Towards Detailed Text-to-Motion Synthesis via Basic-to-Advanced Hierarchical Diffusion Model

Zhenyu Xie[1], Yang Wu[2], Xuehao Gao[3], Zhongqian Sun[2], Wei Yang[2], Xiaodan Liang[1,4*]

[1]Shenzhen Campus of Sun Yat-sen University, Shenzhen, China
[2]Tencent AI Lab, Shenzhen, China
[3] Xi'an Jiao Tong University, Xi'an, China
[4] DarkMatter AI Research, Beijing, China
xiezhy6@mail2.sysu.edu.cn, dylan.yangwu@qq.com, gaoxuehao.xjtu@gmail.com, sallensun@tencent.com, willyang@tencent.com, xdliang328@gmail.com

Abstract

Text-guided motion synthesis aims to generate 3D human motion that not only precisely reflects the textual description but reveals the motion details as much as possible. Pioneering methods explore the diffusion model for text-to-motion synthesis and obtain significant superiority. However, these methods conduct diffusion processes either on the raw data distribution or the low-dimensional latent space, which typically suffer from the problem of modality inconsistency or detail-scarce. To tackle this problem, we propose a novel Basic-to-Advanced Hierarchical Diffusion Model, named B2A-HDM, to collaboratively exploit low-dimensional and high-dimensional diffusion models for high quality detailed motion synthesis. Specifically, the basic diffusion model in low-dimensional latent space provides the intermediate denoising result that to be consistent with the textual description, while the advanced diffusion model in high-dimensional latent space focuses on the following detail-enhancing denoising process. Besides, we introduce a multi-denoiser framework for the advanced diffusion model to ease the learning of high-dimensional model and fully explore the generative potential of the diffusion model. Quantitative and qualitative experiment results on two text-to-motion benchmarks (HumanML3D and KIT-ML) demonstrate that B2A-HDM can outperform existing state-of-the-art methods in terms of fidelity, modality consistency, and diversity.

Introduction

Text-to-motion synthesis, which aims to generate human motion that conforms to the textual descriptions (with result examples of our proposed model shown in Figure 1), has made significant progress in recent years. It has the potential to revolutionize the traditional process of acquiring human motion, which typically requires expert knowledge from artists or expensive motion capture equipment.

However, inferring human motion from textual description is a no-trivial task due to the essential discrepancy between the two data modalities. To address this challenge, some existing works (Ahuja and Morency 2019; Ghosh et al. 2021; Tevet et al. 2022; Petrovich, Black, and Varol 2022) resort to the auto-encoder/VAE for motion synthesis and strive to align the cross-modal information in a shared embedding space. On the other hand, motivated by fruitful attempts of diffusion model (Sohl-Dickstein et al. 2015; Ho, Jain, and Abbeel 2020; Nichol and Dhariwal 2021) in the cross-modal image synthesis (Rombach et al. 2022; Saharia et al. 2022; Ramesh et al. 2022), some pioneering works (Tevet et al. 2023; Zhang et al. 2022; Dabral et al. 2023; Ma, Bai, and Zhou 2022; Chen et al. 2023; Jin et al. 2023) exploit the diffusion model for text-to-motion synthesis, demonstrating significant improvements in fidelity and cross-modal consistency.

In spite of the powerful generative ability, training a diffusion model for text-to-motion synthesis remains challenging, which is mainly attributed to the complexity of the data distribution and the insufficiency of the text-annotated training data. Without adequate data, it is difficult for the neural network to learn the denoising process that converts the Gaussian distribution into the complex motion distribution. To address this problem, MLD (Chen et al. 2023) applies VAE to project the raw motion from the initial 3D pose space into the latent code in the low-dimensional latent space, and then conducts diffusion process on the latent space. Although simplifying the target distribution can ease the learning of the denoising process, the low-dimensional latent code is less expressive, which hinders the diffusion model from generating detailed motion. Specifically, as shown in Fig. 2(a), reducing the dimension of the latent space in VAE results in reconstructed motions with fewer captured details. Additionally, Fig. 2(b) reveals that decreasing the dimension of the latent space leads to an increase in the FID scores of the reconstructed motions, indicating a degradation in their quality. However, simply increasing the dimension of the latent space makes the target distribution complex again, leading to more difficulties in network learning, which will further result in a significant drop in performance in terms of cross-modal consistency, as demonstrated in Fig. 2(c). Nevertheless, the comparisons in Fig. 2(c) provide an intuitive insight that while the low-dimensional diffusion model is ineffective for detail generation, it significantly benefits the modality transformation. This insight further inspires us to integrate the complementary advantages of low-dimensional and high-dimensional diffusion models to enhance cross-modal consistency and facilitate detail-rich motion generation.

*Corresponding Author.

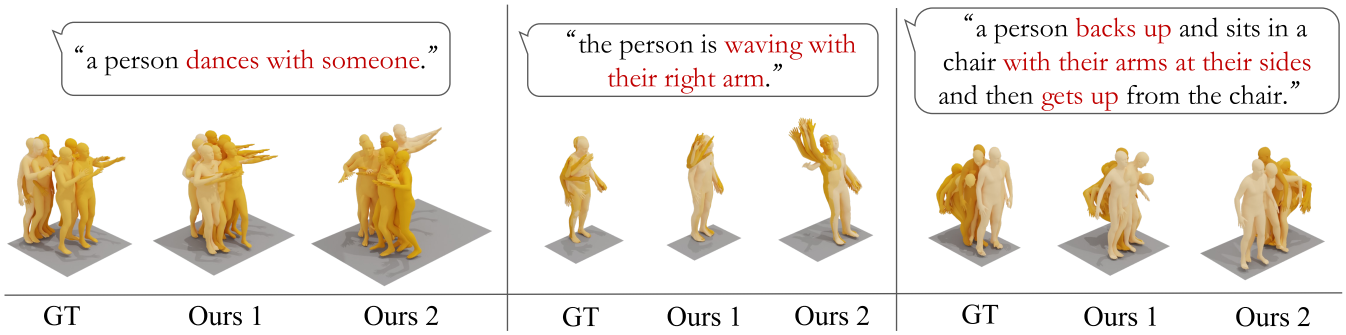

Figure 1: The visual results of our B2A-HDM on HumanML3D (Guo et al. 2022). Our method can generate diverse and high-quality motion sequences that conform to the provided textual descriptions.

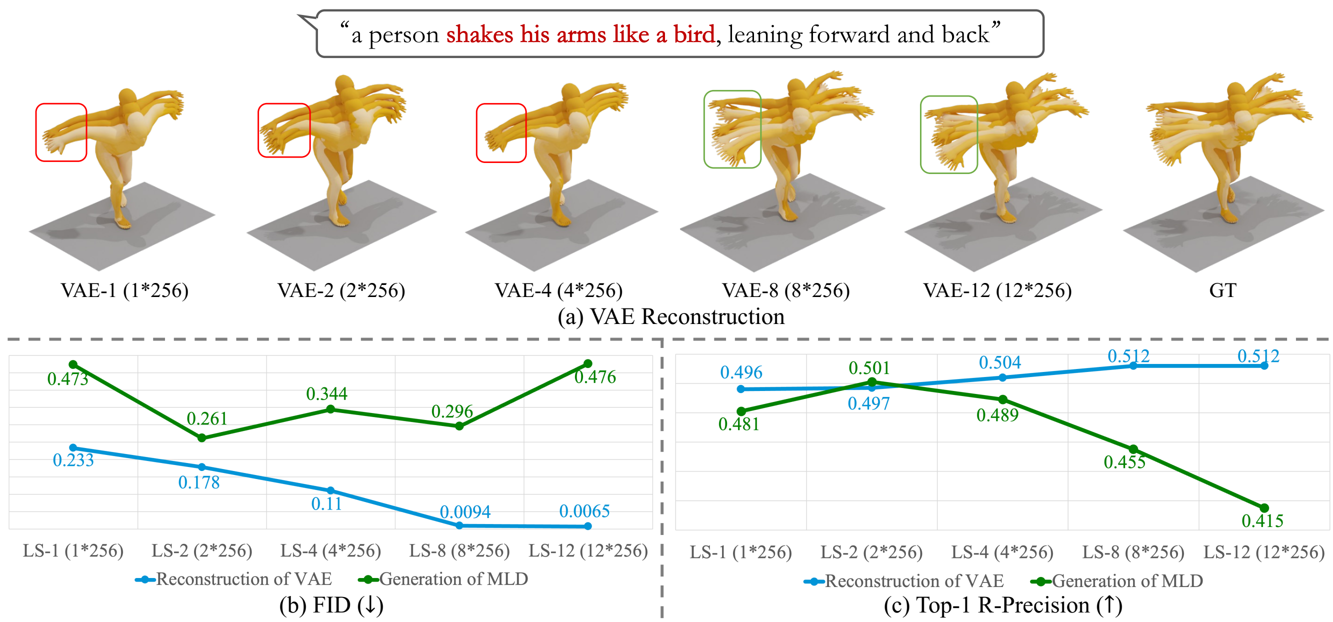

Figure 2: (a) Visual comparisons among the reconstruction results of different VAEs. (b) Comparison of FID scores (lower is better). (c) Comparison of Top-1 R-Precision scores (higher is better).

To this end, we proposed a novel **B**asic-to-**A**dvanced **H**ierarchical **D**iffusion **M**odel, named B2A-HDM, for text-guided motion synthesis, in which the basic diffusion model focuses on consistent but detail-scarce text-to-motion transformation, while the advanced diffusion model aims to conduct a detail-enhancing denoising process based on the intermediate results from the basic model.

Specifically, the Basic Diffusion Model (BDM) is trained in the low-dimensional latent space, in which the data distribution is much simpler than the raw motion distribution, making the learning of text-to-motion transformation easier. However, since the low-dimensional latent space is less expressive, BDM is ineffective to synthesize detail-rich results. On the other hand, the Advanced Diffusion Model (ADM) is trained on the high-dimensional latent space, thus has a larger representation capacity for characterizing more motion details and improving high-fidelity synthesis. However, directly using ADM to conduct the whole text-to-motion denoising process will lead to poor modality consistency. To tackle this problem, B2A-HDM explicitly divides the denoising process into several sub-processes, in which BDM and ADM focus on different denoising stages. To be specific, B2A-HDM first conducts the forward diffusion on the synthesized result from BDM, resulting in the noised motion that will be regarded as the result of the early denoising sub-process. Then, ADM conducts the following denoising process based on the noised motion. Since the noised motion derived from BDM provides a proper initial state (i.e., consistent with the textual description), ADM can focus on the detail-enhancing denoising process. Moreover, to further ease the learning of high-dimensional diffusion model, B2A-HDM exploits the multi-denoisers framework for ADM, in which each denoiser dominates a specific denoising sub-process.

Overall, our contributions can be summarized as follows: (1) We propose a novel Basic-to-Advanced Hierarchical Diffusion Model (B2A-HDM) for text-to-motion synthesis, which jointly incorporates the complementary benefits of low-/high-dimensional diffusion models into detailed motion synthesis. (2) We explicitly divides the denoising process into several sub-processes, which are separately dominated by one basic and two advanced diffusion models. (3) Extensive experiments on two text-to-motion benchmarks (Guo et al. 2022; Plappert, Mandery, and Asfour 2016) show the superiority of B2A-HDM over the existing SOTAs.

Related Work

Conditional Diffusion Models. Diffusion models (Sohl-Dickstein et al. 2015; Ho, Jain, and Abbeel 2020; Nichol and Dhariwal 2021) are a novel class of generative models that have made significant strides in cross-modal synthesis, spanning a diverse range of applications such as text-to-image (Rombach et al. 2022; Saharia et al. 2022; Ramesh et al. 2022), text-to-video (Ho et al. 2022; Esser et al. 2023; Yu et al. 2023), text-to-3d (Poole et al. 2022; Lin et al. 2023), text-to-audio (Popov et al. 2021), among others. Typically, diffusion models consist of the forward and reverse diffusion process, in which the forward process gradually add Gaussian noise into real data to construct training data, while the reverse process involves a neural network to conduct denoising. To adapt diffusion models for conditional generation, Dharival et al. (Dhariwal and Nichol 2019) propose a classifier-guided diffusion model that incorporates conditional information into the reverse diffusion process using additional classifiers. Besides, Ho et al. (Ho and Salimans 2021) propose a classifier-free guidance strategy for conditional diffusion models. This approach strikes a balance between synthesis quality and diversity, which can obtain better results and is widely used by the following works.

Text-Guided Human Motion Synthesis. Following the development of generative models, text-guided motion synthesis has witnessed significant progress in recent years. JL2P (Ahuja and Morency 2019) employs auto-encoder to model a share space for the text and motion embedding, from which the text embedding will be used to reconstruct the corresponding motion during inference. MotionCLIP (Tevet et al. 2022) attempts to improve the auto-encoder's generalization by aligning the shared space with the expressive CLIP (Radford et al. 2021) embedding space, enabling it to handle out-of-distribution motion synthesis. TMEOS (Petrovich, Black, and Varol 2022) and T2M (Guo et al. 2022) use VAE framework to enhance the diversity of the generated results by constraining the share space into a normal distribution. Different from the above encoder-decoder paradigm, T2M-GPT (Zhang et al. 2023) generates motion sequence in an auto-regressive manner by jointly using VQ-VAE (van den Oord, Vinyals, and Kavukcuoglu 2017) and GPT (Radford et al. 2018), which gains improvement in term of fidelity and modality consistency. Building on the great success of diffusion model on image synthesis, some recent works (Tevet et al. 2023; Zhang et al. 2022; Ma, Bai, and Zhou 2022; Chen et al. 2023) explore the generative potential of diffusion for text-to-motion synthesis. However, these methods model the diffusion process either on the raw motion distribution or on a low-dimensional latent space, leading to modality-inconsistent or detail-scarce synthesis. In this paper, we exploit the basic and advanced diffusion models in different latent spaces to conduct the reverse diffusion process, in which basic and advanced models are separately in charge of modality transformation and detail-enhancing denoising process. Note that, while eDiff-I (Balaji et al. 2022) also employs multiple denoisers for reverse diffusion, our method differs in that we we train denoisers on different latent spaces, whereas in eDiff-I various denoisers are all modeled on the same raw data space.

Methodology

Given a textual description $\mathbf{w} = \{w_i\}_{i=1}^{L}$ with L words, text-to-motion synthesis aims to generate the 3D motion $\mathbf{s} = \{s_i\}_{i=1}^{N}$ with N frames that conform to the text input, where $s_i \in \mathbb{R}^J$ denotes a J-dimensional body pose representation at i-th frame. To achieve this, we propose a novel Basic-to-Advanced Hierarchical Diffusion Model (B2A-HDM) to collaboratively exploit the low- and high-dimensional latent diffusion models for modality consistency and detail-rich motion synthesis. A method overview is shown in Fig. 3.

Latent Diffusion for Text-to-Motion Synthesis

Recently, diffusion model (Sohl-Dickstein et al. 2015; Ho, Jain, and Abbeel 2020; Nichol and Dhariwal 2021) has shown its outstanding generative ability for cross-modal synthesis tasks (Rombach et al. 2022; Saharia et al. 2022; Ramesh et al. 2022; Esser et al. 2023), inspiring researchers to explore diffusion model for high-quality text-to-motion synthesis. However, directly modeling the diffusion process on the raw motion distribution typically suffers from inferior synthesis due to the high complexity of raw distribution and the insufficiency of text-annotated training data. To tackle this problem, existing method (Chen et al. 2023) exploits a latent diffusion model to degrade the complexity of target distribution and conduct the diffusion process in low-dimensional latent space, leading to higher quality text-to-motion synthesis.

Specifically, the latent diffusion model is composed of a motion VAE and a diffusion model in VAE latent space. The motion VAE consists of transformer-based encoder $\mathcal{E}$ and decoder $\mathcal{D}$, in which $\mathcal{E}$ is used to encode the raw motion $\mathbf{s} \in \mathbb{R}^{N \times J}$ into latent code $\mathbf{z} \in \mathbb{R}^{K \times D}$ ($K \ll N$) in the latent space, while $\mathcal{D}$ is used to decode sample in latent space back to the real motion. By using the Kullback-Leibler (KL) loss and the Mean Squared Error (MSE) loss for the training procedure, the motion VAE can provide a low-dimensional but representative latent space.

On the other hand, the diffusion model aims to generate a motion latent code in the VAE latent space according to the textual description, which is achieved by a reverse diffusion process that gradually transfers a random noise $\mathbf{n} \in \mathbb{R}^{K \times D}$ from Gaussian distribution to the motion latent code $\mathbf{z}_0 \in \mathbb{R}^{K \times D}$. To train a denoiser ϵ_θ for this reverse process, a forward diffusion process is required to successively add Gaussian noise onto $\mathbf{z}_0$ in a Markov chain manner, which can be formulated as:

$$q\left(\mathbf{z}_{1:T} \mid \mathbf{z}_0\right) := \prod_{t=1}^{T} q\left(\mathbf{z}_t \mid \mathbf{z}_{t-1}\right), \tag{1}$$

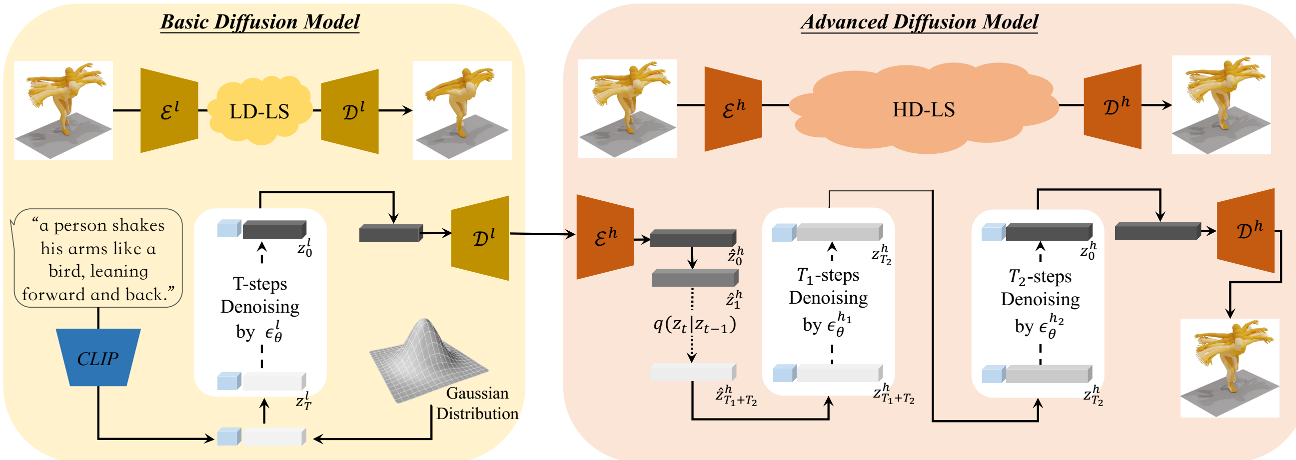

Figure 3: Method Overview. B2A-HDM consists of a Basic Diffusion Model(BDM) and an Advanced Diffusion Model(ADM). BDM comprises a VAE $\{\mathcal{E}_l, \mathcal{D}_l\}$ and a denoiser ϵ_θ^l, in which ϵ_θ^l is in charge of the complete T-steps denoising process in the low-dimensional latent space (LD-LS). ADM comprises a VAE $\{\mathcal{E}_h, \mathcal{D}_h\}$ and two denoisers $\epsilon_\theta^{h_1}$ and $\epsilon_\theta^{h_2}$, in which $\epsilon_\theta^{h_1}$ and $\epsilon_\theta^{h_2}$ are responsible for T_1- and T_2-steps denoising sub-process on the high-dimensional latent space (HD-LS), respectively.

$$q\left(\mathbf{z}_t \mid \mathbf{z}_{t-1}\right) := \mathcal{N}\left(\mathbf{z}_t; \sqrt{1-\beta_t}\mathbf{z}_{t-1}, \beta_t \mathbf{I}\right), \quad (2)$$

where T is the total steps of the forward process and β_t is a hyperparameter of the noising weight. By using the reparameterization trick (Kingma and Welling 2014), we can sample $\mathbf{z}_t$ from $q\left(\mathbf{z}_t \mid \mathbf{z}_0\right)$ at an arbitrary timestep t:

$$\mathbf{z}_t := \sqrt{\bar{\alpha}_t}\mathbf{z}_0 + \epsilon\sqrt{1-\bar{\alpha}_t}, \quad \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \quad (3)$$

where $\bar{\alpha}_t := \prod_{s=1}^t \alpha_s$ and $\alpha_s := 1 - \beta_s$. During training, given the noised data $\mathbf{z}_t$ and the textual description $\mathbf{w}$ as inputs, denoiser ϵ_θ is expected to predict the noise ϵ added at t-th Markov step. The object function for the learning of ϵ_θ only contains the MSE loss:

$$\mathcal{L} := \mathbb{E}_{\epsilon \sim \mathcal{N}(\mathbf{0},\mathbf{I}), t\in[1,T]}\left[\left\|\epsilon - \epsilon_\theta\left(\mathbf{z}_t, \tau_\theta(\mathbf{w}), t\right)\right\|_2^2\right], \quad (4)$$

where τ_θ represents a pre-trained CLIP (Radford et al. 2021) text encoder which is used to extract the text embedding and is frozen during training. Furthermore, denoiser ϵ_θ is trained by using classifier-free guidance (Ho and Salimans 2021). Therefore, during inference, the predicted noise ϵ' is formulated as the linear combination of the conditional and unconditional predictions:

$$\epsilon' := \epsilon_\theta\left(\mathbf{z}_t, \varnothing, t\right) + g \cdot \left(\epsilon_\theta\left(\mathbf{z}_t, \tau_\theta(\mathbf{w}), t\right) - \epsilon_\theta\left(\mathbf{z}_t, \varnothing, t\right)\right), \quad (5)$$

where $\varnothing$ represents a null-text input and g is the hyperparameter of guidance scale.

B2A-HDM

Although using latent diffusion model can ease the learning of diffusion network, the low-dimensional latent space may be under representative (i.e., as shown in Fig 2(a)) and thus constrains the generative upper bound of diffusion model, leading to detail-scarce motion synthesis. Directly increasing the dimension of the VAE latent space will make target distribution complex again and cause a significant performance drop in modality consistency (i.e., as illustrated in Fig 2(c)). To boost the representation capacity of latent motion embedding without degrading the cross-modal mapping consistency, our B2A-HDM employs Basic Diffusion Model (BDM) to provide modality consistent generated results, which will be further processed by Advanced Diffusion Model (ADM) for detail-enhancing synthesis.

To be specific, our BDM and ADM are defined in the low-dimensional and high-dimensional latent space, respectively. To obtain BDM, a motion VAE $\mathcal{V}^l = \{\mathcal{E}^l, \mathcal{D}^l\}$ is trained on the raw motion distribution to obtain a low-dimensional latent space $\mathcal{W}^l$ (with latent code $\mathbf{z}^l \in \mathbb{R}^{K_1 \times D}$). Then, a denoiser ϵ_θ^l on $\mathcal{W}^l$ is trained to handle arbitrary t-th denoising step ($t \in [1, T]$), which will be used to conduct the reverse diffusion process from random noise $\mathbf{n}^l \in \mathbb{R}^{K_1 \times D}$ during inference. For ADM, motion VAE $\mathcal{V}^h = \{\mathcal{E}^h, \mathcal{D}^h\}$ is also required to obtain a high-dimensional latent space $\mathcal{W}^h$ (with latent code $\mathbf{z}^h \in \mathbb{R}^{K_2 \times D}$, $K_2 > K_1$). However, directly training a denoiser ϵ_θ^h on $\mathcal{W}^h$ for arbitrary timestep t is non-trivial and typically results in model degradation.

To tackle this problem, B2A-HDM improves the common reverse diffusion process in two-folds. First, instead of using a single denoiser ϵ_θ^h to conduct the whole reverse diffusion process, B2A-HDM applies BDM to provide the early T^l steps denoising result for ADM, and ADM is only required to conduct the following $T - T^l$ denoising steps. Since BDM is trained on $\mathcal{W}^l$ with lower distribution complexity and performs better in modality consistent synthesis, it can provide an intermediate denoising result with proper modality information. Therefore ADM is designed for the following detail-enhancing denoising process. Furthermore, using a single denoiser ϵ_θ^h for the remaining $T - T^l$ steps denoising process still remains challenging due to the high distribution complexity of $\mathcal{W}^h$ and the significant discrepancy of $\mathbf{z}_t^h$ in various timestep. To further ease the learning of denoiser on $\mathcal{W}^h$, our B2A-HDM assigns k denoisers for ADM, in which

Algorithm 1: Reverse Diffusion Process of B2A-HDM

Require: A textual description $\mathbf{w}$, a random seed r.
Ensure: A motion sequence $\mathbf{s}$.

```
z^l_T ~ N(0, I), ε ~ N(0, I) unit Gaussian random variables with random seed r;
T^h ← T − T^l denoising steps for ADM;
for t = T, T − 1, ..., 1 do
    z^l_{t−1} ← ε^l_θ(z^l_t, τ_θ(w), t);
end for
s^l ← D^l(z^l_0);   ẑ^h_0 ← E^h(s^l);   z^h_{T^h} ← sqrt(ᾱ_{T^h}) ẑ^h_0 + ε sqrt(1 − ᾱ_{T^h});
for t = T^h, T^h − 1, ..., 1 do
    if t > T^h/2 then
        z^h_{t−1} ← ε^{h_1}_θ(z^h_t, τ_θ(w), t);
    else
        z^h_{t−1} ← ε^{h_2}_θ(z^h_t, τ_θ(w), t);
    end if
end for
s ← D^h(z^h_0);
return s
```

each denoiser $\epsilon_\theta^{h_k}$ is only in charge of the specific interval of the denoising process in the high-dimensional latent space.

During inference, the T^l steps denoising result $\mathbf{z}^l_{T^l}$ from BDM can not used by ADM due to dimension inconsistency between $\mathbf{z}^l_{T^l}$ and $\mathbf{z}^h$ in $\mathcal{W}^h$. To address this problem, B2A-HDM first employs BDM to generate a motion sequence $\mathbf{s}^l$, which will be transferred into a latent code $\hat{\mathbf{z}}^h_0$ in $\mathcal{W}^h$. Then, B2A-HDM conducts $T-T^l$ steps forward diffusion to obtained intermediate denoising result for ADM.

We take $k=2$ as an example and show the complete reverse diffusion process in Algorithm 1.

Training Details and Objective Functions

The training procedure for BDM and ADM are similar to that for the common latent diffusion model(Chen et al. 2023; Rombach et al. 2022), except for the training of the diffusion networks for ADM. Specifically, since denoisers in ADM are in charge of different denoising sub-processes, each denoiser is assigned a specific timestep interval during training. Note that although each denoiser is independent of the others, we train all of them in a single training procedure, which enables us to conduct the complete reverse process during training and observe the performance change on the evaluation set.

During training VAE for BDM and ADM, the objective functions can be formulated as:

$$\mathcal{L}_{vae} = \lambda_{kl}\mathcal{L}_{kl} + \lambda^{vae}_{mse}\mathcal{L}^{vae}_{mse}, \tag{6}$$

where λ_{kl} and λ^{vae}_{mse} are the trade-off hyperparameters and are set to 1e-4 and 1.0, respectively. When training the denoiser for BDM and ADM, only MSE loss $\mathcal{L}^{\epsilon_\theta}_{mse}$ in Equation 4 is used. In addition, we introduce a timestep-aware loss weight $\lambda(t)$ for the MSE loss in BDM to increase the penalty for early denoising process learning. The timestep-aware MSE loss can be formulated as:

$$\mathcal{L}^t_{mse} = \lambda(t)\mathcal{L}^{\epsilon}_{mse}, \quad \lambda(t) = (1-\bar{\alpha}_t) * w_1 + w_2, \tag{7}$$

where $\bar{\alpha}_t$ is the diffusion parameter defined in Equation 3, w_1 and w_2 are used to rescale the loss weight to a specific interval (i.e., $\lambda(t) \in [0.5, 5]$) and are set to 4.5 and 0.5, respectively. In Sec. , we will analyse the impact of using timestep-aware MSE loss for BDM, which will demonstrate improved performance in high-dimensional scenarios.

Experiments

Datasets. Our experiments are conducted on two publicly available benchmarks for text-to-motion synthesis, namely KIT-ML (Plappert, Mandery, and Asfour 2016) and HumanML3D (Guo et al. 2022). Specifically, KIT-ML consists of 3,911 motion sequences with 12.5 FPS and 6,278 language annotations. HumanML3D contains 14,616 motion sequences with 20FPS and 44,970 textual descriptions. Regarding the data format, we follow (Guo et al. 2022) to use the redundant representation for each pose frame, which is composed of the local/global joint velocities, joint positions, joint rotations, and the foot contact binary labels.

Baselines. We perform quantitative and qualitative comparisons with four most advanced text-to-motion synthesis methods, including MDM (Tevet et al. 2023), MotionDiffuse (Zhang et al. 2022), MLD (Chen et al. 2023), and T2M-GPT (Zhang et al. 2023). For these methods, we employ the official pre-trained models and strictly adhere to the official instructions to conduct text-to-motion synthesis.

Evaluation Metrics. We employ five evaluation metrics originated from (Guo et al. 2022) to assess the performance of different methods. Specifically, FID (Heusel et al. 2017) measures the distribution difference between the generated and real motion, which is widely used for realism evaluation. R-Precision and MM-Dist are designed to evaluate modality consistency, in which R-Precision calculates the Top-1/2/3 accuracy for the motion-to-text retrieval while MM-Dist calculates the Euclidean distances between the generated motion and its corresponding text. Diversity and MModality are designed for diversity evaluation, which are used to measure the variance of all generated motions and the variance of the particular generations for each text input, respectively.

Implementation Details. The dimension of the latent space for BDM and ADM are 4×256 and 8×256, respectively. ADM is equipped with 2 denoisers. In spite of using different latent space, BDM and ADM share the same network architecture for the motion VAE encoder, decoder, and the diffusion denoiser. In line with MLD (Chen et al. 2023), all of the three modules are composed of 9 transformer layers with skip connection. Our B2A-HDM is implemented using PyTorch (Paszke et al. 2019) and both of motion VAE and diffusion denoiser are trained on 4 Tesla V100 GPUs. During training, for both HumanML3D (Guo et al. 2022) and KIT-ML (Plappert, Mandery, and Asfour 2016) dataset, the batch size on each GPU is set to 96 and the all modules are trained by using AdamW (Loshchilov and Hutter 2019) optimizer with a fixed learning rate 1e-4. For HumanML3D, both VAE and denoiser are trained for 6,000 epochs, while for KIT-ML, the VAE and denoiser are trained for 25,000 epochs and 2,500 epochs, respectively.

	Method	R-Precision ↑			FID ↓	MM-Dist ↓	Diversity →	MModality ↑
		Top-1	Top-2	Top-3				
	Real motion	$0.511^{\pm.003}$	$0.703^{\pm.003}$	$0.797^{\pm.002}$	$0.002^{\pm.000}$	$2.974^{\pm.008}$	$9.503^{\pm.065}$	-
(a)	MDM (Tevet et al. 2023)	$0.320^{\pm.005}$	$0.498^{\pm.004}$	$0.611^{\pm.007}$	$0.544^{\pm.044}$	$5.566^{\pm.027}$	$9.559^{\pm.086}$	$\mathbf{2.799^{\pm.072}}$
	MotionDiffuse (Zhang et al. 2022)	$0.491^{\pm.001}$	$0.681^{\pm.001}$	$0.782^{\pm.001}$	$0.630^{\pm.001}$	$3.113^{\pm.001}$	$9.410^{\pm.049}$	$1.553^{\pm.042}$
	MLD (Chen et al. 2023)	$0.481^{\pm.003}$	$0.673^{\pm.003}$	$0.772^{\pm.002}$	$0.473^{\pm.013}$	$3.196^{\pm.010}$	$9.724^{\pm.082}$	$2.413^{\pm.079}$
	T2M-GPT (Zhang et al. 2023)	$0.491^{\pm.003}$	$0.680^{\pm.003}$	$0.775^{\pm.002}$	$0.116^{\pm.004}$	$3.118^{\pm.011}$	$9.761^{\pm.081}$	$1.856^{\pm.011}$
	B2A-HDM (Ours)	$\mathbf{0.511^{\pm.002}}$	$\mathbf{0.699^{\pm.002}}$	$\mathbf{0.791^{\pm.002}}$	$\mathbf{0.084^{\pm.004}}$	$\mathbf{3.020^{\pm.010}}$	$\mathbf{9.526^{\pm.080}}$	$1.914^{\pm.078}$
	Real motion	$0.424^{\pm.005}$	$0.649^{\pm.006}$	$0.779^{\pm.006}$	$0.031^{\pm.004}$	$2.788^{\pm.012}$	$11.08^{\pm.097}$	-
(b)	MDM (Tevet et al. 2023)	$0.164^{\pm.004}$	$0.291^{\pm.004}$	$0.396^{\pm.004}$	$0.497^{\pm.021}$	$9.191^{\pm.022}$	$10.847^{\pm.109}$	$1.907^{\pm.214}$
	MotionDiffuse (Zhang et al. 2022)	$0.417^{\pm.004}$	$0.621^{\pm.004}$	$0.739^{\pm.004}$	$1.954^{\pm.062}$	$2.958^{\pm.005}$	$\mathbf{11.10^{\pm.143}}$	$0.730^{\pm.013}$
	MLD (Chen et al. 2023)	$0.390^{\pm.008}$	$0.609^{\pm.008}$	$0.734^{\pm.007}$	$0.404^{\pm.027}$	$3.204^{\pm.027}$	$10.80^{\pm.117}$	$2.192^{\pm.071}$
	T2M-GPT (Zhang et al. 2023)	$0.416^{\pm.006}$	$0.627^{\pm.006}$	$0.745^{\pm.006}$	$0.514^{\pm.029}$	$3.007^{\pm.023}$	$10.921^{\pm.108}$	$1.570^{\pm.039}$
	B2A-HDM (Ours)	$\mathbf{0.436^{\pm.006}}$	$\mathbf{0.653^{\pm.006}}$	$\mathbf{0.773^{\pm.005}}$	$\mathbf{0.367^{\pm.020}}$	$\mathbf{2.946^{\pm.024}}$	$10.86^{\pm.124}$	$1.291^{\pm.047}$

Table 1: Quantitative results on (a) HumanML3D (Guo et al. 2022) and (b) KIT-ML (Plappert, Mandery, and Asfour 2016).

Quantitative Results

The quantitative comparison of our B2A-HDM against the existing state-of-the-art methods on HumanML3D (Guo et al. 2022) and KIT-ML (Plappert, Mandery, and Asfour 2016) datasets are reported in Tab. 1 (a) and (b), respectively. Both tables demonstrate that B2A-HDM outperforms other methods in terms of modality consistency and fidelity. Specifically, B2A-HDM achieves the highest R-Precision (Top-1/2/3) and the lowest MM-Dist scores on both datasets, indicating that the generated results of B2A-HDM are more consistent with the input text than those of other methods. Moreover, B2A-HDM obtains the lowest FID score on both datasets, highlighting its superiority over other methods in realistic synthesis. It's worth noting that B2A-HDM is the only approach that consistently improves on the above three metrics, which further validates the effectiveness of the combination of basic and advanced diffusion models. Additionally, B2A-HDM achieves comparable Diversity and Modality scores on both datasets, demonstrating its ability for diverse generation.

Qualitative Results

Fig. 4 shows a qualitative comparison of B2A-HDM against the existing SOTA methods on HumanML3D (Guo et al. 2022) dataset. The visual comparison illustrates B2A-HDM outperforms other methods in generating motion sequences with better modality consistency and detail preservation. For instance, in the first row of Fig. 4, MotionDiffuse (Zhang et al. 2022) and MDM (Tevet et al. 2023) fail to generate motion that coheres with the input text, while T2M-GPT (Zhang et al. 2023) and MLD (Chen et al. 2023) tend to overlook details of the hands. In contrast, B2A-HDM generates motion sequences that conform to the input text and capture the fine-grained motion details in the hand region.

Method	BD No.	AD No.	LD-LS Dim	HD-LS Dim	R-P Top-1 ↑	FID ↓
BDM-4	1	0	4×256	-	0.505	0.284
ADM-8	0	1	-	8×256	0.481	0.171
B2A-HDM⋆	3	0	4×256	-	0.508	0.220
B2A-HDM∗	0	3	-	8×256	0.490	0.225
B2A-HDM	1	2	4×256	8×256	0.511	0.084

Table 2: Quantitative results of the ablation study with different configurations, in which BD/AD No., LD/HD-LS Dim refer to basic/advanced denoiser number and low/high-dimension latent space dimension, respectively

Ablation Study

Impact of the timestep-aware MSE loss. As shown in Fig. 5, when the dimension of the latent space is higher than 2×256, using the timestep-aware MSE loss consistently enhances the performance of diffusion model in terms of lower FID and higher Top-1 R-Precision, which highlights the ability of timestep-aware MSE loss to facilitate the learning of denoisers in high-dimensional latent spaces.

Effectiveness of B2A-HDM. We compare B2A-HDM with BDM in a 4×256 latent space (BDM-4) and ADM in an 8×256 latent space (ADM-8). Additionally, we compare B2A-HDM with two variants (i.e., B2A-HDM⋆ and B2A-HDM∗) that separately include three denoisers in low-dimemsion and high-dimension latent space to demonstrate the effectiveness of combining basic and advanced diffusion models. As reported in Tab. 2, directly using BDM-4 or ADM-8 leads to higher FID or lower R-Precision. Although increasing the denoiser number enables B2A-HDM⋆ and

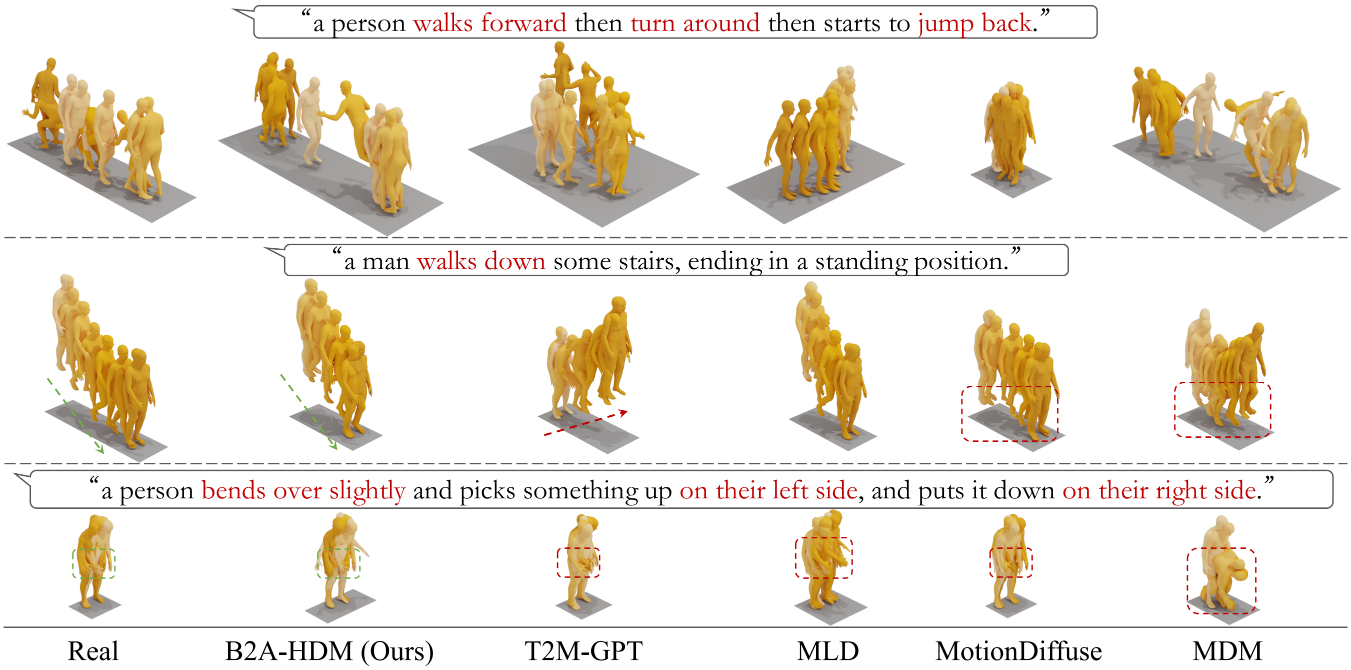

Figure 4: Qualitative comparisons on HumanML3D dataset (Guo et al. 2022). The flow of time is represented by colors, with lighter shades indicating the past. Please zoom in for more details.

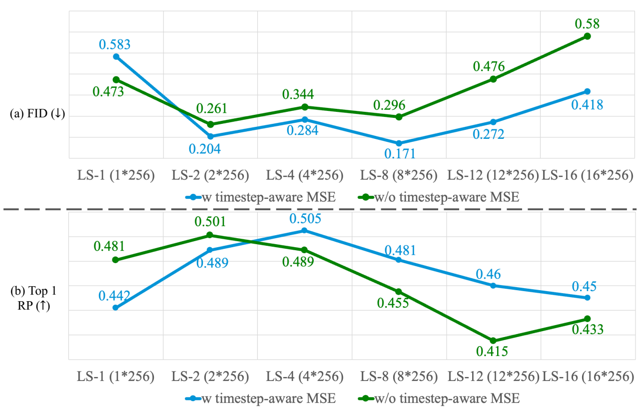

Figure 5: Impact of the timestep-aware MSE loss for BDMs in different latent space (LS).

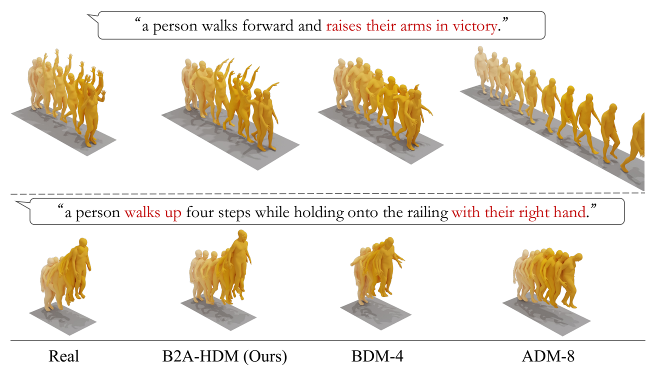

Figure 6: Ablation Study on the effectiveness of B2A-HDM.

B2A-HDM∗ to gain performance improvement against their single-denoiser counterparts (i.e., BDM-4 and ADM-8), they still fall short of B2A-HDM. By collaborativly using basic and advanced diffusion models, B2A-HDM achieves best FID and R-Precision scores, which demonstrate the necessity and effectiveness of combining basic and advanced denoisers. Moreover, Fig. 6 shows that ADM-8 is prone to generate modality inconsistent motions while BDM-4 tends to ignore the motion details. In contrast, our B2A-HDM performs better in modality transformaion and detail preservation, which further validates the effectiveness of our method.

Conclusion

We propose a novel Basic-to-Advanced Hierarchical Diffusion Model (B2A-HDM) for text-to-motion synthesis. B2A-HDM comprises a basic diffusion model (BDM) in low-dimensional latent space and a advanced diffusion model (ADM) with two denoisers in high-dimensional latent space, in which BDM are in charge of the modality-consistent denoising, whereas ADM is responsible for the following detail-enhancing denoising. In this way, B2A-HDM can fully leverage the generative potential of diffusion models to produce high-quality motion sequences that conform to the provided textual descriptions. Extensive experiments on two public text-to-motion benchmarks demonstrate the superiority of B2A-HDM over existing state-of-the-art methods, while ablation studies further validate the effectiveness of our approach.

Acknowledgments

This work was supported in part by National Key R&D Program of China under Grant No.2020AAA0109700, Guangdong Outstanding Youth Fund (Grant No.2021B1515020061), National Natural Science Foundation of China (NSFC) under Grant No.61976233 and No.92270122, Mobility Grant Award under Grant No.M-0461, Shenzhen Science and Technology Program (Grant No.RCYX20200714114642083), Shenzhen Science and Technology Program (Grant No.GJHZ20220913142600001), Nansha Key RD Program under Grant No.2022ZD014 and Sun Yat-sen University under Grant No.22lgqb38 and 76160-12220011.

References

Ahuja, C.; and Morency, L.-P. 2019. Language2pose: Natural Language Grounded Pose Forecasting. In *International Conference on 3D Vision (3DV)*, 719–728.

Balaji, Y.; Nah, S.; Huang, X.; Vahdat, A.; Song, J.; Zhang, Q.; Kreis, K.; Aittala, M.; Aila, T.; Laine, S.; Catanzaro, B.; Karras, T.; and Liu, M.-Y. 2022. eDiff-I: Text-to-Image Diffusion Models with Ensemble of Expert Denoisers. *arXiv preprint arXiv:2211.01324*.

Chen, X.; Biao, J.; Wen, L.; Zilong, H.; Bin, F.; Tao, C.; Jingyi, Y.; and Gang, Y. 2023. Executing your Commands via Motion Diffusion in Latent Space. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Dabral, R.; Mughal, M. H.; Golyanik, V.; and Theobalt, C. 2023. MoFusion: A Framework for Denoising-Diffusion-based Motion Synthesis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Dhariwal, P.; and Nichol, A. 2019. Diffusion models beat gans on image synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 8780–8794.

Esser, P.; Chiu, J.; Atighehchian, P.; Granskog, J.; and Germanidis, A. 2023. Structure and Content-Guided Video Synthesis with Diffusion Models. *arXiv preprint arXiv:2302.03011*.

Ghosh, A.; Cheema, N.; Oguz, C.; Theobalt, C.; and Slusallek, P. 2021. Synthesis of Compositional Animations from Textual Descriptions. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 1396–1406.

Guo, C.; Zou, S.; Zuo, X.; Wang, S.; Ji, W.; Li, X.; and Cheng, L. 2022. Generating Diverse and Natural 3d Human Motions from Text. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5152–5161.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Ho, J.; Chan, W.; Saharia, C.; Whang, J.; Gao, R.; Gritsenko, A.; Kingma, D. P.; Poole, B.; Norouzi, M.; Fleet, D. J.; and Salimans, T. 2022. Imagen Video: High Definition Video Generation with Diffusion Models. *arXiv preprint arXiv:2210.02303*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising Diffusion Probabilistic Models. *arXiv preprint arxiv:2006.11239*.

Ho, J.; and Salimans, T. 2021. Classifier-Free Diffusion Guidance. In *Deep Generative Models and Downstream Applications at NeurIPS (NeurIPSW)*.

Jin, P.; Wu, Y.; Fan, Y.; Sun, Z.; Wei, Y.; and Yuan, L. 2023. Act As You Wish: Fine-Grained Control of Motion Diffusion Model with Hierarchical Semantic Graphs. In *NeurIPS*.

Kingma, D. P.; and Welling, M. 2014. Auto-Encoding Variational Bayes. In *International Conference on Learning Representations (ICLR)*.

Lin, C.-H.; Gao, J.; Tang, L.; Takikawa, T.; Zeng, X.; Huang, X.; Kreis, K.; Fidler, S.; Liu, M.-Y.; and Lin, T.-Y. 2023. Magic3D: High-Resolution Text-to-3D Content Creation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Loshchilov, I.; and Hutter, F. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*.

Ma, J.; Bai, S.; and Zhou, C. 2022. Pretrained Diffusion Models for Unified Human Motion Synthesis. *arXiv preprint arXiv:2212.02837*.

Nichol, A.; and Dhariwal, P. 2021. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning (ICML)*, 8162–8171.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; Desmaison, A.; Köpf, A.; Yang, E.; DeVito, Z.; Raison, M.; Tejani, A.; Chilamkurthy, S.; Steiner, B.; Fang, L.; Bai, J.; and Chintala, S. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *arXiv preprint arXiv:1912.01703v1*.

Petrovich, M.; Black, M. J.; and Varol, G. 2022. TEMOS: Generating diverse human motions from textual descriptions. In *European Conference on Computer Vision (ECCV)*.

Plappert, M.; Mandery, C.; and Asfour, T. 2016. The KIT Motion-Language Dataset. *Big Data*, 4(4): 236–252.

Poole, B.; Jain, A.; Barron, J. T.; and Mildenhall, B. 2022. DreamFusion: Text-to-3D using 2D Diffusion. *arXiv preprint arXiv:2209.14988*.

Popov, V.; Vovk, I.; Gogoryan, V.; Sadekova, T.; and Kudinov, M. A. 2021. Grad-tts: A diffusion probabilistic model for text-to-speech. In *International Conference on Machine Learning (ICML)*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models from Natural Language Supervision. In *International Conference on Machine Learning (ICML)*, 8748–8763.

Radford, A.; Narasimhan, K.; Salimans, T.; and Sutskever, I. 2018. Improving language understanding by generative pre-training. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-Resolution Image Synthesis with Latent

Diffusion Models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10684–10695.

Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E.; Ghasemipour, S. K. S.; Ayan, B. K.; Mahdavi, S. S.; Lopes, R. G.; Salimans, T.; Ho, J.; Fleet, D. J.; and Norouzi, M. 2022. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. *arXiv preprint arXiv:2205.11487*.

Sohl-Dickstein, J.; Weiss, E.; Maheswaranathan, N.; and Ganguli, S. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning (ICML)*, 2256–2265.

Tevet, G.; Gordon, B.; Hertz, A.; Bermano, A. H.; and Cohen-Or, D. 2022. Motionclip: Exposing human motion generation to clip space. In *European Conference on Computer Vision (ECCV)*, 358–374.

Tevet, G.; Raab, S.; Gordon, B.; Shafir, Y.; Bermano, A. H.; and Cohen-Or, D. 2023. Human Motion Diffusion Model. In *International Conference on Learning Representations (ICLR)*.

van den Oord, A.; Vinyals, O.; and Kavukcuoglu, K. 2017. Neural discrete representation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Yu, S.; Sohn, K.; Kim, S.; and Shin, J. 2023. Video Probabilistic Diffusion Models in Projected Latent Space. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhang, J.; Zhang, Y.; Cun, X.; Huang, S.; Zhang, Y.; Zhao, H.; Lu, H.; and Shen, X. 2023. T2M-GPT: Generating Human Motion from Textual Descriptions with Discrete Representations. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhang, M.; Cai, Z.; Pan, L.; Hong, F.; Guo, X.; Yang, L.; and Liu, Z. 2022. MotionDiffuse: Text-Driven Human Motion Generation with Diffusion Model. *arXiv preprint arXiv:2208.15001*.