceedings of the 31st International Conference on Neural Information Processing Systems (Long Beach, California, USA) (NIPS'17). Curran Associates Inc., Red Hook, NY, USA, 6629–6640.

[15] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the Knowledge in a Neural Network. arXiv:1503.02531 [stat.ML] https://arxiv.org/abs/1503.02531

[16] Jonathan Ho and Tim Salimans. 2022. Classifier-Free Diffusion Guidance. arXiv:2207.12598 [cs.LG] https://arxiv.org/abs/2207.12598

[17] Seokhyeon Hong, Chaelin Kim, Serin Yoon, Junghyun Nam, Sihun Cha, and Junyong Noh. 2025. SALAD: Skeleton-aware Latent Diffusion for Text-driven Motion Generation and Editing. arXiv:2503.13836 [cs.CV] https://arxiv.org/abs/2503.13836

[18] Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. 2024. ELLA: Equip Diffusion Models with LLM for Enhanced Semantic Alignment. arXiv:2403.05135 [cs.CV]

[19] Xun Huang and Serge Belongie. 2017. Arbitrary Style Transfer in Real-time with Adaptive Instance Normalization. arXiv:1703.06868 [cs.CV] https://arxiv.org/abs/1703.06868

[20] Yiheng Huang, Hui Yang, Chuanchen Luo, Yuxi Wang, Shibiao Xu, Zhaoxiang Zhang, Man Zhang, and Junran Peng. 2024. StableMoFusion: Towards Robust and Efficient Diffusion-based Motion Generation Framework. arXiv:2405.05691 [cs.CV] https://arxiv.org/abs/2405.05691

[21] Zhihan Huang, Yuting Wei, and Yuxin Chen. 2024. Denoising diffusion probabilistic models are optimally adaptive to unknown low dimensionality. arXiv:2410.18784 [cs.LG] https://arxiv.org/abs/2410.18784

[22] Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. 2023. MotionGPT: Human Motion as a Foreign Language. arXiv:2306.14795 [cs.CV] https://arxiv.org/abs/2306.14795

[23] Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. arXiv:1711.05101 [cs.LG] https://arxiv.org/abs/1711.05101

[24] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. 2022. DPM-Solver: A Fast ODE Solver for Diffusion Probabilistic Model Sampling in Around 10 Steps. arXiv:2206.00927 [cs.LG] https://arxiv.org/abs/2206.00927

[25] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. 2023. DPM-Solver++: Fast Solver for Guided Sampling of Diffusion Probabilistic Models. arXiv:2211.01095 [cs.LG] https://arxiv.org/abs/2211.01095

[26] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. 2019. AMASS: Archive of Motion Capture as Surface Shapes. In International Conference on Computer Vision. 5442–5451.

[27] Zichong Meng, Yiming Xie, Xiaogang Peng, Zeyu Han, and Huaizu Jiang. 2024. Rethinking Diffusion for Text-Driven Human Motion Generation. arXiv:2411.16575 [cs.CV] https://arxiv.org/abs/2411.16575

[28] Mathis Petrovich, Michael J. Black, and Gül Varol. 2021. Action-Conditioned 3D Human Motion Synthesis with Transformer VAE. arXiv:2104.05670 [cs.CV] https://arxiv.org/abs/2104.05670

[29] Mathis Petrovich, Michael J. Black, and Gül Varol. 2022. TEMOS: Generating diverse human motions from textual descriptions. arXiv:2204.14109 [cs.CV] https://arxiv.org/abs/2204.14109

[30] Ekkasit Pinyoanuntapong, Muhammad Usama Saleem, Korrawe Karunratanakul, Pu Wang, Hongfei Xue, Chen Chen, Chuan Guo, Junli Cao, Jian Ren, and Sergey Tulyakov. 2024. ControlMM: Controllable Masked Motion Generation. arXiv:2410.10780 [cs.CV] https://arxiv.org/abs/2410.10780

[31] Ekkasit Pinyoanuntapong, Muhammad Usama Saleem, Pu Wang, Minwoo Lee, Srijan Das, and Chen Chen. 2024. BAMM: Bidirectional Autoregressive Motion Model. arXiv:2403.19435 [cs.CV] https://arxiv.org/abs/2403.19435

[32] Ekkasit Pinyoanuntapong, Pu Wang, Minwoo Lee, and Chen Chen. 2024. MMM: Generative Masked Motion Model. arXiv:2312.03596 [cs.CV] https://arxiv.org/abs/2312.03596

[33] Matthias Plappert, Christian Mandery, and Tamim Asfour. [n. d.]. The KIT Motion-Language Dataset. Big Data ([n. d.]).

[34] Yurui Qian, Qi Cai, Yingwei Pan, Yehao Li, Ting Yao, Qibin Sun, and Tao Mei. 2024. Boosting Diffusion Models with Moving Average Sampling in Frequency Domain. arXiv:2403.17870 [cs.CV] https://arxiv.org/abs/2403.17870

[35] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In International conference on machine learning. PmLR, 8748–8763.

[36] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2023. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. arXiv:1910.10683 [cs.LG] https://arxiv.org/abs/1910.10683

[37] Jiaming Song, Chenlin Meng, and Stefano Ermon. 2022. Denoising Diffusion Implicit Models. arXiv:2010.02502 [cs.LG] https://arxiv.org/abs/2010.02502

[38] Guy Tevet, Brian Gordon, Amir Hertz, Amit H. Bermano, and Daniel Cohen-Or. 2022. MotionCLIP: Exposing Human Motion Generation to CLIP Space. arXiv:2203.08063 [cs.CV] https://arxiv.org/abs/2203.08063

[39] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H. Bermano. 2022. Human Motion Diffusion Model. arXiv:2209.14916 [cs.CV] https://arxiv.org/abs/2209.14916

[40] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H Bermano. 2022. Human motion diffusion model. arXiv preprint arXiv:2209.14916 (2022).

[41] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, Chenliang Li, Yuanhong Xu, Hehong Chen, Junfeng Tian, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. 2024. mPLUG-Owl: Modularization Empowers Large Language Models with Multimodality. arXiv:2304.14178 [cs.CL] https://arxiv.org/abs/2304.14178

[42] Ye Yuan, Jiaming Song, Umar Iqbal, Arash Vahdat, and Jan Kautz. 2023. PhysDiff: Physics-Guided Human Motion Diffusion Model. arXiv:2212.02500 [cs.CV] https://arxiv.org/abs/2212.02500

[43] Ye Yuan, Jiaming Song, Umar Iqbal, Arash Vahdat, and Jan Kautz. 2023. Physdiff: Physics-guided human motion diffusion model. In Proceedings of the IEEE/CVF international conference on computer vision. 16010–16021.

[44] Beichen Zhang, Pan Zhang, Xiaoyi Dong, Yuhang Zang, and Jiaqi Wang. 2024. Long-CLIP: Unlocking the Long-Text Capability of CLIP. arXiv:2403.15378 [cs.CV] https://arxiv.org/abs/2403.15378

[45] Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Shaoli Huang, Yong Zhang, Hongwei Zhao, Hongtao Lu, and Xi Shen. 2023. T2M-GPT: Generating Human Motion from Textual Descriptions with Discrete Representations. arXiv:2301.06052 [cs.CV] https://arxiv.org/abs/2301.06052

[46] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. 2022. MotionDiffuse: Text-Driven Human Motion Generation with Diffusion Model. arXiv:2208.15001 [cs.CV] https://arxiv.org/abs/2208.15001

[47] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. 2024. Motiondiffuse: Text-driven human motion generation with diffusion model. IEEE transactions on pattern analysis and machine intelligence 46, 6 (2024), 4115–4128.

[48] Mingyuan Zhang, Xinying Guo, Liang Pan, Zhongang Cai, Fangzhou Hong, Huirong Li, Lei Yang, and Ziwei Liu. 2023. Remodiffuse: Retrieval-augmented motion diffusion model. In Proceedings of the IEEE/CVF International Conference on Computer Vision. 364–373.

A Spectral Analysis of the Diffusion Process

We analyze the spectral properties of the signal during the diffusion process, focusing on how noise affects different frequency components over time.

A.1 Power Spectral Density Evolution

Consider the forward diffusion process described by the stochastic differential equation (SDE), simplified by assuming zero drift ($\mathbf{f}(\mathbf{m}_t, t) = 0$):

$$\mathrm{d}\mathbf{m}_t = g(t)\mathrm{d}\mathbf{w}_t, \tag{15}$$

where $g(t)$ is the diffusion coefficient and $\mathbf{w}_t$ is a standard Wiener process. The solution integrates to:

$$\mathbf{m}_t = \mathbf{m}_0 + \boldsymbol{\epsilon}_t, \quad \text{where} \quad \boldsymbol{\epsilon}_t = \int_0^t g(s)\mathrm{d}\mathbf{w}_s. \tag{16}$$

Here, $\mathbf{m}_0$ is the initial clean signal and $\boldsymbol{\epsilon}_t$ represents the accumulated noise up to time t.

Theorem A.1 (Power Spectral Density in Simplified Diffusion). *For the process defined by Eq.* (15)*, the power spectral density (PSD) $S_{\mathbf{m}_t}(\omega)$ of the signal $\mathbf{m}_t$ at time t is given by:*

$$S_{\mathbf{m}_t}(\omega) = |\hat{\mathbf{m}}_0(\omega)|^2 + \int_0^t g^2(s)\mathrm{d}s, \tag{17}$$

where $\hat{\mathbf{m}}_0(\omega)$ is the Fourier transform of the initial signal $\mathbf{m}_0$.

Proof Sketch. Applying the Fourier transform $\mathcal{F}$ to Eq. (16), we get:

$$\hat{\mathbf{m}}_t(\omega) = \mathcal{F}[\mathbf{m}_t](\omega) = \hat{\mathbf{m}}_0(\omega) + \hat{\boldsymbol{\epsilon}}_t(\omega). \tag{18}$$

The PSD is defined as $S_{\mathbf{m}_t}(\omega) = \mathbb{E}[|\hat{\mathbf{m}}_t(\omega)|^2]$. Substituting the above:

$$\begin{aligned} S_{\mathbf{m}_t}(\omega) &= \mathbb{E}\left[|\hat{\mathbf{m}}_0(\omega) + \hat{\boldsymbol{\epsilon}}_t(\omega)|^2\right] \\ &= \mathbb{E}\left[|\hat{\mathbf{m}}_0(\omega)|^2 + 2\,\mathrm{Re}\big(\hat{\mathbf{m}}_0(\omega)\hat{\boldsymbol{\epsilon}}_t^*(\omega)\big) + |\hat{\boldsymbol{\epsilon}}_t(\omega)|^2\right]. \end{aligned} \tag{19}$$

Assuming the initial signal $\mathbf{m}_0$ is deterministic or independent of the subsequent noise $\boldsymbol{\epsilon}_t$, and noting that $\mathbb{E}[\hat{\boldsymbol{\epsilon}}_t(\omega)] = 0$, the cross-term vanishes: $\mathbb{E}[\hat{\mathbf{m}}_0(\omega)\hat{\boldsymbol{\epsilon}}_t^*(\omega)] = \hat{\mathbf{m}}_0(\omega)\mathbb{E}[\hat{\boldsymbol{\epsilon}}_t^*(\omega)] = 0$. The expected energy of the noise in the frequency domain is a standard result from the properties of Itô integrals (related to the autocorrelation of the Wiener process):

$$\mathbb{E}[|\hat{\boldsymbol{\epsilon}}_t(\omega)|^2] = \int_0^t g^2(s)\mathrm{d}s. \tag{20}$$

This noise energy term is independent of frequency ω, characteristic of white noise accumulation in the frequency domain. Combining these results yields Eq. (17). □

A.2 Frequency-Dependent Dynamics in Motion Generation

The result from Theorem A.1 helps elucidate the spectral dynamics during diffusion-based motion generation. Natural motion signals $\mathbf{m}_0$ typically exhibit a low-pass characteristic, meaning their power spectral density (PSD) decays with frequency: $|\hat{\mathbf{m}}_0(\omega)|^2 \propto |\omega|^{-\alpha}$ for some $\alpha > 0$.

Signal-to-Noise Ratio (SNR). During the forward diffusion process, the signal-to-noise ratio (SNR) at frequency ω and time t is defined as:

$$\mathrm{SNR}(\omega, t) = \frac{|\hat{\mathbf{m}}_0(\omega)|^2}{\int_0^t g^2(s)\,\mathrm{d}s}, \tag{21}$$

where $g(t)$ denotes the noise schedule. For a given SNR threshold $\gamma > 0$, we define $t_\gamma(\omega)$ as the earliest time at which the SNR at frequency ω drops to γ:

$$\mathrm{SNR}(\omega, t_\gamma(\omega)) = \gamma \quad \Longrightarrow \quad \int_0^{t_\gamma(\omega)} g^2(s)\,\mathrm{d}s = \frac{|\hat{\mathbf{m}}_0(\omega)|^2}{\gamma}. \tag{22}$$

C. Assume the noise schedule is constant, i.e., $g(t) = \sigma$ for some $\sigma > 0$. Then $g^2(s) = \sigma^2$ and the integral simplifies to:

$$\int_0^{t_\gamma(\omega)} \sigma^2\,\mathrm{d}s = \sigma^2 t_\gamma(\omega). \tag{23}$$

Substituting into Equation (22) gives:

$$\sigma^2 t_\gamma(\omega) = \frac{|\hat{\mathbf{m}}_0(\omega)|^2}{\gamma} \quad \Longrightarrow \quad t_\gamma(\omega) = \frac{|\hat{\mathbf{m}}_0(\omega)|^2}{\gamma\sigma^2}. \tag{24}$$

Using the assumption $|\hat{\mathbf{m}}_0(\omega)|^2 = K|\omega|^{-\alpha}$ for some constant $K > 0$, we obtain:

$$t_\gamma(\omega) = \frac{K}{\sigma^2\gamma}|\omega|^{-\alpha}. \tag{25}$$

Interpretation. Equation (25) reveals several key properties: As the frequency ω increases, the corresponding time $t_\gamma(\omega)$ at which the signal-to-noise ratio (SNR) reaches the threshold γ decreases. This implies that high-frequency components (ω_H) hit the SNR threshold earlier and thus become corrupted by noise sooner during the forward diffusion process. In contrast, low-frequency components (ω_L) maintain a higher SNR for a longer period, allowing their structural information to be preserved deeper into the diffusion process.

Forward Process (Corruption). As time t increases, the noise energy $\int_0^t g^2(s)\,\mathrm{d}s$ accumulates, decreasing the SNR across all frequencies. Since high-frequency components exhibit lower spectral energy $|\hat{\mathbf{m}}_0(\omega)|^2$, they reach the SNR threshold γ sooner, implying that fine-grained motion details are lost earlier during forward diffusion.

Reverse Process (Generation/Denoising). The reverse process begins at a high-noise state (large t), where low SNR conditions prevail. During denoising, the model first reconstructs low-frequency components, which retain relatively higher SNR and thus guide the recovery of the coarse semantic structure. As time decreases, the effective noise level drops and the SNR improves across all frequencies, allowing the model to progressively refine the motion with higher-frequency details.

This analysis substantiates prior empirical findings [18] that diffusion models tend to generate coarse, low-frequency structure early in the reverse process, followed by high-frequency refinements as the process evolves.

B More Visualization Comparison

Figure 7 shows the visualization results of ANT and other SOTA models. As can be seen from various examples, ANT outperforms other models in both naturalness and semantic alignment.

C DETAILS OF HUMAN EVALUATION

We utilize the Google Form platform to allow 20 individuals to separately fill out 100 different motion sequences pairs for testing, where we have designed two types of questions. The first type involves directly rating the semantic accuracy of generated motion. Motion is presented in GIF format, accompanied by two evaluation options: yes or no. The second type of question pertains to user preferences between our model and a baseline model. This question aims to obtain a comparison of our method and the original method from the user's perspective regarding motion generation accuracy. Our questionnaire takes the form C:

min \ max	2.5	3.0	3.5	4.0	4.5
0.5	0.5536/0.1034	0.5552/0.0979	0.5511/0.0975	0.5490/0.0954	0.5553/0.0924
1.0	0.5502/0.0945	0.5546/0.0994	0.5578/0.0967	0.5562/0.0960	0.5509/0.1003
1.5	0.5544/0.0974	**0.5623/0.0975**	0.5550/0.0993	0.5529/0.0996	0.5600/0.0926
2.0	0.5528/0.1033	0.5494/0.0954	0.5575/0.0979	0.5535/0.0931	0.5498/0.0936
2.5	0.5566/0.0975	0.5613/0.0992	0.5517/0.0984	0.5513/0.0943	0.5549/0.0962

Table 6: Performance Metrics (Top1/FID) organized in a grid with hyperparameter ω_{min} value as rows and hyperparameter ω_{max} values as columns (best Top1 highlighted)

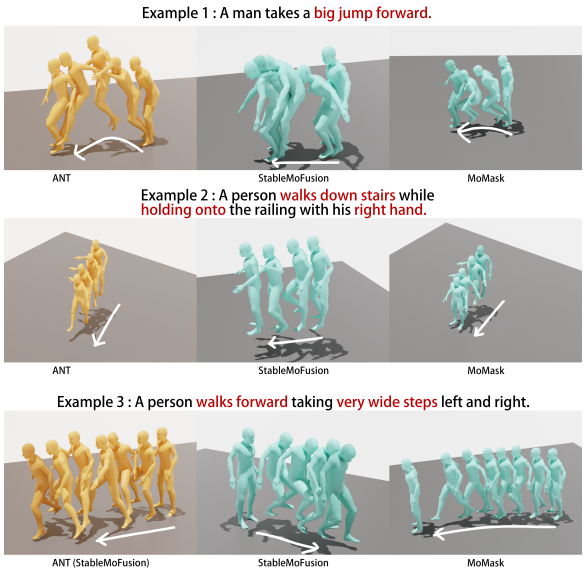

Figure 7: Our ANT can be seamlessly plugged into diffusion-based text-to-motion models to generate semantically rich, fine-grained, and naturally smooth motions with high precision.

[Question1]: Is <motion1.gif> semantically accurate?
(1) Yes
(2) No

[Question2]: Is <motion2.gif> semantically accurate?
(1) Yes
(2) No

[Question2]: Which motion result do you think is generated better?
(1) The first one
(2) The second one

D Discussion of Evaluation Metrics

Previous work has primarily focused on two metrics: FID and R-Precision. FID measures the distance between distributions based on a normality assumption, thereby evaluating the generation quality. However, previous studies [4, 27] have shown that an extremely low FID often contradicts human subjective preferences and is no longer reliable. Considering that the current FID for text-to-motion tasks is already much lower than the average level of text-to-image (T2I) tasks, in this paper, FID is regarded as the second most important reference metric, after R-Precision.

E Pseudo-code

Algorithm 1 shows ANT's Pseudo-code

F Hyperparameter search result

In our study, we conducted a grid search on the validation set to determine the most appropriate values for the hyperparameters ω_{min} and ω_{max}. Specifically, we selected five candidate values for each of ω_{min} and ω_{max}, resulting in a total of 25 distinct parameter combinations. These five groups of hyperparameters were chosen by extending the conventional ranges used in previous studies, allowing us to explore a broader spectrum of possible values. For each combination, we performed tests five times to improve the reliability and statistical significance of our observations.

After completing our grid search methodology, we first analyzed the relationship between the FID and top-k metrics over various hyperparameter configurations. At higher ranges of the parameters, we observed some level of inverse relationship between FID and top-k accuracy. Specifically, settings that yielded lower FID scores tended to correspond with reduced top-k performance, and vice versa. Furthermore, our analysis revealed that, for most of the hyperparameter values (except for the extremely high or low extremes), the overall impact on the results is not significantly sensitive. This insensitivity in the mid-range values suggests that our method demonstrates robustness with respect to these hyperparameter variations. Based on these observations, we carefully selected our hyperparameter to achieve an optimal balance between FID and top-k performance. Our final choice reflects a compromise that leverages the robust region of the parameter space while mitigating the adverse effects observed at the extremes.

Algorithm 1 ANT Diffusion Process for Text-to-Motion Generation

Require:
- Text prompt $\mathcal{T}$
- text encoder (e.g., T5 or CLIP) *Encoder*
- Diffusion denoising network G_θ
- Diffusion timesteps T with noise schedule $\{\alpha_t\}_{t=1}^{T}$
- Guidance parameters: ω_{max}, ω_{min}, decay function $\phi(t)$, threshold time t_{th}

Ensure: Generated motion sequence x_0

Initialization:
$c \leftarrow \text{Encoder}(\mathcal{T})$ {Compute text embedding}
$x_t \sim \mathcal{N}(0, I)$ {Sample initial noise}
for $t = T$ **to** 1 **by** -1 **do**
  Compute Timestep-Specific Features:
  $c_t \leftarrow \text{STA}(c, t)$ {Fuse text embedding e with temporal information}
  Denoising Predictions:
  $\epsilon_{\text{cond}} \leftarrow G_\theta(x_t, t, c_t)$ {Conditional prediction}
  $\epsilon_{\text{uncond}} \leftarrow G_\theta(x_t, t, \emptyset)$ {Unconditional prediction}
  Adaptive Guidance Scale:
  Compute $\omega_t = \omega_{\text{min}} + \phi(t) \cdot (\omega_{\text{max}} - \omega_{\text{min}})$ {e.g., using a cosine schedule}
  Guidance and Update:
  if $t > t_{\text{th}}$ **then**
    $\epsilon_t \leftarrow \epsilon_{\text{uncond}}$ {Late stage: use unconditional branch only}
  else
    $\epsilon_t \leftarrow \epsilon_{\text{uncond}} + \omega_t \cdot (\epsilon_{\text{cond}} - \epsilon_{\text{uncond}})$
  end if
  Update x_{t-1} using the diffusion sampler (e.g., DPM-Solver++):

$$x_{t-1} \leftarrow \text{Update}(x_t, \epsilon_t, t)$$

end for
return x_0

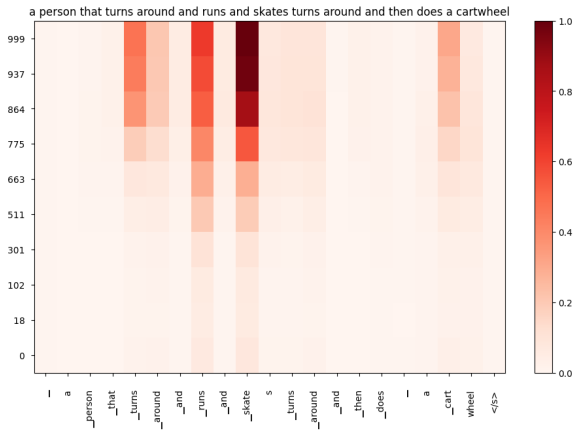

Figure 8: Visualization of attention scores for a sample prompt across diffusion timesteps. The y-axis represents the timestep t from 999 (early, high-noise stage) down to 0 (late, low-noise stage). The x-axis shows the tokens of the input text. Darker shades of red indicate higher attention scores, signifying stronger focus from the model.

G Deeper Analysis of the STA Module's Internal Behavior

To provide a deeper, empirical validation of our proposed Semantic Temporally Adaptive (STA) module, this section delves into its internal working dynamics. We aim to visually substantiate our core hypothesis: the STA module facilitates a two-stage denoising process by dynamically modulating the influence of textual semantics over the diffusion timestep t . We present two complementary pieces of evidence: (1) the temporal dynamics of semantic attention within the cross-attention layers, and (2) the evolution of the AdaLayerNorm modulation parameters that control the strength of semantic injection.

G.1 Temporal Attention Dynamics

To understand what semantic information the model prioritizes at different stages, we visualize the cross-attention scores between the motion features and the input text tokens across the denoising timesteps. As illustrated in Figure 8, we analyze the attention patterns for the prompt: "a person that turns around and runs and skates turns around and then does a cartwheel."

Our analysis reveals a distinct two-stage behavior:

- **Early Stage (Semantic Planning):** During the initial phase of denoising (e.g., for $t > 700$), the model's attention is strongly concentrated on key tokens that define the motion's high-level structure and core actions. For instance, the words `runs`, `skates`, and `cartwheel` receive substantial attention. This observation aligns perfectly with our concept of a "low-frequency semantic planning" phase, where the model establishes the foundational blueprint of the motion sequence based on global textual cues.
- **Late Stage (Detail Refinement):** As the denoising process progresses into its later stages (e.g., $t < 500$), the attention paid to these high-level semantic tokens rapidly decays, with the heatmap becoming significantly lighter. This indicates a functional shift. Once the core motion structure is established, the model reduces its reliance on explicit semantic guidance and transitions to the "high-frequency detail refinement" phase, where it focuses on generating natural, smooth, and coherent transitions between the established keyframes.

G.2 Modulation Strength via AdaLayerNorm Parameters

To provide a more mechanistic understanding of how the STA module controls semantic influence, we analyze the behavior of its core components. The AdaLayerNorm layer modulates motion features using a time-dependent scale parameter (α) and shift parameter (β), which are derived from the textual condition. The magnitudes of these parameters directly govern the strength of semantic conditioning.

Figure 9 illustrates the evolution of the mean α and β values across all denoising timesteps.

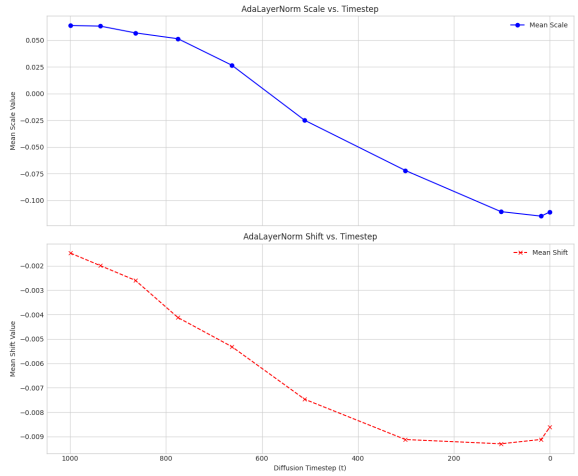

Figure 9: Evolution of the mean `scale` (α, top) and `shift` (β, bottom) parameters of the AdaLayerNorm layer within the STA module. The parameters are plotted against the diffusion timestep t. The decreasing trend in their values indicates a systematic weakening of semantic modulation strength.

The plots show a clear and systematic decrease in the values of both α and β as the timestep t decreases from 1000 to 0. This provides direct quantitative evidence that the STA module programmatically and gracefully weakens the influence of textual semantics as the denoising process unfolds. This behavior is the explicit mechanism behind the "temporal-semantic reweighting" central to our work. It ensures that semantic guidance is strongest when needed for structural planning in the early stages and is attenuated during the later stages to allow for fine-grained, naturalistic motion refinement.

G.3 Synthesis

In summary, these two analyses offer a multi-faceted and cohesive view of the STA module's internal dynamics. The attention heatmap (Figure 8) reveals *what* the model focuses on, demonstrating a shift from high-level concepts to implicit details. Concurrently, the AdaLayerNorm parameter plots (Figure 9) reveal *how strongly* this focus is applied, showing a programmed decay in semantic modulation strength. Together, they provide compelling empirical validation that our STA module successfully implements the intended adaptive, two-stage semantic injection, which is a key contributor to the superior performance and semantic alignment achieved by the ANT model.

Method	Venue	FID ↓	R-Precision ↑			Diversity →	Multimodality ↑
			Top1	Top2	Top3		
Real		$0.006^{\pm0.003}$	$0.335^{\pm0.004}$	$0.513^{\pm0.005}$	$0.628^{\pm0.002}$	$10.098^{\pm0.102}$	-
MLD [5]	CVPR 2023	$0.628^{\pm0.038}$	$0.293^{\pm0.004}$	$0.459^{\pm0.003}$	$0.568^{\pm0.004}$	$9.741^{\pm0.093}$	$3.035^{\pm0.138}$
T2M [10]	CVPR 2022	$1.898^{\pm0.059}$	$0.252^{\pm0.006}$	$0.406^{\pm0.005}$	$0.508^{\pm0.006}$	$8.975^{\pm0.113}$	$4.470^{\pm0.112}$
MoMask [8]	CVPR 2024	$0.383^{\pm0.018}$	$0.301^{\pm0.005}$	$0.481^{\pm0.004}$	$0.597^{\pm0.005}$	$9.689^{\pm0.092}$	$1.968^{\pm0.049}$
T2M-GPT [45]	CVPR 2023	$0.177^{\pm0.016}$	$0.353^{\pm0.005}$	$0.545^{\pm0.006}$	$0.663^{\pm0.005}$	$10.128^{\pm0.132}$	$1.798^{\pm0.041}$
MotionGPT [22]	CVPR 2023	$0.267^{\pm0.017}$	$0.306^{\pm0.004}$	$0.486^{\pm0.006}$	$0.605^{\pm0.006}$	$9.357^{\pm0.133}$	$2.210^{\pm0.137}$
MMM [32]	CVPR 2024	$0.151^{\pm0.013}$	$0.353^{\pm0.004}$	$0.545^{\pm0.004}$	$0.667^{\pm0.005}$	$10.091^{\pm0.086}$	$0.757^{\pm0.042}$
MDM [40]	ICLR 2023	$9.467^{\pm0.217}$	$0.049^{\pm0.003}$	$0.098^{\pm0.005}$	$0.148^{\pm0.005}$	$7.608^{\pm0.100}$	$5.682^{\pm0.203}$
StableMoFusion [20]	ACM MM 2024	$0.460^{\pm0.003}$	$0.312^{\pm0.004}$	$0.494^{\pm0.004}$	$0.607^{\pm0.005}$	$9.546^{\pm0.079}$	$2.157^{\pm0.044}$
ANT (StableMoFusion)	-	$0.149^{\pm0.011}$	$0.347^{\pm0.005}$	$0.541^{\pm0.005}$	$0.666^{\pm0.006}$	$10.034^{\pm0.065}$	$2.094^{\pm0.059}$

Table 7: Evaluation metrics for CMP dataset. ± indicates a 95% confidence interval. Red and Blue indicate the best and the second best result. The right arrow → means the closer to real motion the better. Red and Blue indicate the best and the second best result.

H More Comparison Experiments

To further validate the robustness and generalizability of our proposed ANT model, we provide additional comparison experiments on the Combat Motion Processed (CMP) dataset. This dataset features a distinct motion style compared to the more general HumanML3D and KIT benchmarks, serving as a challenging test case for text-to-motion generation.

CMP Dataset. The Combat Motion Processed (CMP) dataset [**?**] is a benchmark with a combat motion style that includes 8,700 motions and 26,100 text descriptions. It serves as a smaller-scale but more challenging evaluation benchmark for text-to-motion models. For our experiments on this dataset, we employ the same setup used for the main benchmarks: the pose representation follows StableMoFusion, motions are augmented through mirroring, and the data is split into training, validation, and testing sets with a ratio of 0.8 : 0.15 : 0.05.

Results on CMP. Table 7 presents the quantitative results on the CMP dataset. As shown, our ANT model significantly outperforms the baseline (StableMoFusion) across all key metrics, including FID and R-Precision. This demonstrates the effectiveness of our adaptive temporal-aware architecture in a specialized and challenging domain. Furthermore, ANT achieves competitive or superior performance when compared to other state-of-the-art methods, highlighting the strong generalizability of our proposed approach beyond common human actions. These results on the CMP dataset further validate that ANT provides a robust and effective enhancement for diffusion-based text-to-motion models.

I Evaluation Metrics

- **Frechet Inception Distance (FID):** We can evaluate the overall motion quality by measuring the distributional difference between the high-level features of the motions.
- **R-Precision:** We rank Euclidean distances between a given motion sequence and 32 text descriptions (1 ground-truth and 31 randomly selected mismatched descriptions). We report Top-1, Top-2, and Top-3 accuracy of motion-to-text retrieval.
- **Diversity:** From a set of motions, we randomly sample 300 pairs and compute the average Euclidean distances between them to measure motion diversity.
- **Multimodality:** For one text description, we generate 20 motion sequences forming 10 pairs of motion. We then extract motion features and compute the average Euclidean distances of the pairs. We finally report the average over all the text descriptions.
- **Multimodal Distance (MM-Dist)**. The average Euclidean distances between each text feature and the generated motion feature from this text.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009