TIDMG: Text-Image Driven Motion Generation with Unified Attention and Adaptive Conditioning

Author:
ZIXUAN ZHOU

Supervisor:
ZHIDONG XIAO

National Center for Computer Animation
Faculty of Media & Communication
Bournemouth University

A thesis submitted in partial fulfilment of the requirements of
Bournemouth University for the degree of
Master of science in Artificial Intelligence for Media

AUGUST 2025

Abstract

Text-driven 3D human motion generation is broadly useful for animation, robotics, and XR, yet existing diffusion pipelines often rely on denoising alone and struggle with prompt ambiguity and scene consistency; we present TIDMG, a text–image–driven motion generation framework that couples a lightweight LLM for prompt normalization with CLIP encoders for text and optional scene images, and our contributions are: (1) a unified conditioning stack for joint text–image guidance in motion diffusion; (2) an efficient combined-attention block (CAB) with AdaLN-FFN that realizes self- and cross-attention in a single pass; (3) a lightweight prompt-normalization front end; and (4) a scene-augmented evaluation subset for text-and-scene conditioned motion. Specifically, multimodal tokens condition a diffusion denoiser via a single multi-head attention over the concatenated token stream (CAB), followed by a position-wise FFN modulated by adaptive LayerNorm (AdaLN) using the timestep and a pooled text summary; we train with the standard ε-prediction objective and sample by iterative denoising. On HumanML3D and our scene-augmented subset, TIDMG achieves consistent gains in semantic alignment and scene consistency—with lower FID and multimodal distance, higher R-Precision and multimodality, and strong diversity—while the single-pass attention reduces compute compared to dual self/cross stacks; moreover, ablations confirm the benefits of LLM-based prompt normalization, image conditioning, the combined attention versus separate blocks, and AdaLN versus a vanilla FFN.

Keywords: Motion generation, diffusion models, combined attention

Acknowledgements

I am deeply grateful to my supervisor, **Prof. Zhidong Xiao**, for your invaluable guidance and unwavering support throughout this research. My sincere thanks to **Prof. Xiaosong Yang**, **Prof. Hammadi Nait-Charif**, and **Prof. Jon Macey** for your inspiring teaching and mentorship. I also would like to thank the **senior PhD students** who generously shared their expertise and time whenever I needed help. To my wonderful **classmates** - thank you for the collaboration, laughter, and mutual support that made this journey memorable.

Thank you all!

Table of Contents

List of Tables

List of Figures

1 Introduction

Human motion modeling is a critical component for animating virtual characters to imitate vivid and rich human movements, which has been a vital topic for many applications, such as film-making, game development, and virtual reality (Petrovich, Black, and Varol, 2021, Guo, Zou, et al., 2022). To mimic human motions, virtual characters should be capable of moving naturally and expressing sophisticated actions. Despite exciting technological breakthroughs, it often requires sophisticated equipment (e.g., motion capture systems) and domain experts to produce lively and authentic body movements (Mahmood et al., 2019). In order to remove skill prerequisites for layman users and potentially scale to the mass audience, it is vital to create a versatile human motion generation model that could produce diverse, easily manipulable motion sequences.

Various condition signals, including pre-defined motion categories (Petrovich, Black, and Varol, 2021), music pieces (J. Li et al., 2020, R. Li et al., 2021), and natural language (Ahuja and Morency, 2019, Ghosh et al., 2021), have been leveraged in previous human motion generation methods. Among them, natural language is arguably the most user-friendly and convenient input format for motion sequence synthesis, and hence we focus on text-driven motion generation in this work. Recent methods like TEMOS (Petrovich, Black, and Varol, 2022) demonstrate accurate human motion synthesis. However, many existing approaches are limited to short text inputs, single prompts, and often fail to handle complicated motion descriptions or incorporate contextual scene information, which greatly limits users' creativity and practical application.

To tackle the aforementioned challenges, we propose **TIDMG**, a versatile, scene-consistent Text-Image Driven Motion Generation framework. Inspired by the recent progress of the text-conditioned diffusion models (Tevet, Raab, et al., 2022, M. Zhang et al., 2024), we incorporate a Denoising Diffusion Probabilistic Model (DDPM) (Ho, Jain, and Abbeel, 2020) into our motion generation pipeline. Unlike methods that learn a direct, deterministic mapping between the text and motion spaces (Petrovich, Black, and Varol, 2022), we guide the generation process with both text and image conditions via a unified conditioning mechanism, which significantly increases the diversity and contextual relevance of the generated motions.

Furthermore, our model can achieve enhanced controllability by accepting an optional scene image alongside the text description. This allows the generated motion to not only align with the action described in the text but also be spatially plausible within the given environmental context. To efficiently integrate these multimodal inputs, we design a novel **Combined-Attention Block (CAB)** that merges self-attention (Vaswani et al., 2017) and cross-attention into a single, streamlined operation within a transformer-based architecture. This design eliminates the computational overhead of separate attention blocks, as used in frameworks like MotionDiffuse (M. Zhang et al., 2024), while enhancing the interaction between motion, text, and visual modalities. Conditioning is further refined through Adaptive Layer Normalization (AdaLN) (Peebles and Xie, 2023), which modulates the network's activations based on the diffusion timestep and a summary of the text context.

We perform extensive qualitative and quantitative evaluations on the popular HumanML3D benchmark (Guo, Zou, et al., 2022). The results indicate competitive text-driven performance, with consistent improvements on HumanML3D under standard metrics. A qualitative analysis further shows that TIDMG produces more contextually aligned and scene-consistent motions when additional scene

context is provided.

Our main contributions are summarized as follows:

- We propose TIDMG, a unified diffusion framework that extends the inputs from text alone to *text + scene images*; we realize the fusion via combined attention with AdaLN conditioning, enabling semantically aligned and scene-consistent motion generation.
- We design a streamlined FFN together with the proposed combined-attention and AdaLN conditioning, providing a more efficient and effective alternative to the separate attention blocks used in previous diffusion-based motion models.
- Empirically, the model achieves solid text-driven performance and further improves with image conditioning, resulting in high-fidelity, context-aligned motions across quantitative and qualitative assessments.

2 Related Work

2.1 Human Motion Generation

Human motion generation has been studied for decades across graphics, vision, and animation. Early pipelines emphasized model-based editing and statistical interpolation, such as motion graphs and functional analysis, to produce plausible kinematic transitions (Rose, Cohen, and Bodenheimer, 1998) (Mukai and Kuriyama, 2005) (Min and Chai, 2012) (Ormoneit et al., 2005) (Gavrila, 1999) (O'rourke and Badler, 1980). With the advent of deep sequence models, recurrent architectures improved temporal coherence but suffered from exposure bias and error accumulation on long horizons (Hochreiter and Schmidhuber, 1997). To increase realism and diversity, adversarial, flow, and variational formulations were introduced: GANs targeted perceptual plausibility (Goodfellow et al., 2014) (Barsoum, Kender, and Liu, 2018), normalizing flows offered exact likelihood and controllable sampling (Dinh, Krueger, and Y. Bengio, 2014) (Dinh, Sohl-Dickstein, and S. Bengio, 2016) (Durk P Kingma and Dhariwal, 2018) (Henter, Alexanderson, and Beskow, 2020), and VAEs enabled multi-modal dynamics (Diederik P Kingma and Welling, 2013) (Yan et al., 2018). Comprehensive surveys summarize these trajectories and open challenges (Zhu et al., 2023).

More recently, diffusion and transformer paradigms have become dominant due to stable likelihood-based training and scalable attention for long sequences. Diffusion was first formalized for images and audio (Ho, Jain, and Abbeel, 2020) and strengthened by improved training and variational perspectives (A. Q. Nichol and Dhariwal, 2021) (Dhariwal and Alexander Nichol, 2021) (D. Kingma et al., 2021). These ideas were then adapted to human motion, where transformer encoders or CLIP (Contrastive Language–Image Pretraining)-style language backbones condition denoising to align motions with semantics (Radford et al., 2021). (Tevet, Raab, et al., 2022) (Petrovich, Black, and Varol, 2022) (Guo, Zou, et al., 2022) (M. Zhang et al., 2024).

2.2 Diffusion-based Motion Generation Models

Diffusion has rapidly emerged as a leading generative paradigm for motion due to its training stability and strong sample fidelity, extending core advances from image and text-guided synthesis (Ho, Jain, and Abbeel, 2020) (A. Q. Nichol and Dhariwal, 2021) (Dhariwal and Alexander Nichol, 2021) (Alex Nichol et al., 2021) (Ramesh et al., 2022). Compared with pre-diffusion paradigms such as GANs, flows, and VAEs, diffusion typically achieves more reliable optimization and better mode coverage for sequential motion, particularly when language or semantic embeddings condition the denoising trajectory (Tevet, Raab, et al., 2022) (Petrovich, Black, and Varol, 2022) (Guo, Zou, et al., 2022) (M. Zhang et al., 2024). In practice, text-conditioned motion diffusion encodes prompts with transformer-style modules or CLIP-like features and denoises trajectories step-by-step, often integrating efficient attention to scale to longer sequences (Vaswani et al., 2017) (Radford et al., 2021) (Shen et al., 2021).

However, pure diffusion pipelines still face common issues: expensive iterative sampling and latency; weak linguistic disambiguation and limited structural control under open-vocabulary prompts; and long-horizon drift without scene or pose priors—gaps that text-driven or scene-grounded constraints

can explicitly mitigate.

To reduce architectural redundancy while preserving conditional control, recent practice increasingly merges self- and cross-attention into a single combined attention stream and injects conditions through adaptive layer normalization (AdaLN), a pattern popularized in scalable diffusion transformers and compatible with motion denoisers (Peebles and Xie, 2023) (Vaswani et al., 2017) (Shen et al., 2021) (Radford et al., 2021).

2.3 Conditioned Motion Generation

Conditioned motion generation aligns human motion with auxiliary signals such as text, audio, scene, or trajectory cues. Early efforts explored language-grounded pose forecasting and sentence-to-animation synthesis (Ahuja and Morency, 2019) (Ghosh et al., 2021), while music-to-dance methods leveraged adversarial or curriculum-based training to match rhythm and style (Huang et al., 2020) (J. Li et al., 2020) (R. Li et al., 2021) (Siyao et al., 2022). These directions showed that conditioning provides explicit guidance, enabling more semantically meaningful and controllable synthesis than purely stochastic models.

The emergence of motion–language corpora enabled systematic conditioning at scale. HumanML3D provided paired text–motion descriptions for supervised alignment (Plappert, Mandery, and Asfour, 2016) (Guo, Zou, et al., 2022), while BABEL enriched action labels on mocap sequences to cover fine-grained semantics (Punnakkal et al., 2021). Large mocap resources such as AMASS facilitate learning broad kinematic priors for generalization and editing (Mahmood et al., 2019). Together, these datasets catalyzed transformer-based text/action conditioning and improved the evaluation of semantic alignment and kinematic plausibility (Petrovich, Black, and Varol, 2021) (Petrovich, Black, and Varol, 2022) (Guo, Zou, et al., 2022).

Diffusion has recently become the default vehicle for conditioned generation. MotionCLIP projected motions into CLIP space to align with text (Tevet, Gordon, et al., 2022), and MDM adopted diffusion with transformer encoders to condition on natural language (Tevet, Raab, et al., 2022). Building on these, 2024–2025 works broaden control signals and robustness: timeline and trajectory control for precise temporal structure (Petrovich, Litany, et al., 2024) (Wan et al., 2024); discrete or masked diffusion to enhance diversity and mode coverage (Chi et al., 2024) (Guo, Mu, et al., 2024) (Gao et al., 2024); local-to-global decoding for coherent long motions (Sun et al., 2024); scene- and affordance-grounded synthesis for spatial plausibility and interaction (Wang et al., 2024) (Kulkarni et al., 2024) (Cha et al., 2024) (Liang, W. Zhang, et al., 2024) (Mir et al., 2024); intention-guided and editable pipelines that refine or repair generated motions (Diomataris et al., 2024) (Athanasiou et al., 2024); and open-vocabulary or LLM-guided generation that improves prompt coverage and controllability (Liang, Bao, et al., 2024) (Y. Zhang et al., 2024). Efficiency and long-horizon stability are advanced by autoregressive diffusion and state-space style models (Han et al., 2024) (Pinyoanuntapong et al., 2024) (Z. Zhang et al., 2024), while shape- and composition-aware variants improve structural fidelity and reuse (Xue et al., 2025) (Ruiz-Ponce et al., 2025). These trends collectively point to a unified yet flexible conditioned diffusion stack that integrates language, scene, and structure for reliable text-driven motion generation.

3 Methodology

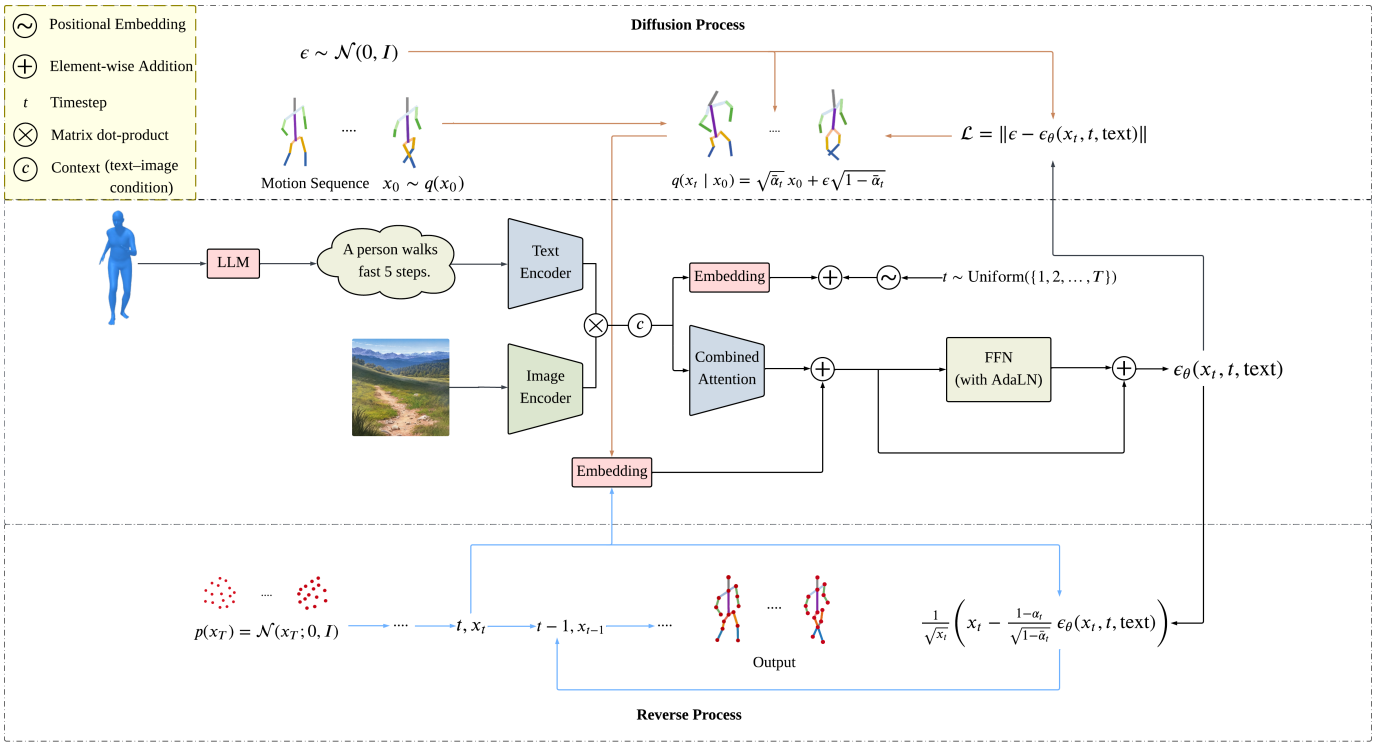

Figure 3.1. LLM converts motion sequences into text, while scene images are encoded as context. This combined information conditions a diffusion model to generate motion that matches both the text and the scene.

3.1 Problem Setup

We propose a *Text-Image-Driven Motion Generation* (TIDMG) framework, which generates a human motion sequence consistent with a natural-language prompt and optionally, a scene image. A motion sequence is defined as:

$$x_0 = \{p_1, p_2, \ldots, p_L\}, \tag{1}$$

where $\{p_\ell\}_{\ell=1}^{L}$ denotes the pose at frame ℓ (e.g., joint coordinates or rotations). Given a prompt y and an optional scene image I, the goal is to produce x_0 that matches the semantics of y and remains consistent with I.

Before encoding, the prompt is normalized by a lightweight instruction-tuned LLM (we use a compact OpenAI GPT-4o-mini), yielding a model-friendly sequence, the text and image features are respectively encoded by the CLIP pre-trained text and image encoders to obtain the corresponding feature representations:

$$T = E_{\text{text}}(y'). \tag{2}$$

$$V = E_{\text{img}}(I). \tag{3}$$

To align modalities to the model width, we further project and position-encode the tokens (where P denotes positional encodings):

$$\tilde{T} = Proj_{\text{text}}(T) + P_{\text{text}} + e^{\text{text}}_{\text{type}}. \tag{4}$$

$$\tilde{V} = Proj_{\text{img}}(V) + P_{\text{img}} + e^{\text{img}}_{\text{type}}. \tag{5}$$

Here $Proj_{\text{text}}$ and $Proj_{\text{img}}$ are linear projections to the model width D; P_{text} and P_{img} are positional encodings for text/image tokens; and $e^{\text{text}}_{\text{type}}$ and $e^{\text{img}}_{\text{type}}$ are learned type embeddings that mark the modality. The resulting sequences satisfy $\tilde{T} \in \mathbb{R}^{N_t \times D}$ and $\tilde{V} \in \mathbb{R}^{N_v \times D}$, where N_t and N_v denote the numbers of text and image tokens, respectively.

For text encoding we adopt a Transformer-style module (Vaswani et al., 2017), and our conditioning protocol follows established text-to-motion practice (M. Zhang et al., 2024).

At diffusion step t, noised motion frames are mapped to motion tokens with positional embeddings; the timestep is embedded via a learned/sinusoidal mapping:

$$M_t = Proj(x_t) + P \in \mathbb{R}^{L \times D}. \tag{6}$$

$$e_t = Emb(t). \tag{7}$$

Here $Emb(\cdot)$ maps the discrete diffusion step $t \in \{1, \ldots, T\}$ to a D-dimensional timestep embedding (sinusoidal/Fourier or learned); $Proj(\cdot)$ is a per-frame linear projection that maps motion features to the model width D; P denotes a positional encoding (learned or sinusoidal) added to motion tokens.

For attention, we form the context from the projected/position-encoded tokens and (with a slight abuse of notation) reuse

$$Q = M_t W_Q. \tag{8}$$

$$K = [M_t; \tilde{T}; \tilde{V}\,] W_K. \tag{9}$$

$$V = [M_t; \tilde{T}; \tilde{V}\,] W_V. \tag{10}$$

The attention result is processed by a Feed-Forward Network (FFN), where conditioning is injected via adaptive LayerNorm (AdaLN). Let $\text{pool}(T)$ be a pooled summary of text tokens, scale and shift are produced by small MLPs from $[e_t; \text{pool}(T)]$:

$$\begin{cases} \gamma = MLP_\gamma\big([e_t; pool(T)]\big), \\ \beta = MLP_\beta\big([e_t; pool(T)]\big). \end{cases}$$

AdaLN then applies

$$AdaLN(u) = \gamma \cdot LN(u) + \beta, \tag{11}$$

yielding a lightweight yet expressive conditioning mechanism effective in diffusion Transformers (Peebles and Xie, 2023). A final linear head maps hidden states back to the motion space to predict the noise residual $\varepsilon_\theta(x_t, t, c)$. The forward/reverse diffusion and training objective are detailed in Section 3.2.

3.2 Combined-attention Block

For consistency with Fig. 3.1, we use P for positional embeddings, $\oplus/\odot$ for element-wise add/multiply, and denote the timestep embedding by e_t.

To clarify the middle block of our pipeline, we refer to it as a unified conditioning stack. On the input side, the prompt y is normalized by a lightweight LLM and encoded as language tokens $T = E_{\text{text}}(y')$ with a CLIP text encoder; the optional scene image I is mapped to visual tokens $V = E_{\text{img}}(I)$ by the CLIP image encoder. Noised motion frames at step t are linearly projected to motion tokens $M_t = Proj(x_t) + P \in \mathbb{R}^{L\times D}$, and the timestep is embedded as $e_t = Emb(t)$. A pooled text summary $e_p = pool(T)$ serves as compact language context. For reference, the scaled dot-product attention kernel used inside the combined-attention block (CAB) is shown in Fig. 3.2.

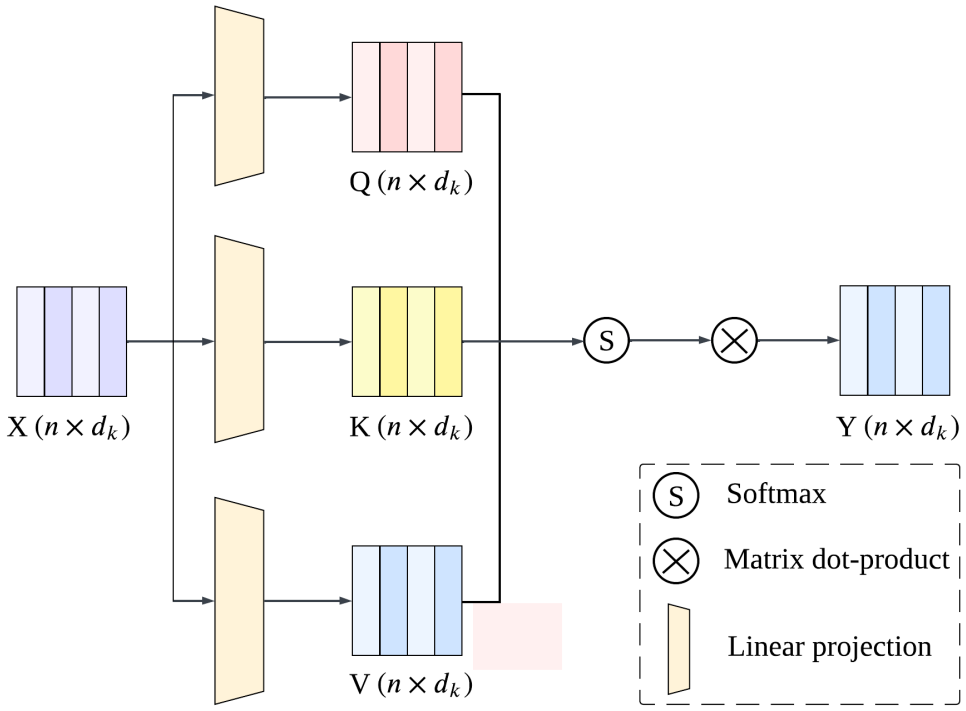

Figure 3.2. Scaled dot-product attention kernel (Q–K–V) used inside the combined-attention block (CAB).

Conditioning and interaction are realized in a single combined attention that merges self- and cross-attention into one pass, avoiding two separate blocks as in MotionDiffuse. We realize this with the combined attention defined in Eqs. (6)–(7), which merges self- and cross-attention in a single pass (see Fig. 3.3).

Motion-to-motion and motion-to-text/image interactions are handled jointly. For long sequences we adopt a linear-complexity implementation (Shen et al., 2021). A schematic of the merged operator is shown in Fig. 3.3.

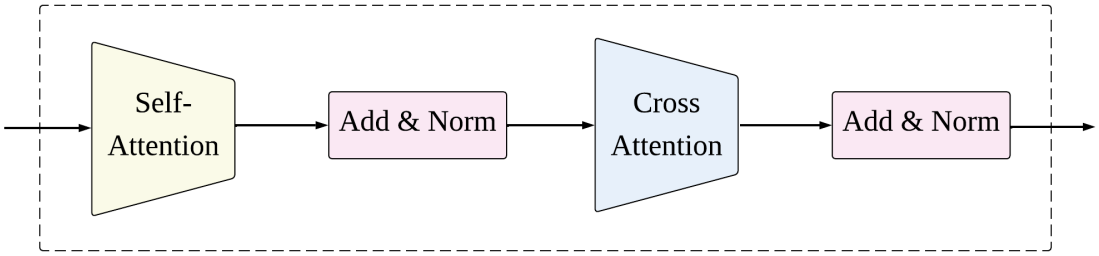

Figure 3.3. Combined attention: one attention over the joint token stream merges self-attention and cross-attention.

For readability we depict the operation as *Self→Add&Norm→Cross→Add&Norm*, in implementation we use a single multi-head attention over the concatenated token stream to realize the same interactions.

Let pool(T) denote a pooled text summary (mean pooling by default, or a CLASS token). The attention output is then transformed by an FFN where conditioning enters via adaptive LayerNorm (AdaLN). The FFN is a position-wise two-layer MLP with a nonlinearity (e.g., GELU). Scale and shift are generated from $[e_t; e_p]$ by two small MLPs and applied at the normalization entry,

The attention output is then passed to a position-wise FFN in which conditioning enters via adaptive LayerNorm (AdaLN), as defined in Eq. (8)–(9). The resulting AdaLN–FFN layout is depicted in Fig. 3.4.

A lightweight yet expressive conditioning mechanism effective in diffusion Transformers. The resulting AdaLN-FFN layout is depicted in Fig. 3.4.

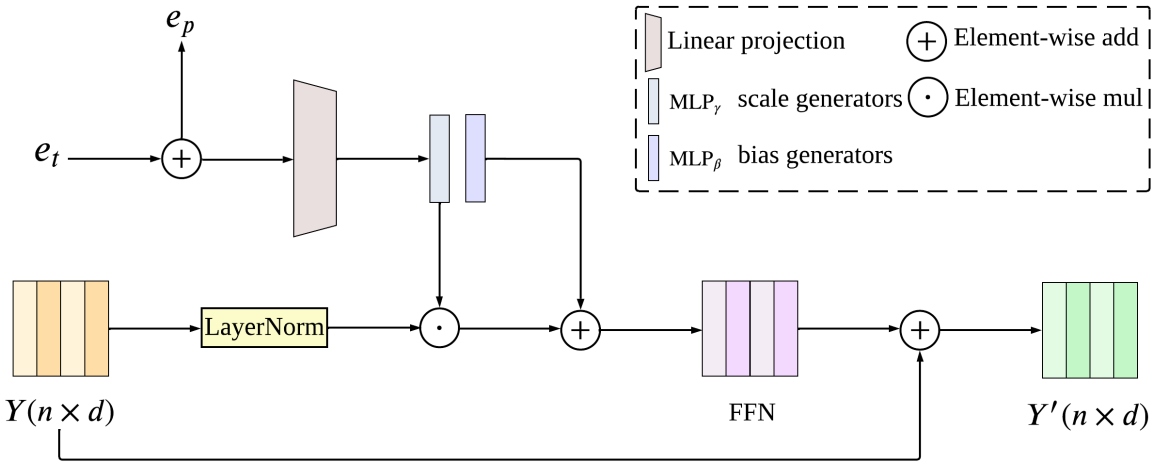

Figure 3.4. AdaLN-FFN: timestep/text summary produces (γ, β) to modulate the normalized hidden state before the FFN.

3.3 Diffusion Process

We adopt a denoising diffusion formulation that comprises a forward noising process, a reverse denoising process, and a noise-prediction objective. Consistent with Fig. 3.1 (top), we initialize from an isotropic Gaussian prior and diffuse a clean motion $x_0 = \{p_1, \ldots, p_L\}$ over T steps:

$$p(x_T) = \mathcal{N}(0, I), \qquad t \in \{1, \ldots, T\}. \tag{12}$$

The forward process perturbs x_0 into x_t by

$$q(x_t \mid x_0) = \sqrt{\bar{\alpha}_t}\, x_0 \; + \; \sqrt{1-\bar{\alpha}_t}\, \varepsilon, \tag{13}$$

$$\varepsilon \sim \mathcal{N}(0, I), \tag{14}$$

where $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$. In practice we use standard noise schedules (e.g., linear or cosine), yielding a smooth decay of the effective signal-to-noise ratio $\mathrm{SNR}(t) = \bar{\alpha}_t/(1-\bar{\alpha}_t)$ so that late steps approach $x_T \sim \mathcal{N}(0, I)$.

At inference (Fig. 3.1, bottom), we iteratively remove noise under the text–image context $c = [\tilde{T}; \tilde{V}]$ defined in Section 3.1 using the learned noise predictor $\varepsilon_\theta(x_t, t, c)$:

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\, \varepsilon_\theta(x_t, t, c) \right). \tag{15}$$

This ancestral update realizes generation as a sequence of small, likelihood-based refinements that are empirically stable for long horizons.

Training follows the standard ε-prediction objective used in text-driven motion diffusion: draw $t \sim \mathcal{U}\{1, \ldots, T\}$, sample $\varepsilon \sim \mathcal{N}(0, I)$, construct x_t via Eq. (13), and minimize

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{x_0, y', I, t, \varepsilon} \left[\left\| \varepsilon - \varepsilon_\theta(x_t, t, c) \right\|_2^2 \right]. \tag{16}$$

where x_0 is the clean motion sequence; y' is the LLM-normalized prompt and I is the optional scene image; $t \in \{1, \ldots, T\}$ is the diffusion step sampled uniformly; $\varepsilon \sim \mathcal{N}(0, I)$ is standard Gaussian noise; x_t is the noised motion obtained by Eq. (13); $c = [\tilde{T}; \tilde{V}]$ is the text–image context defined in Sec. 3.1 (if I is absent, $c = \tilde{T}$); $\varepsilon_\theta(\cdot)$ is the noise predictor with parameters θ; $\| \cdot \|_2^2$ denotes the squared ℓ_2 norm; and $\mathbb{E}[\cdot]$ is the expectation over training triplets (x_0, y', I), the timestep t, and noise ε.

4 Experiments

4.1 Dataset and Evaluation Metrics

4.1.1 Dataset

We conduct all primary training and evaluation on the HumanML3D dataset (Guo, Zou, et al., 2022), the prevailing benchmark for text-to-motion generation. It comprises over 14,616 motion clips, each paired with multiple textual descriptions, spanning a wide spectrum of human actions. Motions are represented using a 22-joint skeleton at 20 FPS. We adhere strictly to the official training, validation, and test splits to ensure fair and direct comparability with established benchmarks (Tevet, Raab, et al., 2022), (Petrovich, Black, and Varol, 2022), (M. Zhang et al., 2024).

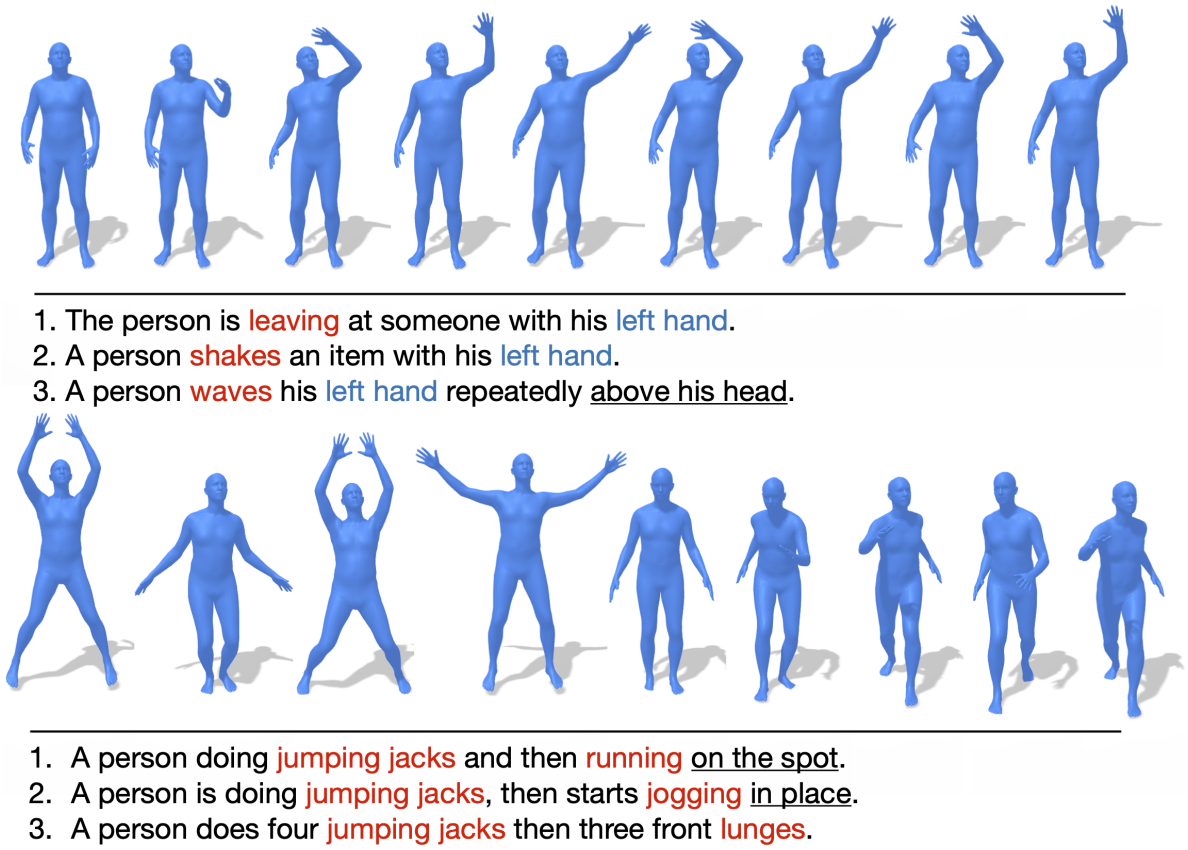

Figure 4.1. HumanML3D overview: text–motion pairs used in our experiments (Guo, Zou, et al., 2022).

To probe our model's capacity for scene-grounded synthesis, we construct a supplementary evaluation subset. For a selection of motions from the HumanML3D test set, we generate static background images using a pre-trained text-to-image diffusion model Ramesh et al., 2022, conditioned on the original captions but employing negative prompts (e.g., "no people, no humans") to ensure scenes contain only environmental context. This procedure yields a curated set of 1,960 high-quality image-motion pairs after manual verification for plausibility and lack of artifacts. Some examples of this scene subset are shown in Fig. 4.2.

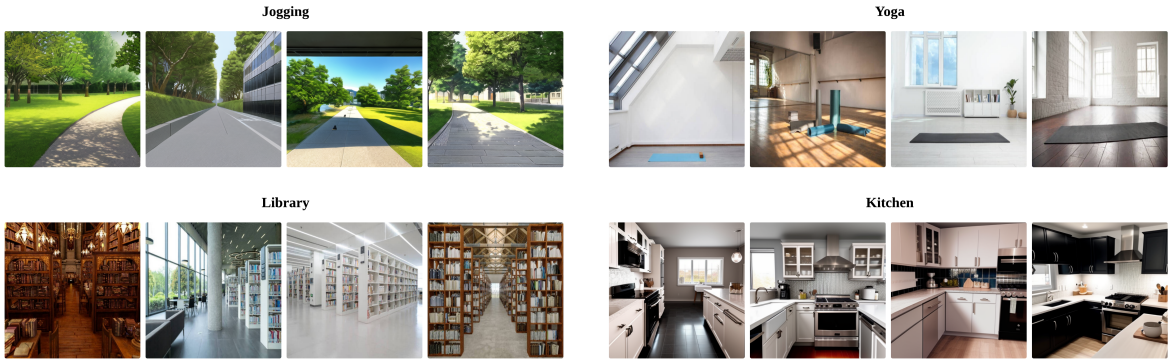

Figure 4.2. Our scene-image mini set: prompts paired with environment images; image is optional at inference.

4.1.2 Evaluation Metrics

Evaluation Metrics: We evaluate all methods with five metrics. 1) Fréchet Inception Distance (FID): a pretrained motion encoder extracts features for generated and ground-truth motions, and FID between the two distributions measures realism (lower is better). 2) R-Precision: for each generated motion and its text, 31 distractor prompts are sampled from the test set; a pretrained text–motion contrastive model computes top-1/top-2/top-3 retrieval accuracy (higher is better). 3) Diversity: generated motions for all test texts are randomly paired and the average inter-pair joint distance is reported (higher is better). 4) Multimodality: for each text, 32 motions are generated; the average pairwise joint distance among these samples quantifies one-to-many variability (higher is better). 5) Multimodal Distance (MD): using the same contrastive model, we report the ℓ_2 distance between the text embedding and the generated motion embedding (lower is better). When image conditioning is available, we additionally report the change (Δ) from text-only to text+image on our scene subset to indicate the benefit of scene context. Full quantitative results are reported in Table 4.1.

4.2 Results

We quantitatively compare TIDMG with representative baselines on HumanML3D under the standard protocol (all methods use the ground-truth motion length; each evaluation is repeated 10 times and we report the mean $\pm$ 80% CI). As shown in Table 4.1, TIDMG achieves the best or tied-best results across all metrics: it attains the highest R-Precision (top-1/top-2/top-3), lower Fréchet Inception Distance (FID), and lower Multimodal Distance (MD) than prior methods, while maintaining strong Diversity. Multimodality is also slightly higher than MotionDiffuse, indicating increased one-to-many variability without sacrificing alignment. Overall, the gap to real motions is narrowed on both retrieval-based and distributional metrics, suggesting that the generated sequences are both realistic and text-consistent.

Table 4.2 reports an ablation over four components of TIDMG: (i) the lightweight LLM for prompt normalization, (ii) optional scene-image conditioning, (iii) the combined-attention block that merges self- and cross-attention in a single pass, and (iv) an FFN with AdaLN conditioning. Removing any single component consistently degrades R-Precision (top-1/top-2/top-3), whereas integrating them yields the best performance, evidencing complementary effects. In particular, the combined-attention design improves both accuracy and efficiency by avoiding two separate attention blocks (see

Table 4.1. Quantitative results on the HumanML3D test set("→" means results are better if the metric is closer to the real motions. All methods use the real motion length from the ground truth. We run all the evaluation 10 times and ± indicates the 80% confidence interval. The best results are in bold).

Methods	**R-Precision ↑**			**FID ↓**	**MultiModal Dist ↓**	**Diversity →**	**Multimodality**
	Top 1	**Top 2**	**Top 3**				
Real motions	$0.511^{\pm.003}$	$0.703^{\pm.003}$	$0.797^{\pm.002}$	$0.002^{\pm.000}$	$2.974^{\pm.008}$	$9.503^{\pm.065}$	–
Language2Pose	$0.246^{\pm.002}$	$0.387^{\pm.002}$	$0.486^{\pm.002}$	$11.02^{\pm.046}$	$5.296^{\pm.008}$	$7.676^{\pm.058}$	–
Text2Gesture	$0.165^{\pm.001}$	$0.267^{\pm.002}$	$0.345^{\pm.002}$	$7.664^{\pm.030}$	$6.030^{\pm.008}$	$6.409^{\pm.071}$	–
MoCoGAN	$0.037^{\pm.000}$	$0.072^{\pm.001}$	$0.106^{\pm.001}$	$94.41^{\pm.021}$	$9.643^{\pm.006}$	$0.462^{\pm.008}$	$0.019^{\pm.000}$
Dance2Music	$0.033^{\pm.000}$	$0.065^{\pm.001}$	$0.097^{\pm.001}$	$66.98^{\pm.016}$	$8.116^{\pm.006}$	$0.725^{\pm.011}$	$0.043^{\pm.001}$
Guo et al.	$0.457^{\pm.002}$	$0.639^{\pm.003}$	$0.740^{\pm.003}$	$1.067^{\pm.002}$	$3.340^{\pm.008}$	$9.188^{\pm.002}$	$2.090^{\pm.083}$
TEMOS	$0.420^{\pm.002}$	$0.607^{\pm.002}$	$0.715^{\pm.002}$	$3.741^{\pm.030}$	$3.718^{\pm.010}$	$8.992^{\pm.079}$	$0.370^{\pm.019}$
MotionDiffuse	$0.491^{\pm.001}$	$0.681^{\pm.001}$	$0.782^{\pm.001}$	$0.630^{\pm.001}$	$3.113^{\pm.001}$	$9.410^{\pm.049}$	$1.553^{\pm.042}$
TIDMG (Ours)	$\mathbf{0.501^{\pm.002}}$	$\mathbf{0.687^{\pm.003}}$	$\mathbf{0.783^{\pm.001}}$	$\mathbf{0.625^{\pm.001}}$	$\mathbf{3.104^{\pm.001}}$	$\mathbf{9.415^{\pm.042}}$	$\mathbf{1.565^{\pm.040}}$

the *Time(s)* column), AdaLN delivers effective conditioning with minimal overhead, and the LLM normalization stabilizes language inputs and brings steady retrieval gains. When scene images are available, text+image conditioning further improves retrieval and MD on the scene subset, showing that the model can exploit visual context without changing the training objective.

Takeaway. Each component contributes measurable improvements, and their combination provides the strongest overall results, confirming both the effectiveness and the complementarity of our design choices.

Table 4.2. Ablation study on key components of our TIDMG framework. The full model integrating all components achieves the best performance.

LLM	**LLM+Scene Image**	**Combined Attention**	**FFN(with AdaLN)**	**Top 1**	**Top 2**	**Top 3**	**Time(s)**
-	-	-	-	$0.164^{\pm0.05}$	$0.218^{\pm0.02}$	$0.319^{\pm0.04}$	0.083
N	N	N	Y	$0.298^{\pm0.03}$	$0.322^{\pm0.04}$	$0.487^{\pm0.03}$	0.258
N	N	Y	Y	$0.356^{\pm0.04}$	$0.411^{\pm0.02}$	$0.342^{\pm0.02}$	0.266
Y	N	Y	Y	$0.482^{\pm0.02}$	$0.512^{\pm0.05}$	$0.621^{\pm0.03}$	0.128
Y	Y	Y	Y	$0.501^{\pm0.02}$	$0.687^{\pm0.03}$	$0.783^{\pm0.01}$	0.131

As shown in Fig. 4.3, under the two example prompts, TIDMG yields slightly more coherent and accurate motions—a steadier wide-stride walk and a low, continuous crawling trajectory—while adhering to the scene context.

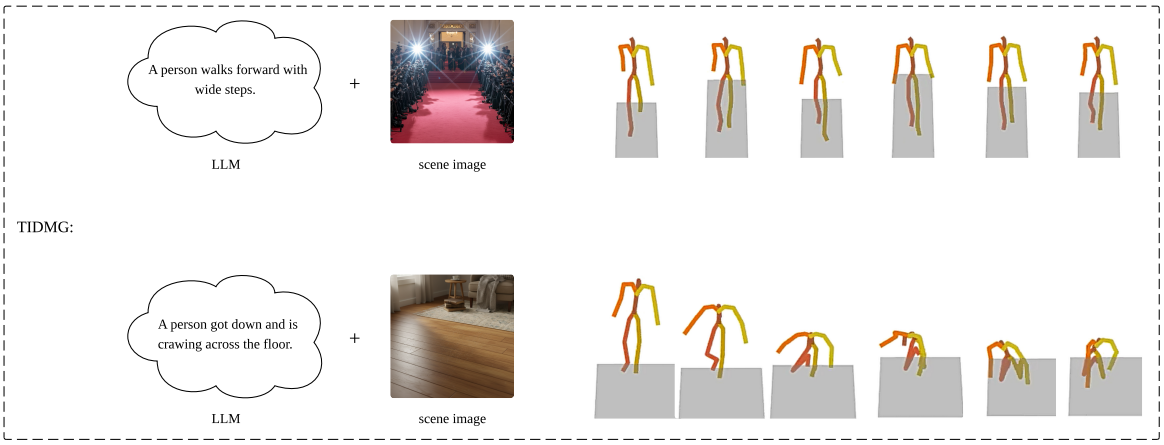

Figure 4.3. Under two prompts (“wide-step walk” and “crawl across the floor”), TIDMG produces slightly more coherent, scene-consistent motions.

5 Conclusion, Limitation and Future Work

In this work, we presented TIDMG, a novel Text-Image-Driven Motion Generation framework for producing 3D human motions that are semantically aligned with natural language prompts and spatially plausible within given scene contexts.

Firstly, we introduced a unified architecture that effectively integrates dual conditioning signals. Unlike prior methods that often rely on text alone or complex cascaded models, TIDMG concurrently processes a text prompt and an optional scene image. We leverage a lightweight LLM for text instruction normalization and a CLIP-based encoder for image context, projecting both modalities into a shared embedding space. The core of our approach is a combined-attention mechanism within a diffusion transformer, which efficiently fuses self-attention (for motion coherence) and cross-attention (for text-image conditioning) into a single, streamlined operation. This design eliminates the computational overhead of separate attention blocks, as used in frameworks like MotionDiffuse, while enhancing the interaction between different modalities. Furthermore, conditioning is refined through Adaptive Layer Normalization (AdaLN), which modulates the network's activations based on the timestep and a summary of the text context, ensuring the generative process is precisely guided by the conditional inputs.

Secondly, comprehensive quantitative and qualitative evaluations on the HumanML3D benchmark demonstrate the effectiveness of our approach. Our method achieves competitive performance, outperforming strong previous works like TEMOS, Guo et al., and MotionDiffuse across key metrics including R-Precision, FID, and MultiModality. The ablation studies further confirm the necessity of each key component in our architecture. Qualitatively, TIDMG generates diverse and semantically accurate motions that not only fulfill the action descriptions but also demonstrate a sensible understanding of the spatial constraints provided by the scene image.

Despite its strong performance, our work has limitations that point to valuable future directions. The current model's inference speed, though improved by our efficient attention design, is still bound by the iterative nature of the diffusion process. Future work could explore distillation techniques or advanced samplers to accelerate generation. Additionally, our framework currently assumes a static scene image; integrating dynamic scene interactions or video sequences would be a significant step toward more complex and realistic motion synthesis. Finally, exploring even larger-scale training and leveraging emerging powerful vision-language models could further enhance the fine-grained understanding and open-vocabulary generalization capabilities of the system.

In summary, TIDMG provides a robust and effective solution for conditional human motion generation. We believe our work offers a solid step towards unified and controllable motion synthesis models that can seamlessly understand and execute instructions from the multifaceted real world.

References

Ahuja, Chaitanya and Louis-Philippe Morency (2019). "Language2pose: Natural language grounded pose forecasting". In: *2019 International conference on 3D vision (3DV)*. IEEE, pp. 719–728 (cit. on pp. 6, 9).

Athanasiou, Nikos et al. (2024). "Motionfix: Text-driven 3d human motion editing". In: *SIGGRAPH Asia 2024 Conference Papers*, pp. 1–11 (cit. on p. 9).

Barsoum, Emad, John Kender, and Zicheng Liu (2018). "Hp-gan: Probabilistic 3d human motion prediction via gan". In: *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 1418–1427 (cit. on p. 8).

Cha, Junuk et al. (2024). "Text2hoi: Text-guided 3d motion generation for hand-object interaction". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1577–1585 (cit. on p. 9).

Chi, Seunggeun et al. (2024). "M2d2m: Multi-motion generation from text with discrete diffusion models". In: *European conference on computer vision*. Springer, pp. 18–36 (cit. on p. 9).

Dhariwal, Prafulla and Alexander Nichol (2021). "Diffusion models beat gans on image synthesis". In: *Advances in neural information processing systems* 34, pp. 8780–8794 (cit. on p. 8).

Dinh, Laurent, David Krueger, and Yoshua Bengio (2014). "Nice: Non-linear independent components estimation". In: *arXiv preprint arXiv:1410.8516* (cit. on p. 8).

Dinh, Laurent, Jascha Sohl-Dickstein, and Samy Bengio (2016). "Density estimation using real nvp". In: *arXiv preprint arXiv:1605.08803* (cit. on p. 8).

Diomataris, Markos et al. (2024). "WANDR: Intention-guided human motion generation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 927–936 (cit. on p. 9).

Gao, Xuehao et al. (2024). "Guess: Gradually enriching synthesis for text-driven human motion generation". In: *IEEE Transactions on Visualization and Computer Graphics* 30.12, pp. 7518–7530 (cit. on p. 9).

Gavrila, Dariu M (1999). "The visual analysis of human movement: A survey". In: *Computer vision and image understanding* 73.1, pp. 82–98 (cit. on p. 8).

Ghosh, Anindita et al. (2021). "Synthesis of compositional animations from textual descriptions". In: *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1396–1406 (cit. on pp. 6, 9).

Goodfellow, Ian J et al. (2014). "Generative adversarial nets". In: *Advances in neural information processing systems* 27 (cit. on p. 8).

Guo, Chuan, Yuxuan Mu, et al. (2024). "Momask: Generative masked modeling of 3d human motions". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1900–1910 (cit. on p. 9).

Guo, Chuan, Shihao Zou, et al. (2022). "Generating diverse and natural 3d human motions from text". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5152–5161 (cit. on pp. 6, 8, 9, 15).

Han, Bo et al. (2024). "Amd: Autoregressive motion diffusion". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 38. 3, pp. 2022–2030 (cit. on p. 9).

Henter, Gustav Eje, Simon Alexanderson, and Jonas Beskow (2020). "Moglow: Probabilistic and controllable motion synthesis using normalising flows". In: *ACM Transactions on Graphics (TOG)* 39.6, pp. 1–14 (cit. on p. 8).

Ho, Jonathan, Ajay Jain, and Pieter Abbeel (2020). "Denoising diffusion probabilistic models". In: *Advances in neural information processing systems* 33, pp. 6840–6851 (cit. on pp. 6, 8).

Hochreiter, Sepp and Jürgen Schmidhuber (1997). "Long short-term memory". In: *Neural computation* 9.8, pp. 1735–1780 (cit. on p. 8).

Huang, Ruozi et al. (2020). "Dance revolution: Long-term dance generation with music via curriculum learning". In: *International conference on learning representations* (cit. on p. 9).

Kingma, Diederik et al. (2021). "Variational diffusion models". In: *Advances in neural information processing systems* 34, pp. 21696–21707 (cit. on p. 8).

Kingma, Diederik P and Max Welling (2013). "Auto-encoding variational bayes". In: *arXiv preprint arXiv:1312.6114* (cit. on p. 8).

Kingma, Durk P and Prafulla Dhariwal (2018). "Glow: Generative flow with invertible 1x1 convolutions". In: *Advances in neural information processing systems* 31 (cit. on p. 8).

Kulkarni, Nilesh et al. (2024). "Nifty: Neural object interaction fields for guided human motion synthesis". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 947–957 (cit. on p. 9).

Li, Jiaman et al. (2020). "Learning to generate diverse dance motions with transformer". In: *arXiv preprint arXiv:2008.08171* (cit. on pp. 6, 9).

Li, Ruilong et al. (2021). "Ai choreographer: Music conditioned 3d dance generation with aist++". In: *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 13401–13412 (cit. on pp. 6, 9).

Liang, Han, Jiacheng Bao, et al. (2024). "Omg: Towards open-vocabulary motion generation via mixture of controllers". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 482–493 (cit. on p. 9).

Liang, Han, Wenqian Zhang, et al. (2024). "Intergen: Diffusion-based multi-human motion generation under complex interactions". In: *International Journal of Computer Vision* 132.9, pp. 3463–3483 (cit. on p. 9).

Mahmood, Naureen et al. (2019). "AMASS: Archive of motion capture as surface shapes". In: *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 5442–5451 (cit. on pp. 6, 9).

Min, Jianyuan and Jinxiang Chai (2012). "Motion graphs++ a compact generative model for semantic motion analysis and synthesis". In: *ACM Transactions on Graphics (TOG)* 31.6, pp. 1–12 (cit. on p. 8).

Mir, Aymen et al. (2024). "Generating continual human motion in diverse 3d scenes". In: *2024 International Conference on 3D Vision (3DV)*. IEEE, pp. 903–913 (cit. on p. 9).

Mukai, Tomohiko and Shigeru Kuriyama (2005). "Geostatistical motion interpolation". In: *ACM SIGGRAPH 2005 Papers*, pp. 1062–1070 (cit. on p. 8).

Nichol, Alex et al. (2021). "Glide: Towards photorealistic image generation and editing with text-guided diffusion models". In: *arXiv preprint arXiv:2112.10741* (cit. on p. 8).

Nichol, Alexander Quinn and Prafulla Dhariwal (2021). "Improved denoising diffusion probabilistic models". In: *International conference on machine learning*. PMLR, pp. 8162–8171 (cit. on p. 8).

O'rourke, Joseph and Norman I Badler (1980). "Model-based image analysis of human motion using constraint propagation". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 6, pp. 522–536 (cit. on p. 8).

Ormoneit, Dirk et al. (2005). "Representing cyclic human motion using functional analysis". In: *Image and Vision Computing* 23.14, pp. 1264–1276 (cit. on p. 8).

Peebles, William and Saining Xie (2023). "Scalable diffusion models with transformers". In: *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4195–4205 (cit. on pp. 6, 9, 11).

Petrovich, Mathis, Michael J Black, and Gül Varol (2021). "Action-conditioned 3d human motion synthesis with transformer vae". In: *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10985–10995 (cit. on pp. 6, 9).

— (2022). "Temos: Generating diverse human motions from textual descriptions". In: *European Conference on Computer Vision*. Springer, pp. 480–497 (cit. on pp. 6, 8, 9, 15).

Petrovich, Mathis, Or Litany, et al. (2024). "Multi-track timeline control for text-driven 3d human motion generation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1911–1921 (cit. on p. 9).

Pinyoanuntapong, Ekkasit et al. (2024). "Bamm: Bidirectional autoregressive motion model". In: *European Conference on Computer Vision*. Springer, pp. 172–190 (cit. on p. 9).

Plappert, Matthias, Christian Mandery, and Tamim Asfour (2016). "The kit motion-language dataset". In: *Big data* 4.4, pp. 236–252 (cit. on p. 9).

Punnakkal, Abhinanda R et al. (2021). "BABEL: Bodies, action and behavior with english labels". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 722–731 (cit. on p. 9).

Radford, Alec et al. (2021). "Learning transferable visual models from natural language supervision". In: *International conference on machine learning*. PmLR, pp. 8748–8763 (cit. on pp. 8, 9).

Ramesh, Aditya et al. (2022). "Hierarchical text-conditional image generation with clip latents". In: *arXiv preprint arXiv:2204.06125* 1.2, p. 3 (cit. on pp. 8, 15).

Rose, Charles, Michael F Cohen, and Bobby Bodenheimer (1998). "Verbs and adverbs: Multidimensional motion interpolation". In: *IEEE Computer Graphics and Applications* 18.5, pp. 32–40 (cit. on p. 8).

Ruiz-Ponce, Pablo et al. (2025). "Mixermdm: Learnable composition of human motion diffusion models". In: *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 12380–12390 (cit. on p. 9).

Shen, Zhuoran et al. (2021). "Efficient attention: Attention with linear complexities". In: *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 3531–3539 (cit. on pp. 8, 9, 12).

Siyao, Li et al. (2022). "Bailando: 3d dance generation by actor-critic gpt with choreographic memory". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11050–11059 (cit. on p. 9).

Sun, Haowen et al. (2024). “Lgtm: Local-to-global text-driven human motion diffusion model”. In: *ACM SIGGRAPH 2024 Conference Papers*, pp. 1–9 (cit. on p. 9).

Tevet, Guy, Brian Gordon, et al. (2022). “Motionclip: Exposing human motion generation to clip space”. In: *European Conference on Computer Vision*. Springer, pp. 358–374 (cit. on p. 9).

Tevet, Guy, Sigal Raab, et al. (2022). “Human motion diffusion model”. In: *arXiv preprint arXiv:2209.14916* (cit. on pp. 6, 8, 9, 15).

Vaswani, Ashish et al. (2017). “Attention is all you need”. In: *Advances in neural information processing systems* 30 (cit. on pp. 6, 8, 9, 11).

Wan, Weilin et al. (2024). “Tlcontrol: Trajectory and language control for human motion synthesis”. In: *European Conference on Computer Vision*. Springer, pp. 37–54 (cit. on p. 9).

Wang, Zan et al. (2024). “Move as you say interact as you can: Language-guided human motion generation with scene affordance”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 433–444 (cit. on p. 9).

Xue, Kebing et al. (2025). “Shape-Conditioned Human Motion Diffusion Model with Mesh Representation”. In: *Computer Graphics Forum*. Wiley Online Library, e70065 (cit. on p. 9).

Yan, Xinchen et al. (2018). “Mt-vae: Learning motion transformations to generate multimodal human dynamics”. In: *Proceedings of the European conference on computer vision (ECCV)*, pp. 265–281 (cit. on p. 8).

Zhang, Mingyuan et al. (2024). “Motiondiffuse: Text-driven human motion generation with diffusion model”. In: *IEEE transactions on pattern analysis and machine intelligence* 46.6, pp. 4115–4128 (cit. on pp. 6, 8, 11, 15).

Zhang, Yaqi et al. (2024). “Motiongpt: Finetuned llms are general-purpose motion generators”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 38. 7, pp. 7368–7376 (cit. on p. 9).

Zhang, Zeyu et al. (2024). “Motion mamba: Efficient and long sequence motion generation”. In: *European Conference on Computer Vision*. Springer, pp. 265–282 (cit. on p. 9).

Zhu, Wentao et al. (2023). “Human motion generation: A survey”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 46.4, pp. 2430–2449 (cit. on p. 8).