LS-GAN: Human Motion Synthesis with Latent-space GANs

Avinash Amballa Gayathri Akkinapalli Vinitra Muralikrishnan
University of Massachusetts Amherst, USA
{aamballa, gakkinapalli, vmuralikrish}@umass.edu

Abstract

*Human motion synthesis conditioned on textual input has gained significant attention in recent years due to its potential applications in various domains such as gaming, film production, and virtual reality. Conditioned Motion synthesis takes a text input and outputs a 3D motion corresponding to the text. While previous works have explored motion synthesis using raw motion data and latent space representations with diffusion models, these approaches often suffer from high training and inference times. In this paper, we introduce a novel framework that utilizes Generative Adversarial Networks (GANs) in the latent space to enable faster training and inference while achieving results comparable to those of the state-of-the-art diffusion methods. We perform experiments on the HumanML3D, HumanAct12 benchmarks and demonstrate that a remarkably simple GAN in the latent space achieves a **FID of 0.482** with more than **91%** in FLOPs reduction compared to latent diffusion model. Our work opens up new possibilities for efficient and high-quality motion synthesis using latent space GANs. Code available at https://github.com/AmballaAvinash/motion-latent-diffusion*

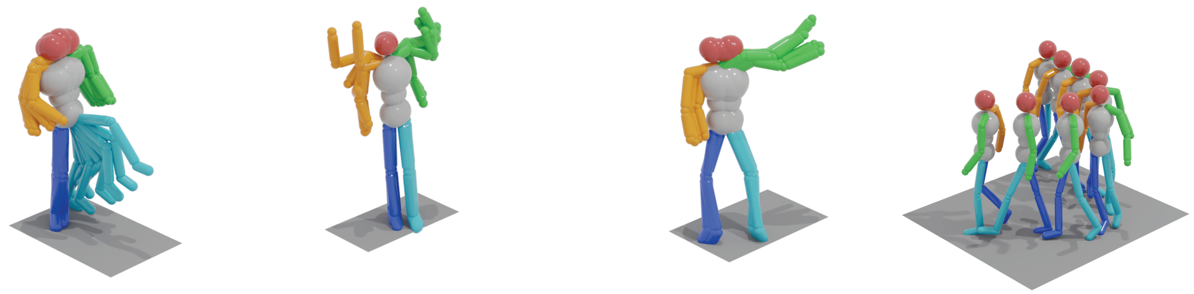
1. Introduction

Human motion synthesis has recently seen rapid advancements in a multi-modal generative fashion, fueled by various conditional inputs such as music [8, 25, 27], action categories [16, 33], and notably, natural language descriptions [2, 12, 13, 23, 34, 49]. This field significantly enhances industries like gaming, film production, and virtual/augmented reality, with text-based conditioning standing out for its convenience and interpretability. However, learning a probabilistic mapping function from textural descriptors to motion sequences is challenging [47] and this mapping often leads to misalignments and high computational demands due to stark differences in distributions between language descriptors and motion sequences, making the task of probabilistic mapping complex.

Conditional diffusion models [23, 49, 55] address this problem by learning a more powerful probabilistic function from the textual descriptors to motion sequences. However, diffusion models in raw sequential data require computational overhead in both traning and inference. To overcome this, motion latent diffusion (MLD) [7] address these issues by encoding motion in a latent space using a Variational Autoencoder (VAE). However, MLD relied on computationally intensive diffusion processes to achieve high-quality image sampling, especially during the training and inference phases.

To efficiently model the motion synthesis, we propose substituting the diffusion model [18] in the latent space with a Generative Adversarial Network (GAN) [10] to capitalize on its efficient adversarial training dynamics. Recognizing the effectiveness of GANs in learning complex representations across diverse modalities [22, 43], and their efficiency in training and inference compared to diffusion models, we propose to utilize them within this latent space. By leveraging GANs, we aim to accelerate the mapping between text embeddings and latent space, thus producing higher-quality motion sequences more efficiently.

Specifically, this work undertakes the task of text-to-motion and action-to-motion synthesis using conditional Generative Adversarial Networks [32] in latent space, as depicted in the accompanying figure:1. We employ a Variational Autoencoder (VAE) to transition from motion space to latent space and utilize pre-trained CLIP models from MLD [7] to condition on textual input. We experiment with various GAN architectures, including vanilla GAN, deep GAN, with loss functions such as cross-entropy and Wasserstein [11] to optimize performance and fidelity in generated motion sequences. Our experiment results on HumanML3D [14] benchmark suggest that a simple GAN architecture achieves an FID of 0.482 with 91% in FLOPS reduction compared to MLD. In addition, our method shows competitive performance on action-to-motion HumanAct12 [16] benchmark. This strategic shift of GANs in latent space not only addresses the computational inefficiencies associated with previous diffusion-based models but also leverages the rapid generative capabilities of GANs to enhance the quality and diversity of motion synthesis, suitable for



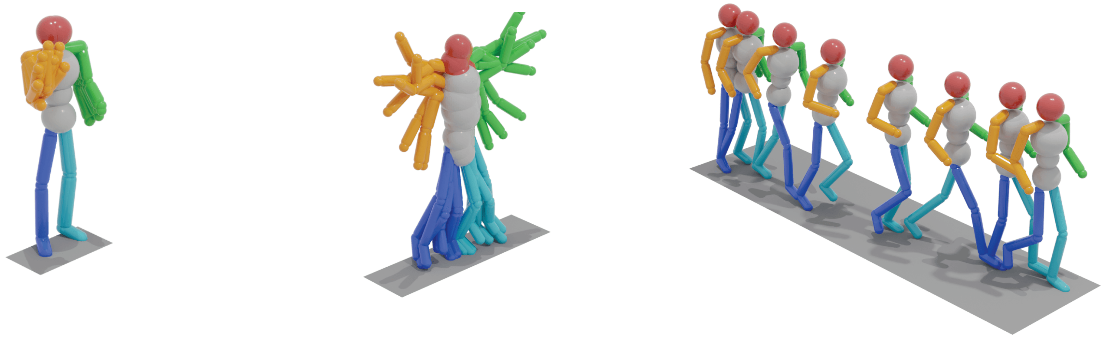
a. a man kicks with something or someone with his left leg. (Vanilla GAN)

b. a person raised arms up and pull them down. (Deep GAN)

c. a person lifts up their left arm to the side. (Vanilla WGAN-GP)

d. a person walks in a circle to their right. (Deep WGAN-GP)

Figure 1. Qualitative results of text-to-motion shown by LS-GAN



a. a person drinks water.

b. a person doing jumping jacks.

c. a person jogs straight forward.

Figure 2. Qualitative results of text-to-motion shown by our best model (Deep WGAN-GP)

real-time applications.

2. Related work.

Motion Synthesis is broadly categorized into conditional and unconditional motion synthesis. Unconditional motion synthesis models the entire motion space without requiring specific annotations, is discussed by Raab et al. [38] in an unsupervised setting using unstructured and unlabeled datasets. Conditional motion synthesis, on the other hand, employs inputs from various modalities such as music [28] and text [24] to generate motion sequences. Text-to-motion synthesis, in particular, has become a dominant area of research due to the user-friendly nature of natural language interfaces. Additional recent advancements in the field include the development of joint-latent models like TEMOS [35] and conditional diffusion models [24, 50, 55], which have led to significant progress. TEMOS, uses a VAE architecture to create a shared latent space for motion and text based on a Gaussian distribution.

Motion diffuse [55] is the first text-based motion diffusion model with fine-grained instructions on body parts. MDM [48] proposes a motion diffusion model on raw motion data to learn the relation between motion and input conditions. Our work closely relates to the Motion Latent Diffusion (MLD) model [7] which utilizes a Variational Autoencoder (VAE) to encode human motion sequences into a low-dimensional latent space and decode them back to motion sequences. The MLD model then employs diffusion processes in this latent space, inspired by other latent diffusion models [41]. To condition the motion sequences on specific inputs like text or actions, the model utilizes CLIP encodings [39], demonstrating robust performance on tasks such as text-to-motion and action-to-motion.

Moreover, approaches like MotionGPT [19] integrates language modeling for both motion and text, treating human motion as a distinct language to construct a generalized model capable of executing various motion tasks through VQ-VAE [51]. T2M-GPT [54] uses a standard 1D convolutional network to map motion sequences to discrete code indices, followed by standard GPT-like model is learned to generate sequences of code indices from pre-trained text embedding. The use of GAN networks for motion synthesis has been done in Ganimator [26] but uses an additional motion sequence as conditional input. On the other hand, Shiobara et al. [46] train Wasserstein GAN directly on the raw motion sequences. Actformer [52] proposed a GAN-based Transformer to generate motion sequence from actions. In addition, Text2Action [1] proposed a generative model which learns the relationship between language and human action in order to generate a human action sequence.

3. Method

While diffusion models have shown tremendous promise and exhibit state-of-the-art performance they are expensive to train, requiring a huge corpus of data. The use of latent space in MLD [7] opens up avenues for other architectures such as GANs to also leverage it. Specifically, given an input condition c describing a motion, our Latent space GAN (LS-GAN) aims to generate a human motion $\hat{x}^{1:L}$ where L represents the motion length.

3.1. VAE and CLIP

Our VAE architecture is borrowed from the MLD [7], which uses transformer model as Encoder $\mathcal{E}$ and Decoder $\mathcal{D}$ with skip connections. The motion encoder $\mathcal{E}$ encodes the motion sequences, $x^{1:L}$ into a latent $z = \mathcal{E}(x^{1:L})$, and the decode z into the motion sequences using the decoder $\mathcal{D}$, i.e., $\hat{x}^{1:L} = \mathcal{D}(z) = \mathcal{D}(\mathcal{E}(x^{1:L}))$. VAE is trained in a similar fashion as MLD with the MSE and KL divergence loss. After training, the VAE is kept fixed. We use pretrained CLIP-ViT-L-14 [40] text encoder to map text prompt. On the other hand to condition on action, we use the learnable embedding for each action category.

3.2. Latent space GAN

We chose GANs for 3 reasons - (1) their effectiveness in learning complex representations across diverse modalities [22, 43], (2) the flexibility of implementing any architecture for the generator and discriminator and the potential adversarial training offers, (3) reduced training and inference time compared to Diffusion models. We discuss the GAN challenges in section 7

Our method overview is shown in figure:3, where we adapt the conditional GAN [32] architecture to latent space. In particular, the generator G takes the latent z and conditioned input c and generates the fake motion latent space $z' = G(z, c)$. On the other hand, discriminator D learns to differentiate between real motion latent space $\mathcal{E}(x)$ and fake motion latent space z' conditioned on c. We write the training objective of LS-GAN as a two-player min-max game with: $\min_G \max_D \mathbb{E}_{z\sim\varepsilon(x)}[\log D(z, c)] + \mathbb{E}_{z\sim p_z(z)}[\log(1 - D(G(z, c), c))]$. During generation, we decode z' into the motion sequences using the decoder $\mathcal{D}$, that is $\hat{x}^{1:L} = \mathcal{D}(z') = \mathcal{D}(G(z, c))$.

3.3. GAN architectures

We experiment with two different LS-GAN architectures in the latent space setting.

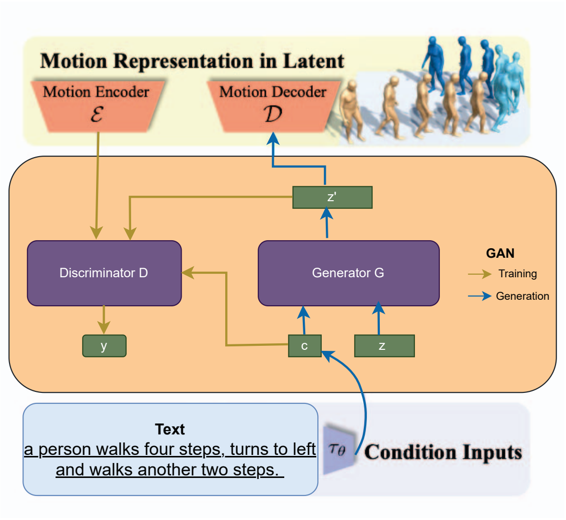


Figure 3. Method overview: The Generator G maps the conditional input $[z, c]$ to a latent z'. The Discriminator D learns to differentiate real $\mathcal{E}(x)$ vs. fake motion latent z' . Finally at generation, we maps the learned latent z' to motion sequence using the decoder $\mathcal{D}(z')$

Vanilla GAN: The Generator comprises three fully connected layers and the Discriminator consists of four fully connected layers. Both models employ leaky ReLU activation to all layers preceding the final layer.

Deep GAN: We add two residual blocks between the fully connected layers in both the Generator and discriminator architectures. Residual connections [17] helps to train deeper networks by overcoming the vanishing gradients.

4. Dataset, Loss and Evaluation metrics

4.1. Dataset

Text-to-motion: HumanML3D [14] is a 3D human motion-language dataset which covers a wide range of human actions including human activities like walking, jumping, swimming, playing golf etc. It contains 14,616 motion sequences from AMASS [31] and annotates 44,970 sequence-level textual descriptions. Here, we employ the motion representation as combination of: 3D joint rotations, positions, velocities, and foot contact.

Action-to-motion: HumanAct12 [16] is a action-to-motion language dataset that provides 1,191 raw motion sequences and 12 action categories.

4.2. Loss

We experiment with Binary Cross entropy (BCE) and Wasserstein loss [4]. We use sigmoid activation on the dis-

Methods	R Precision ↑			FID↓	MM Dist↓	Diversity→	MModality↑
	Top 1	Top 2	Top 3				
Real	$0.511^{\pm.003}$	$0.703^{\pm.003}$	$0.797^{\pm.002}$	$0.002^{\pm.000}$	$2.974^{\pm.008}$	$9.503^{\pm.065}$	-
Seq2Seq [37]	$0.180^{\pm.002}$	$0.300^{\pm.002}$	$0.396^{\pm.002}$	$11.75^{\pm.035}$	$5.529^{\pm.007}$	$6.223^{\pm.061}$	-
LJ2P [3]	$0.246^{\pm.001}$	$0.387^{\pm.002}$	$0.486^{\pm.002}$	$11.02^{\pm.046}$	$5.296^{\pm.008}$	$7.676^{\pm.058}$	-
T2G [5]	$0.165^{\pm.001}$	$0.267^{\pm.002}$	$0.345^{\pm.002}$	$7.664^{\pm.030}$	$6.030^{\pm.008}$	$6.409^{\pm.071}$	-
Hier [9]	$0.301^{\pm.002}$	$0.425^{\pm.002}$	$0.552^{\pm.004}$	$6.532^{\pm.024}$	$5.012^{\pm.018}$	$8.332^{\pm.042}$	-
TEMOS [36]	$0.424^{\pm.002}$	$0.612^{\pm.002}$	$0.722^{\pm.002}$	$3.734^{\pm.028}$	$3.703^{\pm.008}$	$8.973^{\pm.071}$	$0.368^{\pm.018}$
T2M [15]	$0.457^{\pm.002}$	$0.639^{\pm.003}$	$0.740^{\pm.003}$	$1.067^{\pm.002}$	$3.340^{\pm.008}$	$9.188^{\pm.002}$	$2.090^{\pm.083}$
MDM [48]	$0.320^{\pm.005}$	$0.498^{\pm.004}$	$0.611^{\pm.007}$	$0.544^{\pm.044}$	$5.566^{\pm.027}$	$\mathbf{9.559}^{\pm.086}$	$\underline{2.799}^{\pm.072}$
MotionDiffuse [55]	$\mathbf{0.491}^{\pm.001}$	$\mathbf{0.681}^{\pm.001}$	$\mathbf{0.782}^{\pm.001}$	$0.630^{\pm.001}$	$\mathbf{3.113}^{\pm.001}$	$\underline{9.410}^{\pm.049}$	$1.553^{\pm.042}$
MLD [7]	$\underline{0.481}^{\pm.003}$	$\underline{0.673}^{\pm.003}$	$\underline{0.772}^{\pm.002}$	$\mathbf{0.473}^{\pm.013}$	$\underline{3.196}^{\pm.010}$	$9.724^{\pm.082}$	$2.413^{\pm.079}$
Vanilla GAN (Ours)	$0.327^{\pm.002}$	$0.492^{\pm.002}$	$0.599^{\pm.002}$	$1.507^{\pm.017}$	$3.994^{\pm.008}$	$9.320^{\pm.085}$	$0.313^{\pm.020}$
Vanilla WGAN-GP (Ours)	$0.437^{\pm.002}$	$0.622^{\pm.002}$	$0.728^{\pm.002}$	$0.782^{\pm.016}$	$3.395^{\pm.007}$	$9.180^{\pm.085}$	$2.419^{\pm.091}$
Deep GAN (Ours)	$0.352^{\pm.002}$	$0.531^{\pm.002}$	$0.645^{\pm.002}$	$3.036^{\pm.028}$	$3.907^{\pm.006}$	$8.631^{\pm.071}$	$0.308^{\pm.016}$
Deep WGAN-GP (Ours)	$0.391^{\pm.002}$	$0.572^{\pm.002}$	$0.675^{\pm.002}$	$\underline{0.482}^{\pm.013}$	$3.731^{\pm.014}$	$9.249^{\pm.067}$	$\mathbf{3.501}^{\pm.144}$

Table 1. Comparison of text-conditional motion synthesis on HumanML3D dataset. These metrics are evaluated by the motion encoder from [15]. Empty MModality indicates the non-diverse generation methods. The right arrow → means the closer to real motion the better. **Bold** and underline indicate the best and the second best result.

Methods	FLOPs (G) ↓				Parameter	FID ↓			
	DDIM			DDPM		DDIM			DDPM
	50	100	200	1000		50	100	200	1000
MDM	597.97	1195.94	2391.89	11959.44	$x \in \mathbb{R}^{196\times512}$	7.334	5.990	5.936	0.544
MLD	29.86	33.12	39.61	91.60	$z \in \mathbb{R}^{1\times256}$	0.473	0.426	0.432	0.568
Vanilla GAN	**1.581**				$z \in \mathbb{R}^{1\times100}$	0.783			
Deep GAN	**2.665**				$z \in \mathbb{R}^{1\times100}$	0.482			

Table 2. Evaluation of floating-point operations on text-to-motion. We evaluate the FLOPs on 2048 motion clips, counted by THOP library.

criminator with BCE loss. We use Gradient penality [11] instead of weight clipping in Wasserstein GAN.

4.3. Metrics

To assess the performance of our models, we utilize metrics as in MLD [7]: FID, R-precision, Diversity, Multimodality, Multimodal Distance (MM Dist), Average position error(APE), Average variance error (AVE). To measure the computational workload , we use FLOPs.

5. Training details and Results

5.1. Implementation details

We borrow the Motion transformer encoders $\mathcal{E}$ and decoder $\mathcal{D}$ from in MLD [7]. Our VAE model consists of 9 layers and 4 heads with skip connections. To train VAE, we follow the same loss configuration as MLD. All our models are trained on A100 GPU with AdamW optimizer using a fixed learning rate of 10^{-4}. Our batch size is set to 128 during the VAE training stage and 64 during the LS-GAN training stage. We report the test metrics on the training checkpoints with the lowest FID. In all of our experiments, we use the latent dimension $z \in \mathbb{R}^{1\times100}$, $z' \in \mathbb{R}^{1\times256}$. We choose $c \in \mathbb{R}^{1\times768}$ for text-to-motion task and $c \in \mathbb{R}^{1\times10}$ for action-to-motion task.

5.2. Text-to-motion

For text-to-motion, we utilize the VAE checkpoint from iteration 1250 and keep it fixed during GAN training. For detailed evaluation metrics of the VAE, refer table:4. Figures:1, 2 show the qualitative results for the text-to-motion task with LS-GAN (Refer Appendix 9.1 for more results). Table:1 summarizes the test metrics with mean and 95% confidence interval from 20 times running (most of the results are borrowed from MLD [7]). We observe that the vanilla and deep GAN architectures gave the best empirical metrics and qualitative results when used with wasserstein loss with gradient penality. Table:2 depicts the total number of floating-point operations on 2048 motion clips.

Methods	HumanAct12			
	FID$_{\text{train}}$ ↓	ACC ↑	DIV→	MM→
Real	$0.020^{\pm.010}$	$0.997^{\pm.001}$	$6.850^{\pm.050}$	$2.450^{\pm.040}$
ACTOR [33]	$0.120^{\pm.000}$	$0.955^{\pm.008}$	$\underline{6.840}^{\pm.030}$	$\underline{2.530}^{\pm.020}$
INR [6]	$\underline{0.088}^{\pm.004}$	$\underline{0.973}^{\pm.001}$	$6.881^{\pm.048}$	$2.569^{\pm.040}$
MDM [49]	$0.100^{\pm.000}$	$\mathbf{0.990}^{\pm.000}$	$6.680^{\pm.050}$	$\mathbf{2.520}^{\pm.010}$
MLD [7]	$\mathbf{0.077}^{\pm.004}$	$0.964^{\pm.002}$	$6.831^{\pm.050}$	$2.824^{\pm.038}$
Deep WGAN-GP (Ours)	$0.110^{\pm.004}$	$0.942^{\pm.002}$	$\mathbf{6.850}^{\pm.053}$	$2.585^{\pm.057}$

Table 3. Comparison of action-conditional motion synthesis on HumanAct12: FID$_{\text{train}}$ indicate the evaluated splits. Accuracy (ACC) for action recognition. Diversity (DIV), MModality (MM) for generated motion diversity within each action label. The right arrow → means the closer to real motion the better. **Bold** and underline indicate the best and the second best result.

Our Deep WGAN-GP achieves a FID of 0.482 that is near-parity with the state-of-the-art MLD [7] (FID of 0.473) with 91% in FLOPs reduction as shown in Table:2. It outperforms MDM [48] in R precision, FID, MM Dist, MModality. It also achieves state-of-the-art across MModality compared to all the previous models. Furthermore, our LS-GAN outperforms cross-modal models such as Seq2Seq [37], LJ2P [3], T2G [5] , Hier [9], TEMOS [36], T2M [15] across all evaluation metrics. This signifies high-quality motion and high text prompt matching while maintaining a rich motion diversity as evident in Figures:1, 2. These results demonstrate that a simple GAN in latent space can achieve impressive results with minimal compute in both training and inference compared to the Diffusion models.

5.3. Action-to-motion

The action-conditioned task involves generating motion sequences based on an input action label. We compare our Deep WGAN-GP with ACTOR [33], INR [6], MDM [49], and MLD [7]. ACTOR and INR are transformer-based VAE models specifically designed for the action-conditioned task. In contrast, MDM and MLD are diffusion models that utilize the same learnable action embedding module as our method. We report the test metrics as the mean and 95% confidence interval computed from 20 independent runs.

From table:3, we observe that Deep WGAN-GP outperforms all the other models in Diversity while maintaining competitive performance on FID, accuracy and MultiModality (MM). These results indicate that GAN in motion latent can also benefit action-conditioned motion generation task.

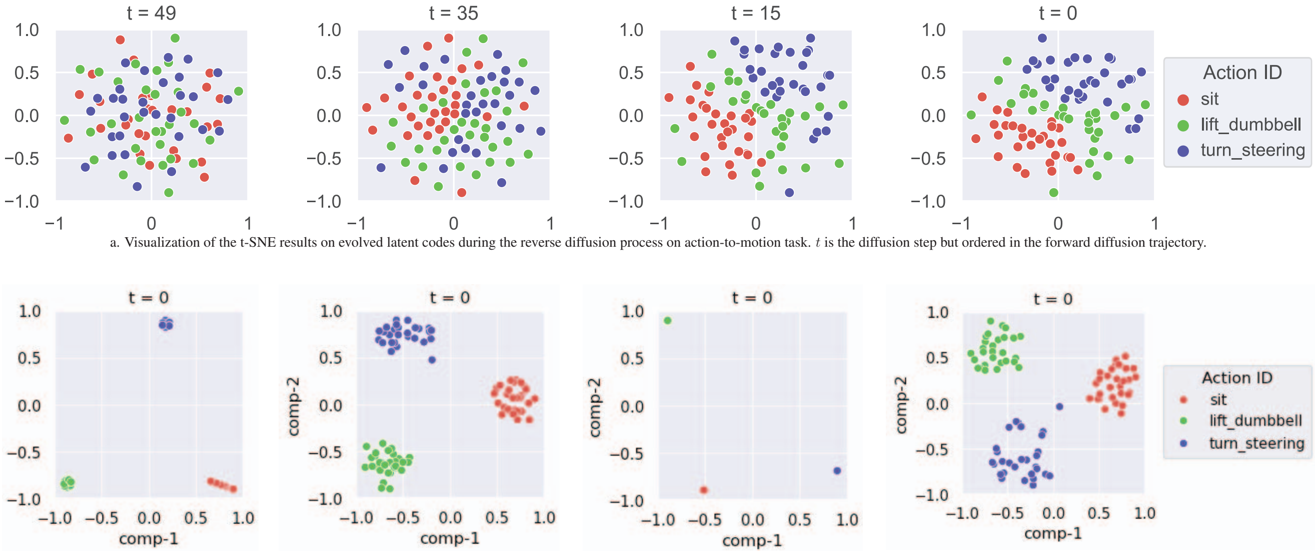
Metric	VAE 250 checkpoint	VAE 1250 checkpoint
APE_root/mean ↓	$0.0897^{\pm 0.0002}$	$\mathbf{0.0756}^{\pm 0.0002}$
APE_traj/mean ↓	$0.0857^{\pm 0.0002}$	$\mathbf{0.0723}^{\pm 0.0002}$
APE_mean_pose/mean ↓	$0.0379^{\pm 0.0000}$	$\mathbf{0.0312}^{\pm 0.0000}$
APE_mean_joints/mean ↓	$0.1008^{\pm 0.0002}$	$\mathbf{0.0845}^{\pm 0.0002}$
AVE_root/mean ↓	$0.0221^{\pm 0.0001}$	$\mathbf{0.0201}^{\pm 0.0001}$
AVE_traj/mean ↓	$0.0220^{\pm 0.0001}$	$\mathbf{0.0200}^{\pm 0.0001}$
AVE_mean_pose/mean ↓	$0.0021^{\pm 0.0000}$	$\mathbf{0.0015}^{\pm 0.0000}$
AVE_mean_joints/mean ↓	$0.0241^{\pm 0.0001}$	$\mathbf{0.0216}^{\pm 0.0001}$
R_precision_top_1 ↑	$0.4422^{\pm 0.0030}$	$\mathbf{0.4891}^{\pm 0.0020}$
R_precision_top_2 ↑	$0.6337^{\pm 0.0020}$	$\mathbf{0.6803}^{\pm 0.0023}$
R_precision_top_3 ↑	$0.7379^{\pm 0.0025}$	$\mathbf{0.7787}^{\pm 0.0021}$
FID ↓	$1.1754^{\pm 0.0030}$	$\mathbf{0.2661}^{\pm 0.0010}$
Diversity →	$\mathbf{9.3856}^{\pm 0.0843}$	$9.6901^{\pm 0.0990}$
MultiModality ↑	$\mathbf{0.2056}^{\pm 0.0095}$	$0.1237^{\pm 0.0058}$

Table 4. Comparison of evaluation metrics (mean and 95% confidence interval from running 20 times) on text-to-motion VAE 250, 1250th checkpoint. **Bold** indicate the best result.

6. Latent space visualization

In this section, we present t-SNE visualizations of the latent space on action-to-motion task, illustrating how our LS-GAN effectively captures and separates different actions within the latent space. These results are compared with the MLD [7] in Figure 4.

From the latent space visualization, it is evident that Vanilla GAN and Deep GAN have low MultiModality scores (measures the generation diversity within the same text or action input), while Vanilla WGAN-GP and Deep WGAN-GP have higher MultiModality highlighting the effectiveness of the Wasserstein loss with gradient penalty. This observation even holds true for the text-to-motion generation task, as evidenced by the results presented in Table 1. Furthermore, our approach shows superior separation of latent code clusters at timestep $t = 0$ compared to MLD. This improved clustering at $t = 0$ indicates that our LS-GAN framework captures a more structured motion latent



a. Visualization of the t-SNE results on evolved latent codes during the reverse diffusion process on action-to-motion task. t is the diffusion step but ordered in the forward diffusion trajectory.

b. Visualization of the t-SNE results on latent codes for Vanilla GAN, Vanilla WGAN-GP, Deep GAN, Deep WGAN-GP (left to right)

Figure 4. Visualization of the t-SNE results on latent codes of LS-GAN compared to MLD. We sample 30 motions for each action label

representation, potentially leading to better interpretability and generation fidelity.

7. Discussion

7.1. Addressing GAN challenges:

Usage of condition information in our model helps us to overcome mode collapse challenge by conditioning the model on additional information. In addition, our generator learns the inherent features of real motion data. This encourages the discriminator to compare the underlying properties instead of the high dimensional real data, similar to Feature Matching [42] that helps to stabilize training. Methods such as regularization, spectral normalization , adaptive learning rates, multiple generators/discriminators, auxiliary loss as discussed by [45] can be further explored to stabilize GAN training.

7.2. Accelerated Diffusion:

Diffusion distillation [30, 44, 53] is a knowledge distillation task, where a student model is trained to distill the multi-step outputs of the original diffusion model into a single or few steps. These prior works, require a separate pretraining and distillation phase. In addition, one-step diffusion models require a greater attention in choosing the training objectives and scheduling mechanism [53]. On the other hand, recent works on GANs [20,43] shows that StyleGAN-T, Giga-Gan outperforms Distilled diffusion models on text-to-image generation. Considering these, we believe GAN in latent space would serve as a solution to accelerated diffusion.

7.3. Limitations

First, similar to most motion generation methods, our approach can generate motion sequences of arbitrary lengths, but still below the maximum length in the dataset. Secondly, LS-GAN specifically targets human body motion, in contrast to works focusing on facial motion [21] or hand motion [29]. Lastly, we limited ourselves to simple prompts for text-to-motion. It may be beneficial to consider the impact of motion outputs on edge cases and ambiguous text descriptions.

8. Conclusion

In this paper, we introduced a novel approach for text-to-motion and action-to-motion synthesis using Generative Adversarial Networks in the latent space. By leveraging the power of GANs and the compact representation of motion sequences in the latent space, our method achieves faster training and inference times compared to previous methods while maintaining high-quality motion synthesis results. Results demonstrate that a simple GAN in latent space is comparable to complex models. This work will open a new direction in exploring latent space GANs that can have faster stable training and inference compared to latent space diffusion.

References

[1] Hyemin Ahn, Timothy Ha, Yunho Choi, Hwiyeon Yoo, and Songhwai Oh. Text2action: Generative adversarial synthesis from language to action. *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1–5, 2017. 3

[2] Chaitanya Ahuja and Louis-Philippe Morency. Language2pose: Natural language grounded pose forecasting. *2019 International Conference on 3D Vision (3DV)*, pages 719–728, 2019. 1

[3] Chaitanya Ahuja and Louis-Philippe Morency. Language2pose: Natural language grounded pose forecasting. In *2019 International Conference on 3D Vision (3DV)*, pages 719–728. IEEE, 2019. 4, 5

[4] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan, 2017. 3

[5] Uttaran Bhattacharya, Nicholas Rewkowski, Abhishek Banerjee, Pooja Guhan, Aniket Bera, and Dinesh Manocha. Text2gestures: A transformer-based network for generating emotive body gestures for virtual agents. In *2021 IEEE Virtual Reality and 3D User Interfaces (VR)*, pages 1–10. IEEE, 2021. 4, 5

[6] Pablo Cervantes, Yusuke Sekikawa, Ikuro Sato, and Koichi Shinoda. Implicit neural representations for variable length human motion generation. In *European Conference on Computer Vision*, 2022. 5

[7] Xin Chen, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, Jingyi Yu, and Gang Yu. Executing your commands via motion diffusion in latent space, 2023. 1, 2, 3, 4, 5

[8] Eduardo de Campos Valadares and Cleber P. A. Anconi. Dancing to the music. *The Physics Teacher*, 38:404–404, 2000.

[9] Anindita Ghosh, Noshaba Cheema, Cennet Oguz, Christian Theobalt, and Philipp Slusallek. Synthesis of compositional animations from textual descriptions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1396–1406, 2021. 4, 5

[10] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014. 1

[11] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. Improved training of wasserstein gans, 2017. 1, 4

[12] Chuan Guo, Xinxin Xuo, Sen Wang, and Li Cheng. Tm2t: Stochastic and tokenized modeling for the reciprocal generation of 3d human motions and texts. *ArXiv*, abs/2207.01696, 2022. 1

[13] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5142–5151, 2022. 1

[14] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5152–5161, 2022. 1, 3

[15] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5152–5161, June 2022. 4, 5

[16] Chuan Guo, Xinxin Zuo, Sen Wang, Shihao Zou, Qingyao Sun, Annan Deng, Minglun Gong, and Li Cheng. Action2motion: Conditioned generation of 3d human motions. *Proceedings of the 28th ACM International Conference on Multimedia*, 2020. 1, 3

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. 3

[18] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020. 1

[19] Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. Motiongpt: Human motion as a foreign language, 2023. 2

[20] Minguk Kang, Jun-Yan Zhu, Richard Zhang, Jaesik Park, Eli Shechtman, Sylvain Paris, and Taesung Park. Scaling up gans for text-to-image synthesis, 2023. 6

[21] Tero Karras, Timo Aila, Samuli Laine, Antti Herva, and Jaakko Lehtinen. Audio-driven facial animation by joint end-to-end learning of pose and emotion. *ACM Transactions on Graphics (TOG)*, 36:1 – 12, 2017. 6

[22] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks, 2019. 1, 3

[23] Jihoon Kim, Jiseob Kim, and Sungjoon Choi. Flame: Free-form language-based motion synthesis & editing. In *AAAI Conference on Artificial Intelligence*, 2022. 1

[24] Jihoon Kim, Jiseob Kim, and Sungjoon Choi. Flame: Free-form language-based motion synthesis & editing, 2023. 2

[25] Buyu Li, Yongchi Zhao, Zhelun Shi, and Lu Sheng. Danceformer: Music conditioned 3d dance generation with parametric motion transformer. In *AAAI Conference on Artificial Intelligence*, 2021. 1

[26] Peizhuo Li, Kfir Aberman, Zihan Zhang, Rana Hanocka, and Olga Sorkine-Hornung. Ganimator: neural motion synthesis from a single sequence. *ACM Transactions on Graphics*, 41(4):1–12, July 2022. 2

[27] Ruilong Li, Sha Yang, David A. Ross, and Angjoo Kanazawa. Ai choreographer: Music conditioned 3d dance generation with aist++. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13381–13392, 2021. 1

[28] Ruilong Li, Shan Yang, David A. Ross, and Angjoo Kanazawa. Ai choreographer: Music conditioned 3d dance generation with aist++, 2021. 2

[29] Yuwei Li, Minye Wu, Yuyao Zhang, Lan Xu, and Jingyi Yu. Piano: A parametric hand bone model from magnetic resonance imaging. In *International Joint Conference on Artificial Intelligence*, 2021. 6

[30] Weijian Luo. A comprehensive survey on knowledge distillation of diffusion models, 2023. 6

[31] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. Amass: Archive of

motion capture as surface shapes. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5441–5450, 2019. 3

[32] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets, 2014. 1, 3

[33] Mathis Petrovich, Michael J. Black, and Gül Varol. Action-conditioned 3d human motion synthesis with transformer vae. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10965–10975, 2021. 1, 5

[34] Mathis Petrovich, Michael J. Black, and Gül Varol. Temos: Generating diverse human motions from textual descriptions. *ArXiv*, abs/2204.14109, 2022. 1

[35] Mathis Petrovich, Michael J. Black, and Gül Varol. Temos: Generating diverse human motions from textual descriptions, 2022. 2

[36] Mathis Petrovich, Michael J. Black, and Gül Varol. TEMOS: Generating diverse human motions from textual descriptions. In *European Conference on Computer Vision (ECCV)*, 2022. 4, 5

[37] Matthias Plappert, Christian Mandery, and Tamim Asfour. Learning a bidirectional mapping between human whole-body motion and natural language using deep recurrent neural networks. *Robotics and Autonomous Systems*, 109:13–26, 2018. 4, 5

[38] Sigal Raab, Inbal Leibovitch, Peizhuo Li, Kfir Aberman, Olga Sorkine-Hornung, and Daniel Cohen-Or. Modi: Unconditional motion synthesis from diverse data, 2022. 2

[39] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 2

[40] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021. 3

[41] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022. 2

[42] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans, 2016. 6

[43] Axel Sauer, Tero Karras, Samuli Laine, Andreas Geiger, and Timo Aila. Stylegan-t: Unlocking the power of gans for fast large-scale text-to-image synthesis, 2023. 1, 3, 6

[44] Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation, 2023. 6

[45] Divya Saxena and Jiannong Cao. Generative adversarial networks (gans survey): Challenges, solutions, and future directions, 2023. 6

[46] Ayumi Shiobara and Makoto Murakami. Human motion generation using wasserstein gan. In *Proceedings of the 2021 5th International Conference on Digital Signal Processing*, ICDSP '21, page 278–282, New York, NY, USA, 2021. Association for Computing Machinery. 2

[47] Guy Tevet, Brian Gordon, Amir Hertz, Amit H. Bermano, and Daniel Cohen-Or. Motionclip: Exposing human motion generation to clip space, 2022. 1

[48] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Amit H Bermano, and Daniel Cohen-Or. Human motion diffusion model. *arXiv preprint arXiv:2209.14916*, 2022. 2, 4, 5

[49] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H. Bermano. Human motion diffusion model. *ArXiv*, abs/2209.14916, 2022. 1, 5

[50] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H. Bermano. Human motion diffusion model, 2022. 2

[51] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning, 2018. 2

[52] Liang Xu, Ziyang Song, Dongliang Wang, Jing Su, Zhicheng Fang, Chenjing Ding, Weihao Gan, Yichao Yan, Xin Jin, Xiaokang Yang, Wenjun Zeng, and Wei Wu. Actformer: A gan-based transformer towards general action-conditioned 3d human motion generation, 2022. 3

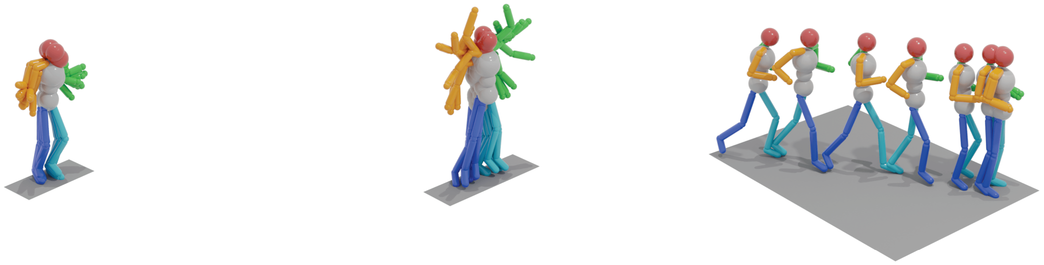
[53] Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Fredo Durand, William T. Freeman, and Taesung Park. One-step diffusion with distribution matching distillation, 2024. 6

[54] Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Shaoli Huang, Yong Zhang, Hongwei Zhao, Hongtao Lu, and Xi Shen. T2m-gpt: Generating human motion from textual descriptions with discrete representations, 2023. 2

[55] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *arXiv preprint arXiv:2208.15001*, 2022. 1, 2, 4

9. Appendix

9.1. LS-GAN qualitative results on text-to-motion:

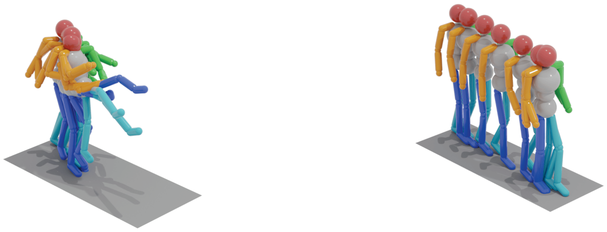


a. A person is skipping rope.

b. a person doing jumping jacks.

c. a person jogs straight forward.

Figure 5. Qualitative results of our method shown by Vanilla GAN

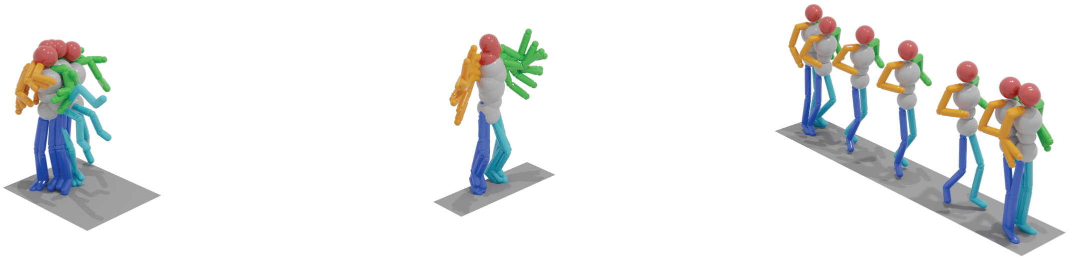


a. a man kicks with something or someone with his left leg.

b. a person walks backward slowly.

c. a person jogs straight forward.

Figure 6. Qualitative results of our method shown by Deep GAN

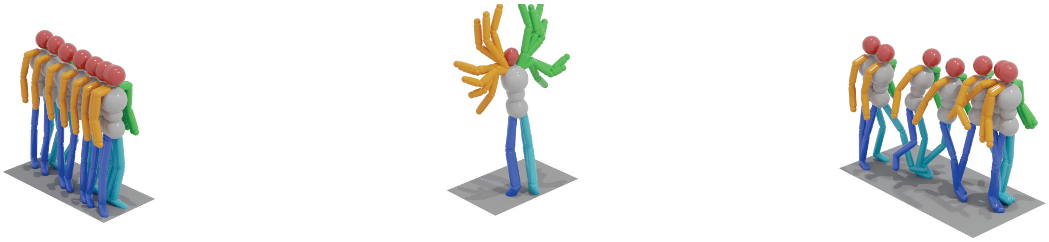


a. a man kicks with something or someone with his left leg.

b. a person doing jumping jacks.

c. a person jogs straight forward.

Figure 7. Qualitative results of our method shown by Vanilla WGAN-GP



a. a person walks backward slowly.

b. a person raised arms up and pull them down.

c. a person walking forward with legs wide apart.

Figure 8. Qualitative results of our method shown by Deep WGAN GP