STYLIZED TEXT-TO-MOTION SYNTHESIS VIA MULTI-CONDITION LATENT DIFFUSION

Fanglu Xie[⋆†] *Tsukasa Shiota*[⋆] *Motohiro Takagi*[⋆] *Edgar Simo-Serra*[†]

[⋆] NTT, Inc., Kanagawa 239-0847, Japan
[†] Waseda University, Tokyo 169-8555, Japan

ABSTRACT

Generating high-quality motion sequences from textual descriptions has become a prominent research area in motion synthesis. For end applications, generated motions need to be diverse, natural, and conform to the textual description. Furthermore, motions include factors such as style and trajectory, which are hard to control. Finding effective ways to manage these factors is crucial for achieving realistic motion generation. To address these challenges, we first propose a multi-condition motion latent diffusion model that integrates style and trajectory information into text-driven generation, enabling diverse stylized motions and precise control with arbitrary trajectories. To preserve text controllability, we apply an adapter that refines a pretrained text-to-motion model by transforming the style and trajectory conditions while fully utilizing the pretrained knowledge. Finally, during inference, we apply explicit trajectory guidance within our classifier-free multi-guidance, ensuring that the produced trajectories follow the intended input path. Our experimental results shows the effectiveness of the proposed approach, achieving state-of-the-art performance in text-to-motion generation and exhibiting high flexibility in stylized motion synthesis. Our work unifies text-driven motion synthesis, style transfer, and trajectory control within a single framework, paving the way for more versatile applications in animation, human interaction, and virtual reality.

Index Terms— motion synthesis, diffusion model, stylized motion, text-to-motion

1. INTRODUCTION

The creation of high-quality motion sequences from textual descriptions has become a significant area of research in motion synthesis. Researchers strive to generate motions that are not only diverse and natural but also accurately reflect the specific details mentioned in the text. This task is complex due to the challenges of controlling various elements of the motion, such as style—how the movement looks—and trajectory—the path that the movement follows. Each of these challenges must be carefully addressed to produce compelling and accurate representations of motion. Existing approaches suffer from limitations when it comes to generating motions due to the complexity and unpredictability of the problem. (1) Text-to-motion challenge: Text-to-motion models [1, 2, 3, 4, 5] directly generate motion from text but struggle with ambiguity and diversity. For example, the word "kick" can refer to very different motions, and a single motion can be described in many ways, making learning difficult. (2) Style transfer limitation: Style transfer methods [6, 7, 8] can generate style-based, and content-based motions by combining input motions, but it cannot generate motions from textual description. While content-style alignment has succeeded with motion data alone, it fails to adapt styled motions to textual content descriptions. (3) Stylized text-to-motion challenge: Stylized text-to-motion approaches [9, 10] aim to generate motion from text descriptions and refine the outputs using additional modality data, such as stylized motion, audio, and more. However, these methods still fall short in achieving high controllability and generation accuracy. (4) Trajectory control issue: Trajectory-aware methods [11, 12, 13] still struggle to achieve effective trajectory control and often suffer from artifacts such as unnatural motion paths, revealing limited spatial control.

We propose a unified text-to-motion generation framework that controls motion based on textual descriptions while incorporating style and trajectory information in the latent diffusion space. Our method introduces a latent-space integration mechanism that jointly integrates style motions and trajectory coordinates with text embeddings. During training, because jointly controlling both style and trajectory is highly challenging, we retain the pretrained MLD [5] and introduce a lightweight adapter for fine-tuning. Rather than training a new model from scratch, this adapter enables diverse stylized motion generation, preserves content integrity, and supports multimodal conditioning without overwriting the original text semantics. During inference, we apply trajectory guidance through our classifier-free multi-guidance mechanism, enabling simultaneous control of style and trajectory and yielding more realistic and faithful motion generation.

In summary, our main contributions are:

- We first fuse multimodal motion styles and trajectories with textual input in the latent space, achieving a unified text-to-motion generation process with multi-factor controllability.
- We apply an adapter that injects multi-conditions into the model without breaking text controllability, thereby achiev-

ing superior performance in stylized text-to-motion generation by striking a more effective balance between content fidelity and style accuracy.
- We incorporate an explicit trajectory guidance mechanism in our proposed classifier-free multi-guidance, allowing the model to generate trajectories that faithfully follow the intended input path.

2. RELATED WORK

Human motion generation includes various methods: text-to-motion generation refers to producing motion sequences directly from textual descriptions; motion style transfer focuses on applying a specific style to existing motions; controllable text-to-motion generation allows for controlling the motion based on additional conditions.

Text-to-Motion Generation. Recent progress in human motion generation has been driven by transformer-based models [2, 4] and diffusion models [1, 3, 5]. Momask [4] improves motion synthesis using a residual VQ-VAE with multiple codebooks, while Guo et al. [2] introduce a VAE-based approach together with the large-scale HumanML3D dataset. MDM [1] establishes a strong transformer-based diffusion baseline for text-to-motion generation, and MLD [5] further improves efficiency by performing diffusion in the latent space. However, these methods often generate motions that deviate from the input text and lack mechanisms to incorporate external controls such as style or trajectory.

Motion Style Transfer. Motion style transfer is a widely used technique for creating stylized movements by transferring the style from a reference motion to a source motion. Aberman et al. [6] propose a generative adversarial network designed to separate motion style from content, allowing for their recombination without needing paired data. Motion Puzzle [7] presents a generative framework that enables control over the style of individual body parts. Additionally, MCMLDM [8] introduces a diffusion-based approach that incorporates trajectory awareness and achieves style transfer through Adaptive Instance Normalization (AdaIN). These methods require two separate motion inputs, which sets a high demand on data and usage. They cannot directly convert text inputs into styled motions, resulting in limited flexibility.

Controllable Text-to-Motion Generation. In text-to-motion generation, numerous methods have been proposed to improve controllability over either style or trajectory. Applying multiple conditions in continuous diffusion space often leads to conflicts between style and trajectory. Style-control methods, Smoodi [10] and [9], combine style with text in latent space, but diffusion-based text-to-motion generation remains highly stochastic, resulting in unstable trajectories. Existing trajectory control methods rely on inpainting constraints during diffusion [11] or explicit control of input motions [8], rather than controlling trajectory in latent diffusion. Generating motions directly from text while controlling both style and trajectory continues to be a major challenge.

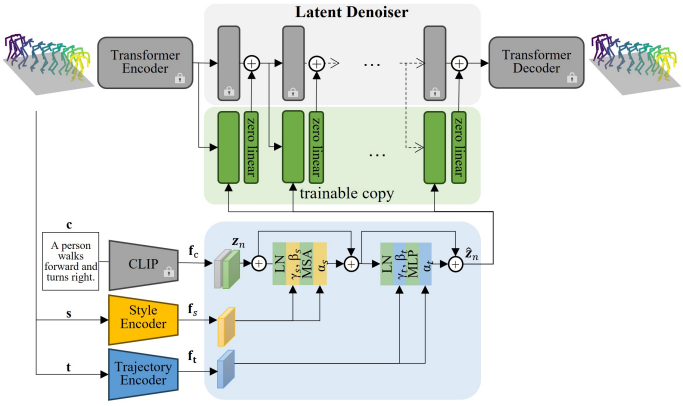

Fig. 1. **Overview of our proposed approach.** Our model generates stylized human motions from content text, style, and trajectory inputs. CLIP, the Style Encoder, and the Trajectory Encoder extract f_c, f_s, and f_t. These features guide the denoising process, where z_n is iteratively updated to $\hat{z_n}$ using a style–trajectory adaptor on pretrained MLD. The motion decoder then produces the final stylized motion.

3. PROPOSED APPROACH

We propose a method that utilizes diffusion models to integrate style motion and trajectory data as additional inputs, enabling control over motion features based on descriptive text in the latent space, shown in Fig. 1. For training, we introduce an adapter-based approach to adapt the pretrained model and design a mechanism that jointly applies stylistic and trajectory constraints to achieve a balanced contribution from each condition, as described in Section 3.1. For generation, we propose a novel guidance strategy that further enhances multi-condition controllability, as detailed in Section 3.2.

We utilize the same motion representations in HumanML-3D [2], $x_0 \in \mathbb{R}^{L\times D}$, where L denotes the frame length of the motion and $D = 263$ indicates the dimension of human motion representations. The style encoder E_s and trajectory encoder E_t process motion inputs to extract style features ($\mathbf{f}_s$) and trajectory features ($\mathbf{f}_t$). The text features are obtained from the CLIP model [14].

$$\mathbf{f}_c = CLIP(\mathbf{c}), \mathbf{f}_s = E_s(\mathbf{s}), \mathbf{f}_t = E_t(\mathbf{t}) \tag{1}$$

In our method, the diffusion process is modeled as a Markov chain that gradually perturbs an initial latent motion feature $\mathbf{z}_0$ into a noisy latent $\mathbf{z}_n \sim \mathcal{N}(0, \mathbf{I})$ via Gaussian noise:

$$q(\mathbf{z}_n \mid \mathbf{z}_{n-1}) = \mathcal{N}\left(\sqrt{\alpha_n}\mathbf{z}_{n-1}, (1-\alpha_n)\mathbf{I}\right), \tag{2}$$

where $n \in 1, \ldots, N$ and α_n controls the noise schedule. The reverse process iteratively denoises $\mathbf{z}_n$ to recover $\mathbf{z}_0$, which is decoded into a stylized motion sequence $\hat{x}_0 \in \mathbb{R}^{L\times D}$.

To guide the denoising with multiple conditions, we design a multi-condition denoiser ϵ_θ, which predicts the noise

at each step n using the noisy latent $\mathbf{z}_n$, the timestep n, and guided features $(\mathbf{f}_c, \mathbf{f}_t, \mathbf{f}_s)$: We define the denoising process at step $n \in (0, N]$ as $\epsilon_n = \epsilon_\theta(\mathbf{z}_n, n, \mathbf{f}_c, \mathbf{f}_s, \mathbf{f}_t)$, where ϵ_θ represents the predicted noise at timestep n.

3.1. Adapter-based Multi-Condition Latent Diffusion

Training a model to simultaneously capture style, trajectory, and text remains highly challenging. Inspired by ControlNet [15], which efficiently controls complex image generation with multi-modal conditions, we design a content-aware style and trajectory adapter for the pretrained MLD [5] model. We create a trainable adapter by copying the MLD denoiser, while keeping the original MLD backbone frozen. In addition, we design a lightweight style encoder E_s and a trajectory encoder E_t whose features are injected into the adapter to modulate the generation process. The adapter output is merged with that of the frozen denoiser through a zero linear layer, thereby preserving the pretrained model's capability. During training, only the adapter and style/trajectory encoder are optimized on the style motion dataset, 100STYLE [16].

In addition, we design an integration mechanism that fuses style and trajectory representations with text embeddings directly in the latent space. Specifically, the text condition feature $\mathbf{f}_c$ is concatenated with the noisy latent feature to form $\hat{\mathbf{z}}_n = \text{concat}(\mathbf{z}_n, \mathbf{f}_c)$, allowing the text condition to consistently influence the denoising process throughout all steps. Subsequently, to incorporate style ($\mathbf{f}_s$) and the trajectory ($\mathbf{f}_t$) and conditions, we first extract modulation parameters from each using separate multi-layer perceptrons:

$$\gamma_s, \beta_s, \alpha_s = \text{MLP}_s(\mathbf{f}_s),\ \gamma_t, \beta_t, \alpha_t = \text{MLP}_t(\mathbf{f}_t) \tag{3}$$

Here, γ_s, β_s, and α_s are style-related modulation parameters, while γ_t, β_t, and α_t correspond to the trajectory condition. $\text{MLP}_s(\cdot)$ and $\text{MLP}_t(\cdot)$ are independent networks used to process $\mathbf{f}_s$ and $\mathbf{f}_t$, respectively. These parameters are incorporated into the denoising model ϵ_θ using AdaLN-Zero [17], which modulates each transformer layer as follows:

$$\begin{aligned} \hat{\mathbf{z}}^{(1)}_{n,k} &= \hat{\mathbf{z}}_{n,k-1} + \alpha_s \text{MSA}(\text{LN}(\hat{\mathbf{z}}_{n,k-1})\gamma_s + \beta_s) \\ \hat{\mathbf{z}}_{n,k} &= \hat{\mathbf{z}}^{(1)}_{n,k} + \alpha_t \text{MLP}(\text{LN}(\hat{\mathbf{z}}^{(1)}_{n,k})\gamma_t + \beta_t) \end{aligned} \tag{4}$$

where $\text{MSA}(\cdot)$ denotes multi-head self-attention, and $\text{LN}(\cdot)$ refers to layer normalization. The variables $\hat{\mathbf{z}}_{n,k-1}$ and $\hat{\mathbf{z}}_{n,k}$ represent outputs from the $(k-1)$-th and k-th layers of the latent denoiser model ϵ_θ. The variable $\hat{\mathbf{z}}^{(1)}_{n,k}$ indicates an intermediate value within the k-th layer, while $\text{MLP}(\cdot)$ stands for a multi-layer perceptron. Using AdaLN-Zero to apply secondary conditions at each layer of ϵ_θ effectively guides the denoising process hierarchically.

Training is based on a denoising score-matching loss:

$$\mathcal{L}_{\text{std}} = \mathbb{E}_{\epsilon,z}\left[\|\epsilon_\theta(\mathbf{z}_n, n, \mathbf{f}_c, \mathbf{f}_s, \mathbf{f}_t) - \epsilon\|_2^2\right] \tag{5}$$

where $\epsilon \sim N(0, I)$ is the ground-truth noise added to $\mathbf{z}_0$. Our design improves motion content preservation and enhances style transfer by capturing stylistic intent while maintaining trajectory structure.

3.2. Classifier-Free Multi-Guided Motion Generation

During the generation process, it is crucial to balance style and trajectory control. Since Smoodi [10] does not explicitly model trajectory constraints, we build upon it and propose an improved classifier-free guidance (CFG) [18] scheme that jointly enforces content, style, and trajectory control. Specifically, we adopt a Classifier-Free Multi-Guidance (CFMG) framework, which decomposes the overall guidance signal into three components: content guidance, style guidance, and trajectory guidance. Each component is derived in a classifier-free manner without relying on external classifiers. The guidance terms are applied according to their functional roles: content and style guidance operate on the motion generation process, while trajectory guidance is treated independently, since trajectory evolution is not directly constrained by other motion attributes. This design enables explicit trajectory control while preserving stylistic expressiveness and content consistency. Formally, the overall guidance is:

$$\begin{aligned} \epsilon_\theta(\mathbf{z}_n, n, \mathbf{f}_c, \mathbf{f}_s, \mathbf{f}_t) = \epsilon_\theta(\mathbf{z}_n, n, \varnothing, \varnothing, \varnothing) + \\ \underbrace{w_c\left(\epsilon_\theta(\mathbf{z}_n, n, \mathbf{f}_c, \varnothing, \varnothing) - \epsilon_\theta(\mathbf{z}_n, n, \varnothing, \varnothing, \varnothing)\right)}_{\text{Classifier-free Content Guidance}} + \\ \underbrace{w_s\left(\epsilon_\theta(\mathbf{z}_n, n, \mathbf{f}_c, \mathbf{f}_s, \varnothing) - \epsilon_\theta(\mathbf{z}_n, n, \mathbf{f}_c, \varnothing, \varnothing)\right)}_{\text{Classifier-free Style Guidance}} + \\ \underbrace{w_t\left(\epsilon_\theta(\mathbf{z}_n, n, \mathbf{f}_c, \mathbf{f}_s, \mathbf{f}_t) - \epsilon_\theta(\mathbf{z}_n, n, \mathbf{f}_c, \mathbf{f}_s, \varnothing)\right)}_{\text{Classifier-free Trajectory Guidance}} \end{aligned} \tag{6}$$

Here, w_c, w_s, and w_t are the guidance weights for the content ($\mathbf{f}_c$), style ($\mathbf{f}_s$), and trajectory ($\mathbf{f}_t$) conditions, respectively. The symbol $\varnothing$ indicates the absence of a specific condition.

This formulation improves generation quality and consistency by combining conditional and unconditional predictions, while enabling independent modulation of content, style, and trajectory influences to maintain flexibility and fine-grained control.

4. EXPERIMENTS

We use the HumanML3D dataset [2] for motion content and the 100STYLE dataset [16] for motion styles. HumanML3D is the largest text-annotated motion dataset, while 100STYLE provides the largest collection of 100 diverse motion styles. We use MLD [5] as the pretrained generative network and train the style/trajectory network with the denoiser. The style/trajectory encoder consists of a single transformer encoder, while the adapter is a multilayer perceptron (MLP). The framework is trained with the AdamW optimizer [19] at a learning rate of 1×10^{-5}. The classifier-free multi-guidance scale w_c is set to 7.5, w_s is 0.8 and w_t is 1.5.

Table 1. Stylized text-to-motion performance for transferring the HumanML3D-text [2] with the 100STYLE-motion [16].

Method	Style Condition	SFID↓	CFID↓	SRA↑ (Top-3)	Diversity↑	Traj. err.↓
Ours	Motion	**2.899**	4.598	**94.403**	**14.826**	**0.848**
Smoodi[10]	Motion	7.372	1.619	73.099	12.429	0.953
MLD[5]+MCMLDM[8]	Motion	6.603	11.254	56.948	14.180	1.703
ChatGPT[20]+MLD[5]	Text	6.012	**1.566**	11.630	12.549	1.039

Table 2. Ablation studies on classifier-free multi-guidance.

Method	SFID↓	CFID↓	SRA↑ (Top-3)	Diversity↑	Traj. err.↓
Ours w/ CFMG	**2.899**	**4.598**	**94.403**	**14.826**	0.848
only w/ style CFG	4.067	4.968	84.362	10.710	0.760
only w/ traj. CFG	7.814	7.591	60.985	10.626	**0.648**
w/o any CFG	13.067	9.812	48.502	5.787	0.884

We evaluate the performance of our method on stylized text-to-motion generation, trajectory control, and motion diversity. We compare our approach with the stylized text-to-motion method Smoodi [10]. In addition, we utilize the text-to-motion model MLD [5] and apply style transfer methods [7] to generate stylized motions. For trajectory control, we adopt the MLD [5] model in combination with trajectory control method, MCMLDM [8]. In scenarios where only text input is available, we employ ChatGPT [20] together with MLD, which relies solely on textual descriptions, and this experiment serves only as a qualitative reference.

Stylized Text-to-motion Performance. The transferred results are evaluated on both the HumanML3D and 100STYLE datasets. The results are obtained by transferring HumanML3D-text [2] with 100STYLE-motion [16] to generate stylized motions, shown in Table. 1. The style features are derived from 100STYLE-motion, while the trajectory comes from HumanML3D-motions. First, we use two Fréchet Inception Distance (FID) [21] metrics to assess motion style transfer quality: SFID (style FID) and CFID (content FID).

$$\text{SFID} = \text{FID}(\phi(x_{\text{transferred}}),\ \phi(x_{\text{100STYLE}})) \tag{7}$$

$$\text{CFID} = \text{FID}(\phi(x_{\text{transferred}}),\ \phi(x_{\text{HumanML3D}})) \tag{8}$$

Here, $\phi(\cdot)$ is a pretrained motion encoder [2] used to extract semantic features for FID computation. Our method achieves the lowest SFID, demonstrating a strong alignment with the target style. The CFID remains comparable to the SFID, indicating a well-balanced trade-off between style fidelity and content preservation. We further evaluate style recognition accuracy (SRA) using a pretrained style classifier on the filtered 100STYLE dataset. Our results show that our approach attains the highest Top-3 performance. In addition, the generated motions exhibit the greatest diversity among all methods. We also evaluate the accuracy of generated global trajectories using trajectory error, which measures absolute deviation. Our method achieves the lowest trajectory error, demonstrating the most accurate trajectory generation.

Ablation Study. Table 2 shows the ablation study on the impact of Classifier-Free Multi-Guidance (CFMG). Our full model, which includes both style and trajectory guidance, CFMG, achieves the best overall performance, striking a strong balance between style fidelity, motion quality, and diversity. Using only style CFG helps style generation but lowers content preservation. Conversely, using only trajectory CFG improves physical plausibility but fails to retain stylistic consistency. Without any CFG, performance drops significantly across all metrics, highlighting the necessity of CFG for generating realistic and style-consistent motion.

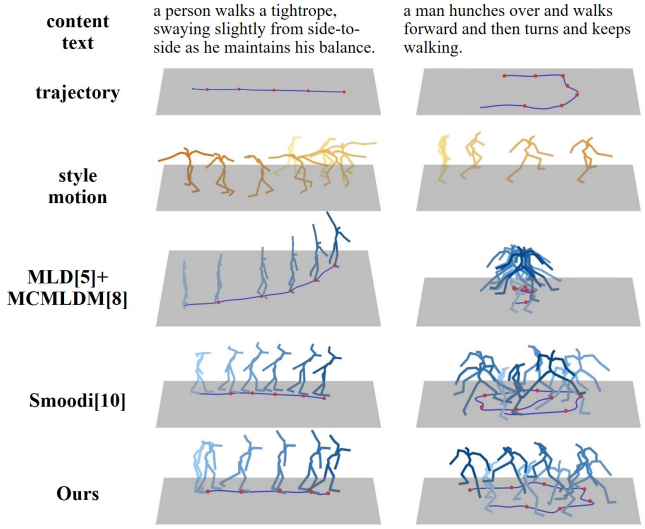

Fig. 2. **Visualization.** The input conditions for the content include text and trajectory, and the style includes motions.

Visualization. Fig. 2 shows motion generation results conditioned on content text, trajectory, and style motion. Compared with prior methods, our approach more faithfully captures the intended style while accurately following the target trajectory. The generated motions clearly inherit stylistic patterns (e.g., arm posture or leg openness) and precisely align with the input paths, including turns and straight movements. In contrast, MLD [5]+MCMLDM [8] often fails to follow the trajectory, and Smoodi [10] produces motions that deviate from both the expected style and path. These results demonstrate that our method more effectively integrates high-level semantic intent with low-level motion constraints.

5. CONCLUSIONS

We propose a multi-condition motion latent diffusion model that addresses the challenges of text-driven motion generation by integrating style and trajectory as distinct conditions. We propose a text-driven latent denoiser with multiple conditions and an adapter to finetune the denoiser. This method enhances the duality of content and style, allowing for diverse, natural, and text-conforming motion synthesis while improving trajectory control. Our experiments demonstrate that our model excels in stylized text-to-motion performance and shows enhanced adaptability for motion trajectory control.

6. REFERENCES

[1] G. Tevet, S. Raab, B. Gordon, Y. Shafir, D. Cohen-Or, and A. H. Bermano, "Human motion diffusion model," in *The Eleventh International Conference on Learning Representations*, 2023.

[2] C. Guo, S. Zou, X. Zuo, S. Wang, W. Ji, X. Li, and L. Cheng, "Generating diverse and natural 3d human motions from text," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 5152–5161.

[3] M. Zhang, Z. Cai, L. Pan, F. Hong, X. Guo, L. Yang, and Z. Liu, "Motiondiffuse: Text-driven human motion generation with diffusion model," *arXiv preprint arXiv:2208.15001*, 2022.

[4] C. Guo, Y. Mu, M. G. Javed, S. Wang, and L. Cheng, "Momask: Generative masked modeling of 3d human motions," *ArXiv*, 2023.

[5] X. Chen, B. Jiang, W. Liu, Z. Huang, B. Fu, T. Chen, and G. Yu, "Executing your commands via motion diffusion in latent space," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 18 000–18 010.

[6] K. Aberman, Y. Weng, D. Lischinski, D. Cohen-Or, and B. Chen, "Unpaired motion style transfer from video to animation," *CoRR*, vol. abs/2005.05751, 2020.

[7] D.-K. Jang, S. Park, and S.-H. Lee, "Motion puzzle: Arbitrary motion style transfer by body part," *ACM Trans. Graph.*, vol. 41, no. 3, June 2022.

[8] W. Song, X. Jin, S. Li, C. Chen, A. Hao, X. Hou, N. Li, and H. Qin, "Arbitrary motion style transfer with multi-condition motion latent diffusion model," *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 821–830, 2024.

[9] T. Ao, Z. Zhang, and L. Liu, "Gesturediffuclip: Gesture diffusion model with clip latents," *ACM Trans. Graph.*, 2023.

[10] L. Zhong, Y. Xie, V. Jampani, D. Sun, and H. Jiang, "Smoodi: Stylized motion diffusion model," *ArXiv*, vol. abs/2407.12783, 2024.

[11] K. Karunratanakul, K. Preechakul, S. Suwajanakorn, and S. Tang, "Guided motion diffusion for controllable human motion synthesis," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 2151–2162.

[12] Y. Shafir, G. Tevet, R. Kapon, and A. H. Bermano, "Human motion diffusion as a generative prior," in *The Twelfth International Conference on Learning Representations*, 2024.

[13] Y. Yuan, J. Song, U. Iqbal, A. Vahdat, and J. Kautz, "Physdiff: Physics-guided human motion diffusion model," *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15 964–15 975, 2022.

[14] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning*, 2021, pp. 8748–8763.

[15] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in *IEEE International Conference on Computer Vision (ICCV)*, 2023.

[16] I. Mason, S. Starke, and T. Komura, "Real-time style modelling of human locomotion via feature-wise transformations and local motion phases," *Proceedings of the ACM on Computer Graphics and Interactive Techniques*, vol. 5, no. 1, 2022.

[17] W. Peebles and S. Xie, "Scalable diffusion models with transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4195–4205.

[18] W. Peebles and A. Jain, "Classifier-free diffusion guidance," *arXiv preprint arXiv:2207.12598*, 2022.

[19] I. Loshchilov and F. Hutter, "Fixing weight decay regularization in adam," *ArXiv*, vol. abs/1711.05101, 2017.

[20] OpenAI, "Chatgpt," 2024, [Online]. Available: https://chat.openai.com, Accessed on: Jul. 9, 2025.

[21] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *Advances in neural information processing systems*, vol. 30, 2017.