Computers, Materials & Continua

Tech Science Press

https://doi.org/10.32604/cmc.2026.080992

ARTICLE

Global-Local Embedding Gating Network for Part-Wise Text-to-Motion Generation

Chanyoung Kim, Jion Kim and Byeong-Seok Shin*

Department of Electrical and Computer Engineering, Inha University, Incheon, Republic of Korea

*Corresponding Author: Byeong-Seok Shin. Email: bsshin@inha.ac.kr

Received: 20 February 2026; Accepted: 20 April 2026; Published: 15 June 2026

ABSTRACT: Diffusion-based methods have substantially improved the performance of full-body Text-to-Motion (T2M) generation from natural language descriptions. Despite this progress, accurately capturing the fine-grained semantics of composite prompts remains challenging. Approaches that rely solely on a single global text condition often fail to retain part-specific semantic cues, leading to deviations in the motions of certain body parts from the intended descriptions. Recent methods have attempted to address this by incorporating both global and local conditions, yet these are typically combined using fixed ratios or applied in separate stages, which restricts their adaptability to evolving semantic requirements during generation. To address these constraints, this work proposes the Embedding Gating Network (EGN), which dynamically modulates the contributions of global and local information according to the current noisy motion state and the diffusion timestep. By conditioning the gating mechanism on the intermediate noisy motion estimate, EGN adjusts the relative importance of global and local information to emphasize semantics that remain underrepresented at each denoising step. The conditioned signals are processed through independent part-wise generation pathways to minimize semantic interference, while a lightweight fusion module enables inter-part information exchange to preserve structural coherence across the full body. Experiments on the HumanML3D benchmark show that the proposed method consistently improves text-motion alignment over existing full-body and part-based baselines, without compromising motion quality or diversity. Analysis of the learned gating coefficients reveals that local conditions primarily contribute to the formation of part-wise structural outlines during early denoising stages, whereas global conditions become increasingly influential, integrating cross-part semantics and refining full-body consistency as denoising advances. These findings indicate that dynamically modulating conditioning signals during generation is an effective alternative to fixed-ratio conditioning.

KEYWORDS: Motion generation; diffusion model; human motion synthesis; text-to-motion; condition embedding

1 Introduction

Text-to-Motion (T2M) synthesis aims to generate coherent and realistic 3D human motion sequences from natural language descriptions, serving as a foundational technology across virtual reality, animation, robotics, and gaming [1–3]. The primary challenge arises from the intricate spatiotemporal structures of human motion and the inherent flexibility of natural language, both of which hinder precise semantic alignment [4–6].

Recent advances in T2M research have been driven by diverse generative paradigms, including diffusion-based approaches [7,8], autoregressive token-based approaches [9,10], and mask-based token modeling approaches [11,12]. These approaches have improved full-body motion quality, temporal coherence, and

text-motion alignment, as well as the ability to process complex natural language descriptions. Nevertheless, higher generation quality does not guarantee fine-grained controllability. Most current methods condition on text using a single global embedding [7,8,13], which often fails to preserve part-level semantics present in composite sentences. Consequently, although the generated motion may appear natural overall, the movements of specific body parts can diverge from the intended text. Recent approaches have sought to address this limitation through part-wise generation [3,14] or by employing finer-grained condition separation strategies [15,16]. However, these methods typically combine global and part-level conditions with fixed contributions or treat part-wise generation and global integration as separate stages. Such designs implicitly assume that the relative importance of global context and part-level information remains constant throughout the generation process. In practice, when generating composite actions, the balance between maintaining overall motion structure and capturing part-specific semantics varies over time. Therefore, a conditioning mechanism capable of flexibly adjusting the contributions of global context and part-level semantics throughout the generation process is required.

To address this, we propose a diffusion-based T2M framework incorporating an Embedding Gating Network (EGN), which dynamically adjusts the relative contributions of global and part-level embeddings based on the current generation state. In contrast to existing methods that combine global and local conditions with fixed weights or process them in separate stages, EGN ensures that the global context continuously informs part-wise generation throughout the entire process. The gating mechanism is conditioned on both static text embeddings and the current noisy motion state, enabling dynamic adjustment of the global-local balance based on semantics that remain underrepresented in the current motion estimate. We further observe that part-level semantics contribute not only to late-stage refinement but also to early structure formation. Based on this, the framework incorporates local embeddings from the earliest generation stages. The modulated embeddings are routed through dedicated part-wise generation pathways, which limit inter-part interference inherent in shared pathways and allow each generator to specialize in its assigned part-level semantics. Simultaneously, the global context is maintained through EGN, supporting both part-wise specialization and full-body coherence.

Our main contributions are as follows:

- We introduce the EGN, which dynamically adjusts the contributions of global and part-level embeddings to reflect the current generation state, achieving precise part-level semantic alignment while preserving global motion consistency.
- We demonstrate that local embeddings, when explicitly incorporated from the early generation stages, contribute to initial structure formation beyond late-stage refinement.
- We design dedicated part-wise generation pathways with a fusion architecture that enables each generator to focus on part-specific semantics without inter-part interference, while maintaining structurally coherent full-body motion through controlled inter-part information exchange.

2 Related Work

2.1 Text-to-Motion Generation

T2M generation aims to synthesize 3D human motion sequences from natural language descriptions. Early studies relied on conditioning with simple action classes or text labels [17,18]. As the field expanded to include free-form natural language instructions and required precise semantic alignment between text and motion, controllability became a key challenge. Initial approaches employed variational autoencoders (VAEs) [19] to align text and motion in a shared embedding space [5,6] or to learn text–motion correspondence via conditional decoding [4]. Tevet et al. [7] used visual priors by encoding motion descriptions

with a frozen Contrastive Language-Image Pre-training (CLIP) [20] text encoder and aligning the motion latent space accordingly. More recently, diffusion models [21] have been widely adopted in T2M due to their strong generation quality and temporal stability. The Motion Diffusion Model (MDM) [7] introduced a Transformer-based [22] denoising network that directly models temporal dependencies. Chen et al. [13] improved efficiency and quality by leveraging latent-space diffusion with optimized sampling. However, these methods often apply the same text condition across all body parts, which makes it difficult to capture accurate part-level semantic mappings and leads to mismatches in part motion [14,15]. In parallel, token-based approaches tokenize continuous motion into codebook indices using vector quantization (VQ) [23] and treat it as a discrete sequence modeling problem. Guo et al. [9] formulated text-to-motion as a discrete sequence generation problem, and Zhang et al. [10] improved token prediction via autoregressive generation. More recently, Guo et al. [11] combined hierarchical quantization with bidirectional masked prediction, mitigating the error accumulation inherent in sequential autoregressive generation. However, tokenization can result in information loss when compressing continuous motion into a limited codebook, potentially restricting motion diversity and fidelity [24]. Although generation architectures have evolved in diverse directions, from a conditioning perspective, most methods still share the limitation of relying on fixed-length global embeddings, such as the CLIP [CLS] token. This reliance dilutes fine-grained semantics for specific body parts in composite prompts, motivating the need for conditioning mechanisms that faithfully reflect part-level semantics while maintaining expressiveness.

Ghosh et al. [3] began with coarse divisions such as upper and lower body, treating each part through a separate generation pathway. Athanasiou et al. [16] automated spatial composition by using GPT-3 to extract action-body part mappings and combining independently generated part-wise motions post-hoc. Subsequent research has focused on increasing the number of segmented parts to achieve finer motion representation. Part-Coordinating (ParCo) [14] decomposes the full body into multiple semantic parts and performs part-wise generation while considering inter-part coordination. More recent work has combined part-based modeling with text-based motion generation, directly connecting detailed elements to part-level representations. The Local-to-Global pipeline for Text-to-Motion generation (LGTM) [15] decomposes global text into part-specific descriptions to strengthen semantic alignment at the part level. Wang et al. [25] extract body-part-related cues through large language model (LLM)-based semantic parsing and reflect sentence structure to condition detailed semantics more precisely. Fan et al. [26] identify salient body parts and enforce semantic alignment for interaction motion. Improvements have also been attempted from the perspective of the conditioning mechanism itself. Chang et al. [27] introduce a composite-aware text encoder and a text-motion aligner, enabling dynamic word-level correspondence rather than fixed-length global embeddings. Li and Feng [28] improve the generation accuracy of composite actions by simultaneously leveraging coarse- and fine-grained descriptions. These studies supply fine-grained cues that would otherwise be lost in a global summary embedding, enabling body-part actions specified in text to be more faithfully reflected in the generated motion. However, existing methods combine global and local conditions at fixed ratios or perform partial generation and global optimization as separate stages, which can cause one condition to overshadow the other. Additionally, post-hoc composition approaches do not incorporate inter-part coordination into the generation process itself, and learning-based part-wise generation methods that inject part conditions independently through dedicated pathways struggle to capture full-body context. A mechanism that dynamically adjusts the relative contributions of global and local conditions as generation progresses has yet to be explored.

2.2 Conditioning Mechanism in Generative Models

The challenge of fine-grained semantics being diluted under global conditioning has been widely recognized in conditional generative modeling. In text-to-image synthesis, efforts to address this issue have included decomposing composite prompts into per-concept diffusion processes [29] and leveraging linguistic structure to restructure cross-attention [30]. Zarei et al. [31] demonstrated that the output space of CLIP is suboptimal for compositional prompts, showing that attention contributions from unrelated tokens are mixed into the final token embeddings, leading to failures in attribute-object binding. In the facial synthesis domain, Song et al. [32] improved attribute-level alignment under multi-attribute conditions by introducing a module that dynamically balances global text features and local attribute features through learnable gating. It has also been observed that the role of conditioning changes qualitatively across stages of the diffusion process. Balaji et al. [33] empirically showed that text-to-image diffusion models rely heavily on text conditioning during early denoising but largely ignore it in later stages. More recently, Cho et al. [34] argued that static conditioning cannot flexibly adapt to the dynamic nature of multi-stage denoising, which evolves from coarse structure to fine detail, and proposed TC-LoRA to dynamically adjust conditioning based on both the denoising timestep and the control signal. These studies consistently demonstrate that hierarchical semantic decomposition and dynamic conditioning are critical to generation quality, yet these insights remain largely unexplored in T2M. In this paper, we combine such adaptive conditioning strategies with part-level semantic decomposition for T2M and show that timestep-conditioned gating improves motion-text alignment.

3 Methodology

3.1 Part-Wise Motion Representation

Recent text-to-motion research, particularly following HumanML3D [4], typically represents full-body motion using a canonical pose that incorporates root information and joint features relative to the root. A motion sequence consists of F frames with J joints, where each frame $\mathbf{x}^i = [\dot{r}_a, \dot{\mathbf{r}}_{xz}, r_h, \mathbf{j}_p, \mathbf{j}_r, \mathbf{j}_v, \mathbf{c}_f]$. In this formulation, $\dot{r}_a \in \mathbb{R}^1$ denotes the root yaw angular velocity, $\dot{\mathbf{r}}_{xz} \in \mathbb{R}^2$ denotes the root linear velocity on the XZ plane, and $r_h \in \mathbb{R}^1$ denotes the root height. The features $\mathbf{j}_p \in \mathbb{R}^{3\times(J-1)}$, $\mathbf{j}_r \in \mathbb{R}^{6\times(J-1)}$, and $\mathbf{j}_v \in \mathbb{R}^{3\times J}$ represent joint position, rotation, and velocity in the root coordinate system, respectively, while $\mathbf{c}_f \in \mathbb{R}^4$ is a binary feature indicating foot-ground contact. This representation encodes both global trajectory and relative joint motion, and is widely adopted for stable modeling of diverse motions. Building on this canonical representation, our approach reorganizes motion into semantic parts, reflecting the fact that text descriptions frequently target specific body parts. We partition joints into six groups (root, backbone, left arm, right arm, left leg, right leg), and construct part-wise motion by selecting only the joint features (position, rotation, velocity, and foot contact) for each group.

3.2 Embedding Gating Network

We propose the EGN, which consists of a transformation module and a *GlocalGate*, as illustrated in Fig. 1a. In diffusion-based T2M models, text is typically condensed into a single global embedding, which can obscure fine-grained cues in complex sentences. Motivated by this, we separate text conditions into global and local components. The global text $\mathbf{T}_g$ represents the original full-body description, providing overall action identity and coarse context. Part-specific local texts $\{\mathbf{T}_{l,p}\}_{p=1}^{P}$ are derived by decomposing $\mathbf{T}_g$ into short action phrases for each of the six body parts using an LLM. For instance, given the global description *"a person bends forward and picks up an object in their right hand,"* the LLM assigns *"bends forward"* to the pelvis and *"picks up an object"* to the right arm, while the remaining parts receive a null descriptor indicating no part-specific action. Each text input is independently encoded by the pretrained

CLIP text encoder, extracting the end-of-sequence token representation as a fixed-dimensional embedding. This process yields a global embedding $\mathbf{e}_g^{\text{CLIP}} \in \mathbb{R}^D$ and part-specific local embeddings $\{\mathbf{e}_{l,p}^{\text{CLIP}} \in \mathbb{R}^D\}_{p=1}^{P}$. Since the text encoder's embedding space may not align with the conditioning space required for motion generation, a transformation module is introduced to expand the representational capacity of both global and local embeddings. This module is implemented as a residual feed-forward adapter that learns motion-relevant correction terms on top of the original CLIP embeddings. We formulate this transformation as a residual function, allowing the embeddings to be progressively adapted for motion conditioning while retaining the original CLIP embeddings as a semantic anchor. The final conditioning embeddings are obtained as

$$\mathbf{e}_g = \mathbf{e}_g^{\text{CLIP}} + f_g(\mathbf{e}_g^{\text{CLIP}}), \quad \mathbf{e}_l = \mathbf{e}_l^{\text{CLIP}} + f_l(\mathbf{e}_l^{\text{CLIP}}) \tag{1}$$

where f_g and f_l denote the learnable nonlinear mappings for global and local conditions, respectively.

Diffusion models are known to recover global structures at high noise levels and fine-grained details at low noise levels [35,36]. We therefore hypothesize that the relative importance of global and local conditions shifts across timesteps and generation states. Based on this hypothesis, we dynamically modulate their contributions rather than combining them at a fixed ratio. Accordingly, GlocalGate predicts global and local weights from the noisy motion $\mathbf{x}_t$ and timestep t. This module is defined as:

$$[\mathbf{w}_g, \mathbf{w}_l] = \text{Linear}(\text{SiLU}(\text{Linear}([\mathbf{x}_t, \text{emb}(t), t]))), \tag{2}$$

where $[\cdot]$ denotes concatenation and $\text{emb}(t)$ represents the timestep embedding. The output logits $\mathbf{w}_g$ and $\mathbf{w}_l$ are subsequently normalized via softmax to yield the final gating coefficients:

$$\boldsymbol{\alpha}_g = \frac{\exp(\mathbf{w}_g)}{\exp(\mathbf{w}_g) + \exp(\mathbf{w}_l)}, \quad \boldsymbol{\alpha}_l = \frac{\exp(\mathbf{w}_l)}{\exp(\mathbf{w}_g) + \exp(\mathbf{w}_l)}. \tag{3}$$

These gating coefficients, $\boldsymbol{\alpha}_g(t)$ and $\boldsymbol{\alpha}_l(t)$, determine the relative contribution of global and local semantics at each timestep. Using these coefficients, we compute the part-specific condition $\mathbf{c}_p(t)$ for part p as the weighted sum of global and local embeddings:

$$\mathbf{c}_p(t) = \boldsymbol{\alpha}_g(t) \cdot \mathbf{e}_g + \boldsymbol{\alpha}_{l,p}(t) \cdot \mathbf{e}_{l,p}. \tag{4}$$

GlocalGate thus adaptively balances global and local influences based on its inputs. Specifically, the noisy motion $\mathbf{x}_t$ conveys the current geometric state and pose configuration, while the timestep t encodes the prevailing noise level. The learned coefficients $\boldsymbol{\alpha}_g$ and $\boldsymbol{\alpha}_l$ therefore shift the balance between global coherence and local specificity at each generation step. The EGN integrates both the transformation module and GlocalGate to produce the final part-specific conditioning embedding $\mathbf{c}_p(t)$.

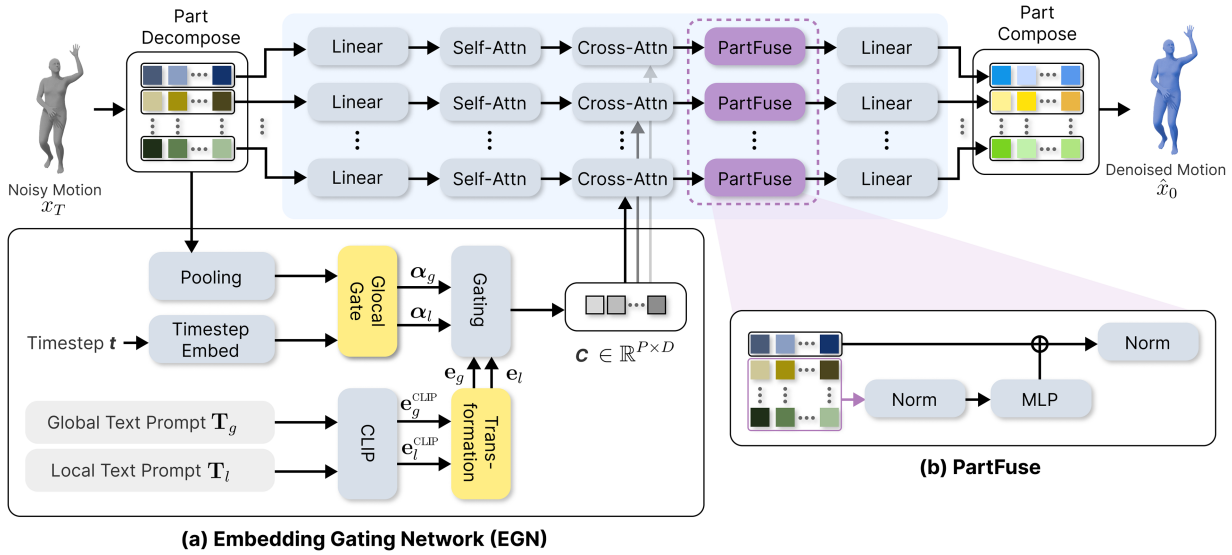


Figure 1: The overall architecture of the proposed framework. (**a**) The final embeddings computed by the EGN are applied to part-wise generation modules. (**b**) The generated part-wise motions are fused within PartFuse blocks to maintain overall motion consistency.

3.3 Conditional Generation with Part-Wise Pathways

Our approach maintains part-wise pathways within a single model: each part representation is generated under its global-local embedding, after which a lightweight fusion step exchanges inter-part information, as illustrated in Fig. 1b. Accordingly, we separate the network into part-wise attention for conditional generation and PartFuse$(\cdot)$ for part-wise fusion. At each timestep t, the generation process is divided into two specialized stages: semantic synthesis and structural coordination. Initially, part-wise attention utilizes the part-specific condition $\mathbf{c}_p(t)$ to update the latent representation of each part p as follows:

$$\hat{h}_t^p = \text{Attention}(h_t^p, \mathbf{c}_p(t)), \tag{5}$$

where $\hat{h}_t^p$ represents the semantically updated state focusing on synthesizing the local action structure. Subsequently, PartFuse$(\cdot)$ enables controlled inter-part coordination while maintaining the separation of generation pathways. It is defined for each part i as follows:

$$\text{PartFuse}\left(\hat{h}_t^i; [\hat{h}_t^1 \mid \cdots \mid \hat{h}_t^j]_{j\neq i}\right) = \text{LN}\left(\hat{h}_t^i + \text{MLP}_i \circ \text{LN}\left([\hat{h}_t^1 \mid \cdots \mid \hat{h}_t^j]_{j\neq i}\right)\right), \tag{6}$$

where MLP_i functions as a part-specific alignment module that incorporates features from other parts into the current pathway.

A key advantage is the explicit separation, at the layer level, between part-wise attention and PartFuse$(\cdot)$. Decoupling semantic interpretation from inter-part coordination introduces an inductive bias that keeps part representations distinct. This architectural choice stabilizes training and provides consistent part-level control, even for complex textual conditions involving multiple parts.

4 Experiments

4.1 Experiment Settings

Dataset. We evaluate our method on HumanML3D [4], a widely used public benchmark for text-to-motion generation. HumanML3D consists of 3D human motion sequences paired with natural language annotations and has served as a standard dataset in prior T2M studies. It contains 14,616 motion sequences and 44,970 text descriptions. Since global text annotations in HumanML3D are often insufficient for part-level semantic alignment, we augment the dataset with local texts derived from the global annotations using the decomposition procedure described in Section 3.2.

Metrics. We evaluate motion quality and text–motion alignment using five metrics: (1) R-Precision measures motion–text retrieval accuracy in the feature space of a pretrained T2M evaluation network. For each motion sequence, we compute precision based on whether the ground-truth text is ranked within Top-1/2/3 among 32 candidate texts. (2) Fréchet Inception Distance (FID) measures the distributional gap between generated and real motions by computing FID on motion features extracted by the T2M evaluator. (3) Multi-Modal Distance (MM-Dist) computes the average Euclidean distance between each text feature and the motion features generated from that text. (4) Diversity splits generated motions into two random subsets of equal size and computes the average Euclidean distance between their motion features [4,5,37]. (5) Part-level Multi-Modal Similarity (PMM Sim), adopted from Sun et al. [15], trains part-level text and motion encoders with contrastive learning following Petrovich et al. [38] and measures the correspondence between part-specific text descriptions and generated part motions.

Baselines. We select three baselines to evaluate our part-wise diffusion generation framework, which jointly uses global and local text conditions: (1) MDM, a diffusion-based T2M model conditioned on global text, as a reference for full-body diffusion performance; (2) ParCo, a part-wise generation approach with separate body-part pathways, to assess the impact of part-wise generation on performance and consistency; and (3) LGTM, which decomposes global text into part-wise descriptions, to evaluate the effect of part-wise text conditioning.

Implementation Details. We employ Qwen2.5-7B-Instruct [39] as the LLM for decomposition, prompting it to produce a local text for each of the six body parts given the global annotation. Our model consists of 155M total parameters, of which 21.2M are trainable. Training was conducted on a single NVIDIA RTX 4090 GPU for 83.3 h on the HumanML3D dataset. At inference, the model requires 0.45 s per sample (DDIM [40], 50 steps) and peaks at 1184 MB of GPU memory.

4.2 Qualitative Results

Fig. 2 compares results on composite prompts in which a single sentence specifies a full-body action alongside multiple part-level instructions. MDM largely preserves natural full-body motion, but tends to converge to a similar global motion pattern, yielding limited changes when part instructions vary and occasionally omitting part-specific constraints. ParCo can reflect some parts' instructions due to part-wise generation, but under composite prompts, inter-part interactions often become unstable, leading to motion collapse or suppression where constraints on one part inhibit the motions of other parts. LGTM improves responsiveness to part instructions by providing part-wise text conditions, but when part conditions conflict or when the fusion with global context is not sufficiently stable, some constraints are weakened, or the overall action consistency degrades. In contrast, our method injects part-wise signals progressively while preserving the global condition throughout denoising, producing motions that preserve full-body coherence while simultaneously satisfying multiple part instructions under composite prompts.

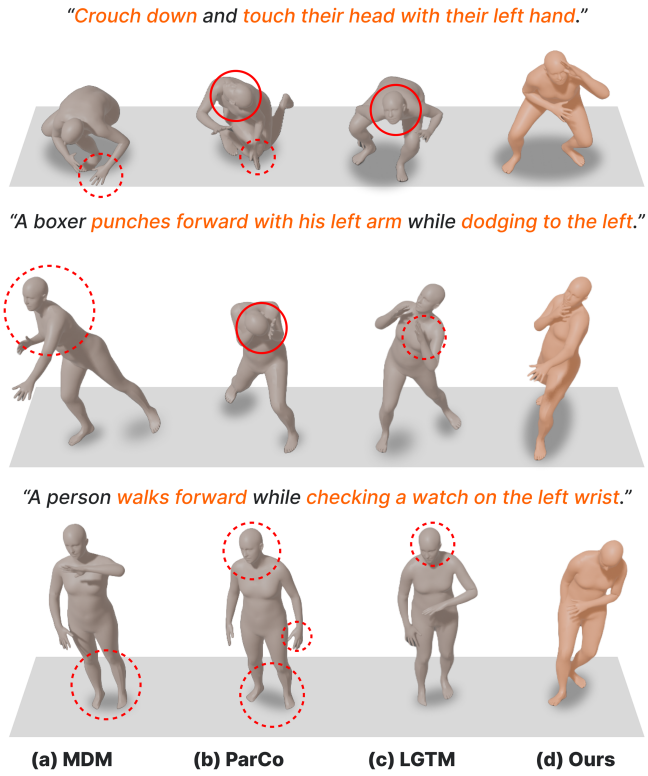


Figure 2: Qualitative comparison results with baseline models. Red dashed lines indicate missing instructions, and red solid lines indicate motion collapse artifacts. (**a**) MDM, (**b**) ParCo, and (**c**) LGTM exhibit missing or incorrectly reflected instructions along with body distortion artifacts, whereas (**d**) the proposed method reflects the given instructions while generating natural body motion.

Fig. 3 evaluates how well part-level text descriptions, decomposed from input sentences using an LLM, are preserved during full-body motion synthesis. For example, given the prompt *"A person sits down and stretches his legs straight."*, LGTM generates part-wise motions from decomposed part texts and then applies strong global conditioning during the full-body motion optimization stage [15]. In this process, features induced by part conditions (e.g., stretches legs straight) are partially suppressed by the global context (e.g., sits down), leading to cases where the legs converge to a bent posture rather than being fully extended. In contrast, our method maintains the global condition as an anchor while explicitly injecting and aligning part-wise conditions during generation, ensuring full-body consistency while preserving part-level semantics. As a result, local semantics such as *"stretches legs straight"* are more consistently reflected in the final motion, even in composite sentences.

Overall, our method reflects part-wise conditions more consistently when satisfying multiple part constraints simultaneously in composite prompts. Even in sentences with a strong global context, local prompts decomposed at the part level are stably preserved in the final motion, ensuring that key actions of specific parts appear without being weakened. This observation qualitatively supports the effectiveness of the global-local conditioning scheme, which stabilizes the integration of part-wise conditions throughout the diffusion process while preserving the global condition.

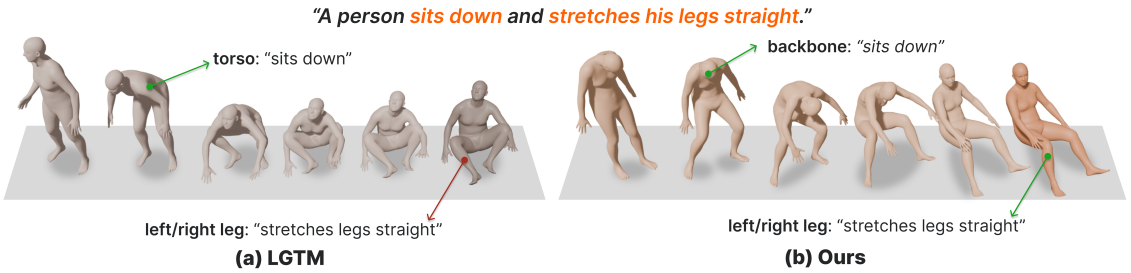


Figure 3: Generation results according to different embedding methods for part-wise instructions. Green lines indicate parts where instructions are well-reflected, and red lines indicate parts where instructions are missing. (**a**) LGTM fails to reflect the "stretches legs straight" instruction in both legs, whereas (**b**) the proposed method correctly captures both "sits down" and "stretches legs straight" across the corresponding body parts.

4.3 Quantitative Results

Table 1 presents the quantitative evaluation results of the proposed method and baselines (MDM, ParCo, LGTM) on the HumanML3D test set. From the perspective of text-motion alignment, our method demonstrates consistent improvements across all metrics. Our method achieves substantial gains in R-Precision and MM-Dist compared with MDM and outperforms ParCo and LGTM in text-motion alignment. This quantitatively confirms that part-wise conditioning improves alignment over global-only conditioning, which tends to obscure fine-grained elements in complex sentences. In terms of generation quality (FID), our method shows slightly higher FID than ParCo, yet R-Precision and MM-Dist consistently show improvements. Meng et al. [24] have shown that the standard evaluation protocol can disproportionately favor VQ-based methods over diffusion-based methods. Accordingly, the FID gap between the proposed diffusion-based method and the VQ-based ParCo likely reflects intrinsic differences between the two generation paradigms. Among diffusion-based methods, our method achieves the best FID, suggesting that dynamically modulating global and local conditions throughout generation is effective for improving generation quality. Finally, for Diversity, our model maintains a level of diversity comparable to that of real motions, suggesting that it improves text-motion alignment without sacrificing variation in the generated results.

Table 1: Quantitative comparison with baseline models on HumanML3D dataset. ↑ indicates higher is better, ↓ indicates lower is better, and → indicates closer to real motion is better. Bold indicates the best performance, and underline indicates the second-best performance.

Method	R-Precision ↑			FID ↓	MM-Dist ↓	Div. →
	Top-1	Top-2	Top-3			
Real	0.4616	0.6726	0.7746		3.2408	9.3808
MDM	$0.4078^{\pm 0.007}$	$0.6065^{\pm 0.008}$	$0.7194^{\pm 0.007}$	$0.8482^{\pm 0.081}$	$3.5088^{\pm 0.024}$	$9.1045^{\pm 0.099}$
ParCo	$\underline{0.4514}^{\pm 0.003}$	$\underline{0.6552}^{\pm 0.003}$	$0.7635^{\pm 0.003}$	$\mathbf{0.1427}^{\pm 0.006}$	$\underline{3.2785}^{\pm 0.010}$	$\mathbf{9.4201}^{\pm 0.068}$
LGTM	$0.4486^{\pm 0.003}$	$0.6551^{\pm 0.003}$	$\underline{0.7676}^{\pm 0.003}$	$0.8015^{\pm 0.003}$	$3.3048^{\pm 0.003}$	$8.7743^{\pm 0.003}$
Ours	$\mathbf{0.4751}^{\pm 0.003}$	$\mathbf{0.6823}^{\pm 0.003}$	$\mathbf{0.7874}^{\pm 0.003}$	$\underline{0.1567}^{\pm 0.005}$	$\mathbf{3.1396}^{\pm 0.009}$	$\underline{9.5130}^{\pm 0.072}$

Table 2 reports the evaluation of part-level semantic alignment using PMM Sim. Our method achieves the highest scores on Left Arm and Right Arm, and remains competitive on Head, ranking second only to ParCo by a narrow margin. This is because arm movements are explicitly described in text annotations, yielding semantically rich local embeddings via LLM-based decomposition, which enables EGN's gating

and dedicated generation pathways to operate effectively. In contrast, performance on Torso, Left Leg, and Right Leg is slightly below LGTM. These parts are less frequently described in text, often receiving null descriptors, which makes the local embeddings semantically sparse and causes EGN's gating to shift toward the global condition, reducing the benefit of part-specific conditioning. Nevertheless, our method consistently surpasses MDM and ParCo on these parts, maintaining balanced alignment performance across all body parts.

Table 2: Quantitative comparison of part-level semantic alignment with baseline models on the HumanML3D dataset, evaluated using PMM Sim. Higher values indicate better performance. Bold indicates the best performance, and underline indicates the second-best performance.

Method	Head	Left Arm	Right Arm	Torso	Left Leg	Right Leg
Real	0.8037	0.7160	0.7210	0.7583	0.7531	0.7578
MDM	0.7867	0.7000	0.6925	0.7425	0.7347	0.7231
ParCo	**0.8081**	<u>0.7206</u>	<u>0.7293</u>	0.7070	0.6380	0.6243
LGTM	0.7967	0.7198	0.7256	**0.7656**	**0.7575**	**0.7631**
Ours	<u>0.8069</u>	**0.7248**	**0.7349**	<u>0.7603</u>	<u>0.7516</u>	<u>0.7592</u>

Overall, by preserving the global condition while adaptively injecting part-level conditions during generation, our method improves text-motion semantic alignment over existing baselines and remains stable in generation quality and diversity.

4.4 Ablation Study

This section analyzes the impact of the key components of the proposed framework on performance. We first examine the contributions of individual components within EGN, then investigate how framework-level design choices—including global descriptions, local descriptions, the PartFuse module, and part-wise generation pathways—affect overall performance.

Table 3 reports the results of separately removing the two core components of EGN: the Transformation module and the Noisy Motion conditioning. *"w/o Transformation"* removes the residual feed-forward adapter f_g and f_l (Eq. (1)), directly using the raw CLIP embeddings as conditioning inputs. *"w/o Noisy Motion"* removes $\mathbf{x}_t$ from the GlocalGate input (Eq. (2)), so that the gating coefficients are determined solely by timestep information. Removing the Transformation module degrades R-Precision and MM-Dist, while FID rises sharply from 0.1567 to 0.4152. Without a learned transformation, the raw embeddings are insufficient for effective conditioning, destabilizing the generation distribution. Removing the Noisy Motion conditioning also results in a decline in R-Precision and MM-Dist, along with a modest increase in FID. This indicates that reflecting the current noisy motion state in the gating contributes to adaptive condition adjustment as generation progresses. Collectively, these results confirm that both the embedding transformation and the motion-state-based gating within EGN contribute to text-motion alignment and generation quality, and that neither component alone is sufficient.

Table 4 analyzes the contributions of four framework-level components—global descriptions, local descriptions, the PartFuse module, and part-wise generation pathways—by evaluating their combinations. Removing PartFuse results in a slight decline in R-Precision and MM-Dist, while FID increases notably from 0.1567 to 0.2973. This suggests that simply combining independently generated part motions without a dedicated fusion mechanism degrades the quality of the full-body distribution, confirming that PartFuse

helps ensure inter-part coherence. When the global description is removed and the generation is based solely on local descriptions, a sharp performance drop is observed across all metrics. This can be attributed to two factors. First, text annotations in the current dataset do not explicitly describe every body part, so relying on local descriptions alone leaves certain parts without conditioning information. Second, without the global context itself, there is no basis for capturing inter-part relationships and the overall semantic structure of the motion, making it difficult for part-wise generators to form coherent full-body motion. When local descriptions are removed, each part-wise generation pathway receives only the same global summary as its condition. Although the pathways are physically separated, the absence of differentiated semantic information prevents part-wise separation from being fully exploited, leading to degradation in both text-motion alignment and generation quality. When part-wise generation pathways are removed and generation relies solely on the global description, text-motion alignment metrics (MM-Dist, R-Precision) decline, and FID also increases. This indicates that relying solely on a global description through a single generation pathway limits the model's ability to capture fine-grained semantics, and that explicitly separating part-wise generation pathways is effective for improving both alignment and generation quality.

Table 3: Ablation results for EGN component variants on the HumanML3D dataset, including the removal of the Transformation module and Noisy Motion. ↑ indicates higher is better, ↓ indicates lower is better, and → indicates that values closer to the real-motion Diversity score (9.3808) are better. Bold indicates the best performance.

Method	R-Precision ↑			FID ↓	MM-Dist ↓	Div. →
	Top-1	Top-2	Top-3			
Full EGN	**0.4751**$^{\pm 0.003}$	**0.6823**$^{\pm 0.003}$	**0.7874**$^{\pm 0.003}$	**0.1567**$^{\pm 0.005}$	**3.1396**$^{\pm 0.009}$	9.5130$^{\pm 0.105}$
w/o Transformation	0.4670$^{\pm 0.002}$	0.6665$^{\pm 0.002}$	0.7735$^{\pm 0.002}$	0.4152$^{\pm 0.008}$	3.2141$^{\pm 0.009}$	**9.3378**$^{\pm 0.064}$
w/o Noisy Motion	0.4655$^{\pm 0.003}$	0.6700$^{\pm 0.003}$	0.7800$^{\pm 0.003}$	0.1835$^{\pm 0.008}$	3.1893$^{\pm 0.009}$	9.4955$^{\pm 0.085}$

Table 4: Ablation study of framework-level conditioning and architectural components on the HumanML3D dataset. ✓ and ✗ indicate the inclusion and exclusion of each component, respectively. ↑ indicates higher is better, ↓ indicates lower is better, and → indicates that values closer to the real-motion Diversity score (9.3808) are better. Bold indicates the best performance.

Method				R-Precision ↑			FID ↓	MM-Dist ↓	Div. →
Global Description	Local Description	Part Fuse	Part-wise Pathways	Top-1	Top-2	Top-3			
✓	✓	✓	✓	**0.4751**$^{\pm 0.003}$	**0.6823**$^{\pm 0.003}$	**0.7874**$^{\pm 0.003}$	**0.1567**$^{\pm 0.005}$	**3.1396**$^{\pm 0.009}$	9.5130$^{\pm 0.105}$
✓	✓	✗	✓	0.4664$^{\pm 0.003}$	0.6717$^{\pm 0.003}$	0.7808$^{\pm 0.003}$	0.2973$^{\pm 0.009}$	3.1718$^{\pm 0.006}$	**9.3709**$^{\pm 0.107}$
✗	✓	✓	✓	0.1762$^{\pm 0.003}$	0.2632$^{\pm 0.005}$	0.3234$^{\pm 0.005}$	10.632$^{\pm 0.099}$	6.6868$^{\pm 0.017}$	7.6367$^{\pm 0.083}$
✓	✗	✓	✓	0.4581$^{\pm 0.006}$	0.6525$^{\pm 0.006}$	0.7640$^{\pm 0.006}$	0.6030$^{\pm 0.068}$	3.2864$^{\pm 0.021}$	9.3459$^{\pm 0.068}$
✓	✗	✗	✗	0.4639$^{\pm 0.003}$	0.6670$^{\pm 0.002}$	0.7750$^{\pm 0.002}$	0.2274$^{\pm 0.009}$	3.2098$^{\pm 0.009}$	9.3021$^{\pm 0.070}$

In summary, the global context serves as the structural foundation for overall motion, while local descriptions and part-wise pathways function complementarily to achieve fine-grained semantic alignment. PartFuse integrates independently generated part motions at the full-body level. When these components are combined, both text-motion alignment and generation quality improve jointly.

4.5 Analysis of Timestep-Dependent Gating

To quantify the relative contribution of local conditioning with respect to global conditioning at diffusion timestep t, we define $\alpha(t)$ as the ratio of the local gating coefficient to the global gating coefficient,

i.e., $\boldsymbol{\alpha}(t) = \boldsymbol{\alpha}_l(t)/\boldsymbol{\alpha}_g(t)$. Figs. 4 and 5 visualize the changes in $\boldsymbol{\alpha}(t)$ across diffusion timesteps. $\boldsymbol{\alpha}(t)$ is high in the early generation stages and gradually decreases as denoising progresses, indicating that local conditions contribute strongly to part-level structure formation at the beginning and then shift toward refining full-body context under global conditions. This suggests that local information is not merely used as a late-stage detail refinement signal; rather, it serves as a structural constraint in the early steps to shape the outline and skeleton of part-specific actions. As sampling proceeds, the relative contribution of global conditions increases, strengthening overall motion context and full-body consistency. $\boldsymbol{\alpha}(t)$ is not uniform across body parts: parts that are essential for establishing the global action tend to exhibit larger $\boldsymbol{\alpha}(t)$. This discrepancy is particularly pronounced in the early diffusion steps, implying that the gating network prioritizes injecting local signals into globally critical parts to stabilize the formation of the initial motion structure. As timesteps progress, local contributions decrease across all parts, while the early-stage prioritization for globally important parts remains consistent.

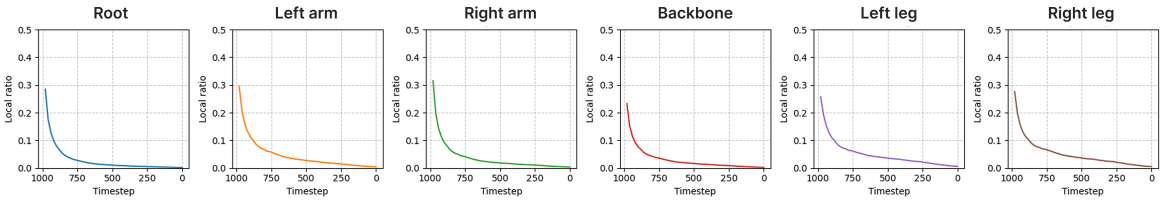
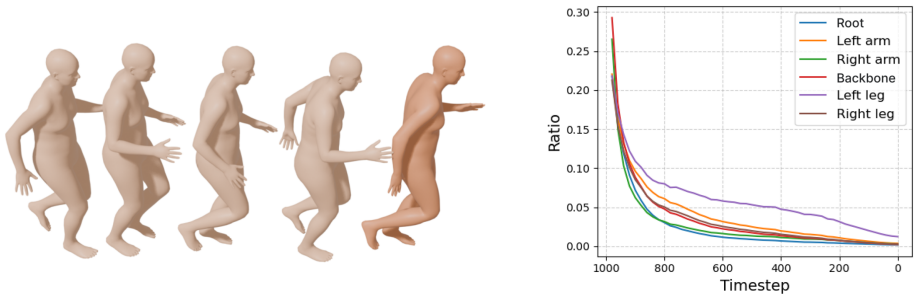
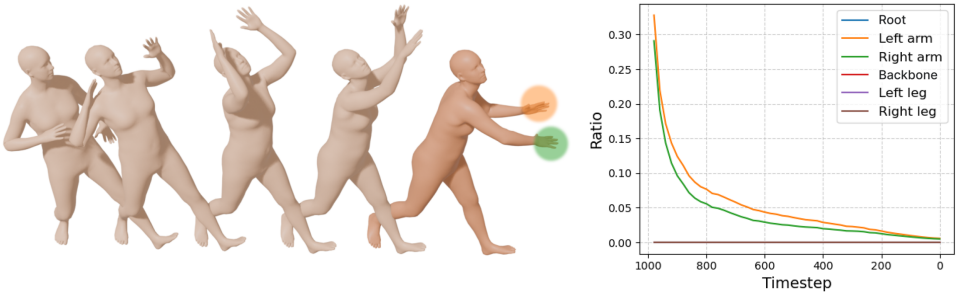


Figure 4: The part-wise variation of the ratio of local gating coefficient to global gating coefficient across timesteps. Across all body parts, the local ratio is high in the early stages of generation and gradually decreases as denoising progresses, indicating that local conditions contribute primarily to initial structure formation and their influence diminishes in later stages.



(a) *"person begins to move their arms and legs in a motion that indicates jogging forward."*



(b) *"a person uses both hands to throw an object."*

Figure 5: The part-wise variation of the ratio of local gating coefficient to global gating coefficient across timesteps according to different prompts. (**a**) When the instructions for all parts are clear, the ratios show similar patterns across them. (**b**) When instructions for specific parts are critical, the ratios of those parts appear relatively higher.

Table 5 provides quantitative evidence supporting this interpretation. Removing local injection ($\boldsymbol{\alpha}(t) = 0$) degrades part-level semantic alignment, confirming that explicit local conditioning is necessary for part-level semantic alignment under composite prompts. Mirroring $\boldsymbol{\alpha}(t)$ over timesteps—so that early local contributions shrink and late ones grow—causes an overall performance drop, demonstrating that sufficiently strong local signals in the early stage are crucial. Furthermore, using a fixed $\boldsymbol{\alpha}$ also underperforms the learned schedule, suggesting that a static global–local mixture is insufficient and that timestep-dependent role transition is necessary.

Table 5: Performance metrics according to local gating coefficients across timesteps. ↑ indicates higher is better, ↓ indicates lower is better, and → indicates that values closer to the real-motion Diversity score (9.3808) are better. Bold indicates the best performance.

Method	R-Precision ↑			FID ↓	MM-Dist ↓	Div. →
	Top-1	Top-2	Top-3			
$\boldsymbol{\alpha}(t) = 0$	$0.4575^{\pm 0.002}$	$0.6631^{\pm 0.003}$	$0.7718^{\pm 0.003}$	$0.1738^{\pm 0.005}$	$3.2190^{\pm 0.009}$	$\mathbf{9.3201}^{\pm 0.072}$
$\boldsymbol{\alpha}(t) = 0.5$	$0.4295^{\pm 0.003}$	$0.6418^{\pm 0.003}$	$0.7470^{\pm 0.003}$	$0.6598^{\pm 0.005}$	$3.4003^{\pm 0.009}$	$9.1112^{\pm 0.071}$
Mirrored	$0.3616^{\pm 0.002}$	$0.5355^{\pm 0.003}$	$0.6395^{\pm 0.003}$	$2.0028^{\pm 0.036}$	$4.2265^{\pm 0.012}$	$8.1536^{\pm 0.071}$
Learnable (Ours)	$\mathbf{0.4751}^{\pm 0.003}$	$\mathbf{0.6823}^{\pm 0.003}$	$\mathbf{0.7874}^{\pm 0.003}$	$\mathbf{0.1567}^{\pm 0.005}$	$\mathbf{3.1396}^{\pm 0.009}$	$9.5130^{\pm 0.105}$

In summary, the proposed EGN functions as a mechanism that uses local signals as early structural constraints to form part-wise outlines and progressively refines full-body consistency through global context.

5 Conclusion

This study proposes an EGN that dynamically modulates the contributions of global and local conditions to enhance part-level semantic alignment in composite prompts. Additionally, we design a pathway-separated generation structure composed of part-wise attention and PartFuse modules, enabling each part to maintain its separated pathway while ensuring full-body consistency through inter-part coordination. Quantitative and qualitative evaluations on HumanML3D showed overall improvements in text-motion alignment and generation quality compared to existing full-body and part-based methods. Furthermore, modulating global-local fusion based on timestep and motion state produced more consistent alignment than either fixed-ratio or schedule-based alternatives. These results suggest that condition modulation reflecting generation progress can effectively improve fine-grained motion consistency and controllability in diffusion-based T2M.

Acknowledgement: Not applicable.

Funding Statement: This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2025-02214780, Generative Haptics and Fine Response Inference for Flexible Tactile Interfaces) and the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2026-25479030).

Author Contributions: The authors confirm contribution to the paper as follows: Conceptualization, Chanyoung Kim; methodology, Chanyoung Kim; software, Chanyoung Kim; validation, Chanyoung Kim; formal analysis, Chanyoung Kim; investigation, Chanyoung Kim; resources, Chanyoung Kim; data curation, Chanyoung Kim; writing—original draft preparation, Chanyoung Kim; writing—review and editing, Jion Kim, Byeong-Seok Shin; visualization, Chanyoung Kim; supervision, Byeong-Seok Shin; project administration, Byeong-Seok Shin; funding acquisition, Byeong-Seok Shin. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The HumanML3D dataset used in this study is publicly available at https://github.com/EricGuo5513/HumanML3D.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Plappert M, Mandery C, Asfour T. The KIT motion-language dataset. Big Data. 2016;4(4):236–52. doi:10.1089/big.2016.0028.
2. Ahn H, Ha T, Choi Y, Yoo H, Oh S. Text2Action: generative adversarial synthesis from language to action. In: Proceedings of the 2018 IEEE International Conference on Robotics and Automation (ICRA); 2018 May 21–25; Brisbane, QLD, Australia. p. 1–5.
3. Ghosh A, Cheema N, Oguz C, Theobalt C, Slusallek P. Synthesis of compositional animations from textual descriptions. In: Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV); 2021 Oct 10–17; Montreal, QC, Canada. p. 1396–406.
4. Guo C, Zou S, Zuo X, Wang S, Ji W, Li X, et al. Generating diverse and natural 3D human motions from text. In: Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2022 Jun 18–24; New Orleans, LA, USA. p. 5152–61.
5. Petrovich M, Black MJ, Varol G. TEMOS: generating diverse human motions from textual descriptions. In: Comput vision–ECCV 2022. Vol. 13682. Cham, Switzerland: Springer; 2022. p. 480–97.
6. Ahuja C, Morency LP. Language2Pose: natural language grounded pose forecasting. In: Proceedings of the 2019 International Conference on 3D Vision (3DV); 2019 Sep 16–19; Québec City, QC, Canada. p. 719–28.
7. Tevet G, Raab S, Gordon B, Shafir Y, Cohen-Or D, Bermano AH. Human motion diffusion model. arXiv:2209.14916. 2022. doi:10.48550/arXiv.2209.14916.
8. Zhang M, Cai Z, Pan L, Hong F, Guo X, Yang L. MotionDiffuse: text-driven human motion generation with diffusion model. IEEE Trans Pattern Anal Mach Intell. 2024;46(6):4115–28. doi:10.1109/TPAMI.2024.3355414.
9. Guo C, Zuo X, Wang S, Cheng L. TM2T: stochastic and tokenized modeling for the reciprocal generation of 3D human motions and texts. In: Comput vision–ECCV 2022. Vol. 13695. Cham, Switzerland: Springer; 2022. p. 580–97.
10. Zhang J, Zhang Y, Cun X, Zhang Y, Zhao H, Lu H, et al. Generating human motion from textual descriptions with discrete representations (T2M-GPT). In: Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2023 Jun 17–24; Vancouver, BC, Canada. p. 14730–40.
11. Guo C, Mu Y, Javed MG, Wang S, Cheng L. MoMask: generative masked modeling of 3D human motions. In: Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2024 Jun 16–22; Seattle, WA, USA. p. 1900–10.
12. Pinyoanuntapong E, Wang P, Lee M, Chen C. MMM: generative masked motion model. In: Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2024 Jun 16–22; Seattle, WA, USA. p. 1546–55.
13. Chen X, Jiang B, Liu W, Huang Z, Fu B, Chen T, et al. Executing your commands via motion diffusion in latent space. In: Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2023 Jun 17–24; Vancouver, BC, Canada. p. 18000–10.
14. Zou Q, Yuan S, Du S, Wang Y, Liu C, Xu Y, et al. ParCo: part-coordinating text-to-motion synthesis. In: Comput vision-ECCV 2024. Cham, Switzerland: Springer Nature Switzerland; 2025. p. 126–43.
15. Sun H, Zheng R, Huang H, Ma C, Huang H, Hu R. LGTM: local-to-global text-driven human motion diffusion model. In: SIGGRAPH '24: Special Interest Group on Computer Graphics and Interactive Techniques Conference; 2024 Jul 27–Aug 1; Denver, CO, USA. p. 1–9. doi:10.1145/3641519.3657422.

16. Athanasiou N, Petrovich M, Black MJ, Varol G. SINC: spatial composition of 3D human motions for simultaneous action generation. In: Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV); 2023 Oct 1–6; Paris, France. p. 9984–95.
17. Guo C, Zuo X, Wang S, Zou S, Sun Q, Deng A, et al. Action2Motion: conditioned generation of 3D human motions. In: Proceedings of the 28th ACM International Conference on Multimedia; 2020 Oct 12–16; Seattle, WA, USA. p. 2021–9.
18. Petrovich M, Black MJ, Varol G. Action-conditioned 3D human motion synthesis with Transformer VAE. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV); 2021 Oct 10–17; Montreal, QC, Canada. p. 10985–95.
19. Kingma DP, Welling M. Auto-encoding variational Bayes. arXiv:1312.6114. 2013.
20. Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. Learning transferable visual models from natural language supervision. In: Proceedings of the 38th International Conference on Machine Learning; 2021 Jul 18–24; Virtual. p. 8748–63.
21. Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models. In: Proceedings of the 34th International Conference on Neural Information Processing Systems; 2020 Dec 6–12; Vancouver, BC, Canada. p. 6840–51.
22. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems; 2017 Dec 4–9; Long Beach, CA, USA. p. 6000–10.
23. van den Oord A, Vinyals O, Kavukcuoglu K. Neural discrete representation learning. In: Proceedings of the 31st International Conference on Neural Information Processing Systems; 2017 Dec 4–9; Long Beach, CA, USA. p. 6309–18.
24. Meng Z, Xie Y, Peng X, Han Z, Jiang H. Rethinking diffusion for text-driven human motion generation: redundant representations, evaluation, and masked autoregression. In: Proceedings of the 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2025 Jun 10–17; Nashville, TN, USA. p. 27859–71.
25. Wang Y, Li M, Liu J, Leng Z, Li FWB, Zhang Z, et al. Fg-T2M++: LLMs-augmented fine-grained text driven human motion generation. Int J Comput Vis. 2025;133(7):4277–93. doi:10.1007/s11263-025-02392-9.
26. Fan S, Du B, Cai X, Peng B, Sun L. TextIM: part-aware interactive motion synthesis from text. arXiv:2408.03302. 2024.
27. Chang CJ, Liu QT, Zhou H, Pavlovic V, Kapadia M. CASIM: composite aware semantic injection for text to motion generation. arXiv:2502.02063. 2025. doi:10.48550/arxiv.2502.02063.
28. Li K, Feng Y. Motion generation from fine-grained textual descriptions. In: LREC-COLING 2024-The 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation; 2024. p. 11625–41.
29. Liu N, Li S, Du Y, Torralba A, Tenenbaum JB. Compositional visual generation with composable diffusion models. In: Comput vision-ECCV 2022. Cham, Switzerland: Springer Nature Switzerland; 2022. p. 423–39.
30. Feng W, He X, Fu TJ, Jampani V, Akula A, Narayana P, et al. Training-free structured diffusion guidance for compositional text-to-image synthesis. arXiv:2212.05032. 2022. doi:10.48550/arxiv.2212.05032.
31. Zarei A, Rezaei K, Basu S, Saberi M, Moayeri M, Kattakinda P, et al. Improving compositional attribute binding in text-to-image generative models via enhanced text embeddings. arXiv:2406.07844. 2024. doi:10.48550/arxiv.2406.07844.
32. Song W, Ye Z, Sun M, Hou X, Li S, Hao A. AttriDiffuser: adversarially enhanced diffusion model for text-to-facial attribute image synthesis. Pattern Recognit. 2025;163(8):111447. doi:10.1016/j.patcog.2025.111447.
33. Balaji Y, Nah S, Huang X, Vahdat A, Song J, Zhang Q, et al. eDiff-I: text-to-image diffusion models with an ensemble of expert denoisers. arXiv:2211.01324. 2022. doi:10.48550/arxiv.2211.01324.
34. Cho M, Ohana R, Jacobsen C, Jothi A, Chen MH, Mao ZM, et al. TC-LoRA: temporally modulated conditional LoRA for adaptive diffusion control. arXiv:2510.09561. 2025. doi:10.48550/arXiv.2510.09561.
35. Rissanen S, Heinonen M, Solin A. Generative modelling with inverse heat dissipation. In: Proceedings of the Eleventh International Conference on Learning Representations; 2023 May 1–5; Kigali, Rwanda. p. 1–5.

36. Choi J, Lee J, Shin C, Kim S, Kim H, Yoon S. Perception prioritized training of diffusion models. In: Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2022 Jun 18–24; New Orleans, LA, USA. p. 11472–81.
37. Heusel M, Ramsauer H, Unterthiner T, Nessler B, Hochreiter S. 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; 2017 Dec 4–9; Long Beach, CA, USA. p. 6629–40.
38. Petrovich M, Black MJ, Varol G. TMR: text-to-motion retrieval using contrastive 3D human motion synthesis. In: Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV); 2023 Oct 1–6; Paris, France. p. 9488–97.
39. Yang A, Yang B, Zhang B, Hui B, Zheng B, Yu B, et al. Qwen2.5 technical report. arXiv:2412.15115. 2024. doi:10.48550/arxiv.2412.15115.
40. Song J, Meng C, Ermon S. Denoising diffusion implicit models. arXiv:2010.02502. 2020. doi:10.48550/arXiv.2010.02502.