Spatio-Temporal Graph Diffusion for Text-Driven Human Motion Generation

BMVC 2023 Submission # 722

Abstract

Text-based human motion generation is challenging due to the complexity and context-dependency of natural human motions. In recent years, an increasing number of studies have focused on using transformer-based diffusion models to tackle this issue. However, an over-reliance on transformers has resulted in a lack of adequate detail in the generated motions. This study proposes a novel graph network-based diffusion model to address this challenging problem. Specifically, we use spatio-temporal graphs to capture local details for each node and an auxiliary transformer to aggregate the information across all nodes. In addition, the transformer is also used to process conditional global information that is difficult to handle with graph networks. Our model achieves competitive results on currently the largest dataset HumanML3D and outperforms existing diffusion models in terms of FID and diversity, demonstrating the advantages of graph neural networks in modeling human motion data. The source code will be publicly available.

1 Introduction

Human motion generation aims to generate realistic human motions for various applications, such as video games, film production and virtual reality. For instance, in virtual reality, creating natural and believable human motions is crucial for providing an immersive experience that mimics real-life scenarios. One important aspect is the ability to generate motions that are conditional on various inputs, such as action labels, music, and text. Text-driven generation, in particular, has gained significant attention for its advantage of being intuitive and easy to use, as it does not require users to master preliminary knowledge or operate complex software. In this approach, the input to the motion generation system is a natural language description of the desired motion, and the output is a sequence of human motions that corresponds to the description.

Recently, there have been promising results in human motion generation using diffusion models based on transformers [33, 39, 41]. Nevertheless, several studies [20, 36, 37] have shown that while transformers excel at capturing long-range correlations, they may not be as efficient at capturing fine-grained local details. This is particularly evident in human motion generation, as even a minor change in the joint angles, timing, or phasing can significantly compromise the quality of the generated motion.

On the other hand, graph neural networks (GNNs) [22, 29] have demonstrated good performance in modeling human motion but struggle to capture global context. This issue arises because graph neural networks rely on a message-passing mechanism to obtain information from neighboring nodes. To capture a broader context, increasing the number of GNN layers is often necessary [19], which leads to the GNN repeatedly aggregating information from neighboring nodes. Consequently, nodes become progressively more similar, leading to what is known as the over-smoothing problem [3]. Also, handling conditional global information such as text description and action labels in GNN remains challenging. This is because this information belong to all nodes, and the structure of the graph does not provide a natural way to efficiently broadcast this information to all nodes. Therefore, previous GNN-based approaches [7] rarely focus on conditional generation. Finally, graph-based diffusion models have received relatively less attention compared to other types of diffusion models such as UNet or transformer based models. Nonetheless, they have shown great success in various fields, most notably in traffic flow forecasting [24, 31, 35], where they have been able to capture the complex relationships between different traffic nodes and predict traffic patterns more accurately. In the field of human motion generation, as far as we know there have not been studies using graph-based diffusion models.

To tackle the aforementioned challenges, we propose Spatio-Temporal Graph Motion Diffusion (STGMD), a human motion generation framework that can generate diverse motions from natural language descriptions. To incorporate both local and global context, we employ a two-stage approach in our model. First, we utilize a GNN to capture the local interactions between human body joints. Then, an auxiliary transformer is employed to aggregate the information from each joint and capture the global context. The combination of GNN and transformer allows our model to effectively process both local and global context. Furthermore, the auxiliary transformer enables our graph network-based model to easily handle conditional global information, such as text descriptions.

In summary, the main contributions of this paper are as follows.

- Our proposed STGMD model leverages the advantages of both GNN and Transformer, and the combination enables our model to effectively process both fine-grain and coarse-grain information, resulting in more realistic human motion generation.

- By incorporating an auxiliary transformer, our proposed graph network-based model is capable of effortlessly processing conditional global information (*i.e.*, text description) without the need to broadcast this information to every node.

- Extensive experiments have evinced that the proposed method STGMD exhibits a 30% improvement in the Frechet Inception Distance (FID) metric when compared to the prior state-of-the-art diffusion approaches. Our visualization results highlight the efficacy of our model in generating more realistic human motions.

2 Related Work

Text-Driven Human Motion Generation Human motion generation can be categorized by the type of input signal, including music-driven, speech-driven, text-driven, and others. Among these, text-driven human motion generation has received the most attention because it is highly user-friendly and accessible without the need for specialized equipment or expertise. Text-driven human motion generation involves interpreting text descriptions and converting them into corresponding motion sequences. Ahuja and Morency [1] employ a curriculum learning approach to learn a joint embedding of text and human motion. Tevet

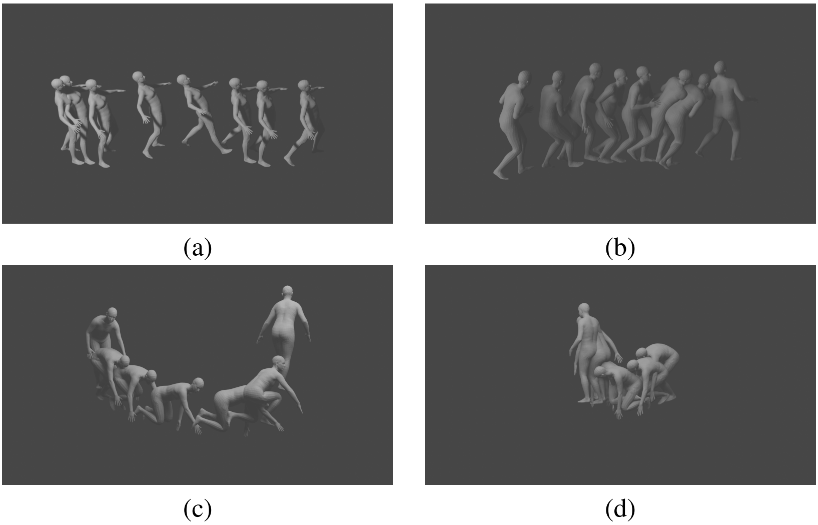

Figure 1: The proposed model is a graph neural based diffusion model for text-driven motion generation method, capable of generating diverse human motion from the same natural text description. (a) and (b): A person twists her body to avoid bumping into someone while walking down a crowded street. (c) and (d): A person crawls on her hands and knees to search for a lost object on the floor.

et al. [32] present a motion auto-encoder based on the transformer architecture, and designed to reconstruct motion sequences while aligned to its text label in the CLIP space. Petrovich et al. [26] propose a transformer-based VAE to learn an action-aware latent representation for human motions. Guo et al. [12] also consider the inverse motion-to-text task by leveraging a discrete and compact motion representation.

Graph Neural Network Graph Neural Networks (GNNs) have been applied in various fields such as natural language processing, social network analysis, and recommendation systems. In human motion modeling, multi-scale spatio-temporal maps are typically employed because the fine details of human motion is difficult to be predicted, and the human model tends to be more stable at a coarser scale. The key for multiple scales is to build a suitable graph structure for each scale. [5] and [17] are examples of methods that manually select nodes for different scales. Another branch of methods of creating different scales is through node drop pooling. The node drop pooling is achieved by removing relatively unimportant nodes, and the importance is usually computed by a node score calculation function. Graph-UNet [8] employs a simple linear layer to estimate the node score. By using graph convolution in self-attention, Lee et al. [16] are able to take into account both the features of the nodes and the topology of the graph. Zhang et al. [40] propose hierarchical graph pooling with structure learning, a method that integrates graph pooling and structure learning into a single module to create hierarchical graph representations. Due to the inherent limitations of graph networks in processing conditional global information (*i.e.*, text description), most previous works that utilize GNN focus on prediction or recognition task [18, 23], and there has been limited research using graph network for conditional human motion generation. Our proposed STGMD narrows this gap by providing an easy way to process conditional global information.

Diffusion Model The diffusion model belongs to a class of score-based generative models that use denoising score matching through annealed Langevin dynamic sampling. It was first introduced by Sohl-Dickstein et al. [30], and later on, Ho et al. [14] demonstrated its ability to generate high-quality samples. Dhariwal and Nichol [6] proposed classifier guidance for conditional generation, which involves using an auxiliary classifier to enable conditional sampling in diffusion models. However, Ho and Salimans [13] demonstrated that classifier guidance is not necessary. They introduced a new approach called classifier-free guidance, which is a technique for guiding diffusion models that does not require a separate classifier model. Their results demonstrate that guidance can be performed by a pure generative model without a classifier.

Recently, diffusion models have gained significant attention in human motion research due to their ability to produce high-quality and diverse samples. Zhang et al. [39] propose MotionDiffuse, the first diffusion model-based text-driven motion generation framework that surpasses existing methods in diversity and quality. Tevet et al. [33] introduce Motion Diffusion Model, a carefully adapted classifier-free diffusion-based generative model for human motion generation. To better capture the representations of various human motion sequences while reducing computational costs, Chen et al. [4] design a variational autoencoder (VAE) to extract representative and low-dimensional latent embedding for human motion sequences and perform diffusion processes in the latent space. While these studies all benefit from the attention mechanism, they may not perform as well on local spatial temporal relationships among skeleton joints. Our STGMD approach leverages a spatio-temporal graph network to address this limitation and enhance performance by better capturing local details.

3 Method

The overview of the proposed STGMD is shown in Figure 2. In particular, the design of the spatio-temporal graph UNet (STG-UNet) is elaborated in Section 3.1, and the details of the diffusion process is explained in Section 3.2.

3.1 Spatio-Temporal Graph UNet

As shown in Figure 3, the STG-UNet consists of multiple Spatio-Temporal Graph Blocks (STG-Block), as well as pooling and unpooling operations.

3.1.1 Spatio-Temporal Graph Block

Before delving into the complete spatio-temporal graph block (STG-Block), let us first explore the functioning of graph convolutional networks at each time point. By representing the human body as a graph, with each joint as a node and the body segments as edges, the purpose of the spatial graph convolutional network is to capture the complex interactions between different body parts and learn to recognize the inherent patterns, such as the position of the knee is dependent on the position of the hip and ankle.

The underlying structure of human motion sequence X is represented by a graph $X = (V, A)$. $V \in \mathbb{R}^{D\times N\times J}$ is the vertex matrix, where D is the dimension of motion feature, N is number of frames and J is number of joints. $A \in \mathbb{R}^{J\times J}$ is the symmetric adjacency matrix with self connection where $a_{i,k}$ denotes the edge weight between nodes v_i and v_k. $D_{ii} = \sum_j A_{ij}$ is the degrees matrix of A. $W^{(l)}$ is the layer-specific trainable weight matrix parameters. $H^{(l)}$

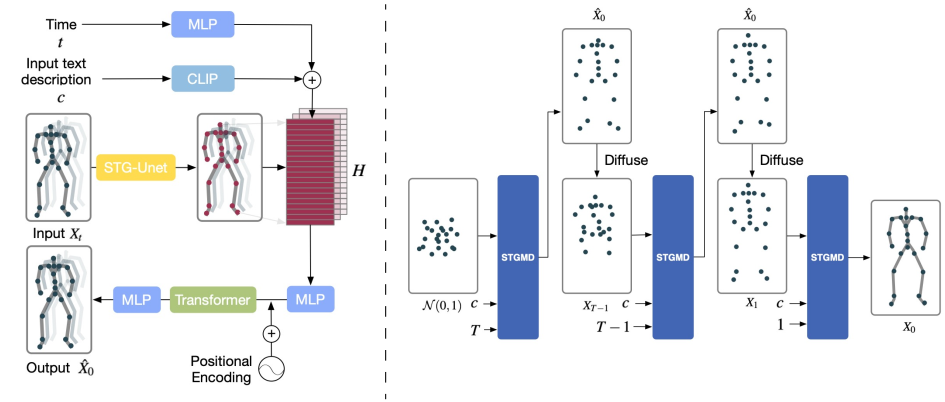

Figure 2: **(Left) STGMD Overview.** The input to the diffusion model consists of a spatio-temporal graph X_t, the noising time step t, and the corresponding text description c. The STG-UNet (Section 3.1) extracts the fine-grained local details of human motion, while the frozen CLIP extracts text embeddings. In addition to aggregating local details, the transformer processes text embeddings and time t. **(Right) Sampling Process.** Starting with random Gaussian noise, a text description c and diffusion step T, STGMD gradually anneals the noise to a sample X_0. At each time step t, our model predicts the sample's initial state $\hat{x}_0$ and then diffuses it back to X_{t-1}. By repeating these operations T times, the diffusion process finally yields the sample X_0. The figure only shows one frame of the graph.

is the input feature map of layer l, and $H^{(0)} = V$. Following [15, 25], the spatial graph convolution in layer l is defined as

$$\tilde{H}^{(l)} = \sigma\left(D^{-\frac{1}{2}} A D^{-\frac{1}{2}} H^{(l)} W^{(l)}\right),$$

where σ is the ReLU activation function.

Spatio-temporal GNN extends the idea of GNN by incorporating temporal information, which is particularly important in human motion tasks, as human motion is inherently dynamic and changes over time. The construction of the spatio-temporal graph is based on the spatial graph, where the same joints across consecutive frames are treated as adjacent nodes. To model the temporal information, we design a 1-D temporal convolution module, which is inspired by the work of Yu et al. [38] and Cao et al. [2]. Specifically, we perform such temporal convolution independently to each human joint. Formally, let $\tilde{H}^{(l)}$ be the feature map after the spatial graph convolution, the temporal graph convolution is defined as

$$H^{(l+1)} = \sigma(CONV_{\Omega}(\tilde{H}^{(l)})),$$

where $CONV_{\Omega}$ denotes the 1-D convolutional operation with trainable parameter Ω.

3.1.2 Multiple Scale Graph

In human motion, the information from different scales (*i.e.*, the coarse scale and the fine scale) usually represent different properties. Coarse scales describe the classes of human motion and are more stable [5], making them useful for generating representative features. Fine scales, on the other hand, complement the details of human motion on top of the coarse scales. Therefore, the relationship within and between the coarse and fine scales are essential

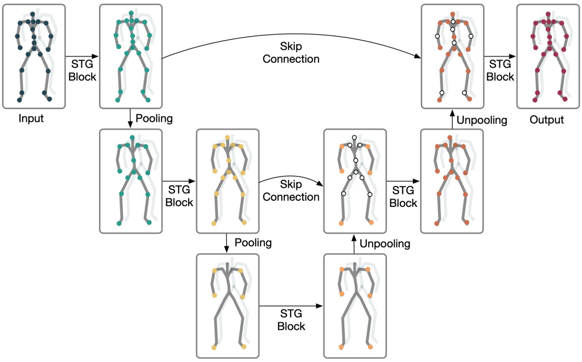

Figure 3: **STG-UNet Overview.** The input is a spatio-temporal graph $X = (V,A)$, which consists of a vertex matrix V and an adjacency matrix A. The STG-Block processes the vertex matrix and transforms it into a graph representation. The pooling operation is then applied to select nodes for the coarser scale. This process is repeated for a coarser scale. In reverse, the unpooling operation refills the previously excluded nodes with empty nodes for finer scales. Additionally, skip connections are employed to connect graph representations on the same scale.

to understand the human motion. To enhance the capacity of our graph network in capturing fine and coarse scales, we design a multi-scale UNet structure, as shown in Figure 3.

On each scale, we choose a different number of nodes to represent the human skeleton through pooling operation adapted from [8]. Let p be the trainable projection vector, the indices for the selected nodes in layer l are computed via a top-k operation idx $=$ topk$(\sigma(\frac{H^{(l)}p^{(l)}}{\|p^{(l)}\|}))$. The unpooling operation is the reverse operation of pooling that refills the previously excluded nodes with empty nodes.

3.1.3 Auxiliary Transformer for Global Context and Information

Although GNNs are useful for capturing fine-grained local details in human motion, they are less effective in extracting long-range dependencies. This would leads to an sub-optimal problem when handling the global context. In addition, since it is struggle for GNNs to model the conditional global information (*i.e.*, the text descriptions), we propose to apply an auxiliary transformer encoder in the network to address these challenges.

As shown in Figure 2, given the graph $X = (H,A)$ with H is the graph embedding generated by STG-UNet, we concatenate the diffusion time step t and the text embeddings with H first. Afterward, with the help of the standard transformer encoder proposed in [34], the local details, the diffusion time step, as well as the text embeddings can be aggregated together, effectively.

3.2 Diffusion Process

The diffusion process starts with a Gaussian noise, and gradually anneals it into a realistic output by applying a series of diffusion steps. The diffusion model has been successfully applied in several fields, such as image synthesis, data augmentation, and human motion synthesis.

Given samples from a data distribution $q(X_0)$, The diffusion model on a graph is modeled as a Markov nosing process using: $q\left(X_{1:T} \mid X_0\right)=\prod_{t=1}^T q\left(X_t \mid X_{t-1}\right)$ where $q(X_t|X_{t-1})$ is a Gaussian distribution as $q(X_t|X_{t-1}) = \mathcal{N}(X_t; \sqrt{\alpha_t}X_{t-1}, (1-\alpha_t)I)$, X_t stand for the temporal graph at t-step diffusion process, and the $\alpha_t \in (0,1)$ is a constant hyper-parameters for sampling. The goal of our model is to approximate the reverse process that recovers X_T from X_0 recurrently: $p_\theta\left(X_{0:T}\right)=p\left(X_T\right) \prod_{t=T}^1 p_\theta\left(X_{t-1} \mid X_t\right)$. Then, the transition between two adjacent time stamps is denoted by $p_\theta\left(X_{t-1} \mid X_t\right)=\mathcal{N}\left(X_{t-1} ; \mu_\theta\left(X_t, t\right), \sigma_\theta\left(X_t, t\right)\right)$, with shared parameters θ. We follow the objective suggested by Ramesh et al. [28], and in text-driven motion generation, the goal is to learn a network D that fits $D_\theta(X_t, t, c)$ with the simple mean squared loss:

$$\min_\theta \mathcal{L}(\theta) = \min_\theta \mathbb{E}_{X_0 \sim q(X_0, c), t \sim [1,T]} \left\| X_0 - D_\theta(X_t, t, c) \right\|_2^2 .$$

4 Experiments

This section presents the experimental results of our proposed method for text-conditional human motion generation. Section 4 describes the dataset used and implementation details. Section 4.1 then compares the performance of our method to state-of-the-art approaches.

Dataset HumanML3D [11] is currently the largest 3D human motion dataset that combines two publicly available large-scale 3D human motion datasets, HumanAct12 [10] and AMASS [21]. The two datasets contain a variety of human actions, such as daily activities, sports, acrobatics, and artistry. Each motion sequence is annotated by three native English speakers with at least five words, and a manual post-processing step filters away abnormal text descriptions. As a result, the HumanML3D dataset comprises 14,616 motions and 44,970 descriptions, and the average motion length is 7.1 seconds.

Implementation Details The number of human motion frames is set to 196, which is the maximum frame length for HumanML3D. The depth of STG-UNet is set to three, and we select 21,10, and 5 nodes for each scale, and the ablation study is provided in Section 4.2. We use AdamW optimizer with a fixed learning rate of 10^{-4}, and the batch size is set to 64. The number of diffusion steps is 1K. For runtime, training takes about 36 hours on RTX A6000 GPU.

Evaluation metrics To evaluate our proposed model, objective metrics are necessary to assess the generation quality, generation diversity, and text conditional matching [11]. We use a pre-trained network [11] to extract motion and text representation and evaluate the results based on the extracted representation.

Generation quality is evaluated using the *Frechet Inception Distance (FID)* metric, which measures the distributional similarity between the generated and real motion sequences. Lower FID scores indicate better similarity.

Generation diversity is evaluated using two metrics: *Diversity* and *Multimodality*. Diversity is measured by the variety of generated motions, and the closer it is to the real distribution, the better. Multimodality measures the ability to generate diverse motions given the same text description, with higher scores indicating greater diversity.

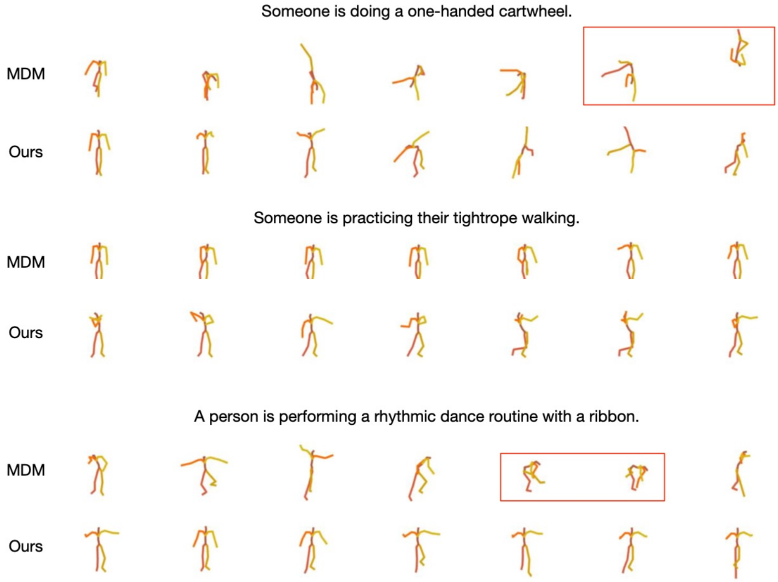
Someone is doing a one-handed cartwheel.
MDM
Ours
Someone is practicing their tightrope walking.
MDM
Ours
A person is performing a rhythmic dance routine with a ribbon.
MDM
Ours

Figure 4: We present qualitative results on the HumanML3D dataset, where we compare our approach with MDM Tevet et al. [33] and visualize the outcomes for given prompts. Distorted motions are highlighted in red

Text conditional matching is evaluated using two metrics, *i.e.*, *R-Precision* and *Multi-modal Distance (MM-Dist)*. R-Precision measures the accuracy of the generated motion sequences in matching text descriptions, with higher scores indicating better alignment. MM-Dist measures the distance between the text description and the generated motion, with lower scores indicating better similarity.

4.1 State-of-the-Art Comparisons

We compare proposed STGMD to existing state-of-the-art methods [4, 9, 11, 27, 33, 39] on the test set of HumanML3D, and the results are shown in Table 1. The results indicate that the STGMD outperforms existing methods in terms of FID, diversity, and multimodality. Specifically, the proposed method attains the lowest FID score, indicating superior quality of generated motions compared to other methods, as also shown in Figure 4. Moreover, the proposed method can generate more diverse results, as evidenced by higher multimodality scores and a similar Diversity to the ground truth. Although STGMD does not achieve the best performance in terms of R precision and multimodal distance, it is worth noting that the proposed method still delivers competitive performance in both metrics while surpassing other methods in other metrics.

4.2 Ablation Studies

In this part, we firstly analyze the effectiveness of the STG-UNet, then we investigate the impact of various depths and the number of nodes in each scale on the model's performance.

The results of experiments with and without STG-UNet are presented in Table 2. The findings demonstrate that removing STG-UNet results in a performance drop across all met-

Methods	FID ↓	R Precision ↑ Top 3	Diversity →	Multimodal Dist ↓	Multimodality ↑
Real	$0.002^{\pm.000}$	$0.797^{\pm.002}$	$9.503^{\pm.065}$	$2.974^{\pm.008}$	-
Hier [9]	$6.532^{\pm.024}$	$0.552^{\pm.004}$	$8.332^{\pm.042}$	$5.012^{\pm.018}$	-
TEMOS [27]	$3.734^{\pm.028}$	$0.722^{\pm.002}$	$8.973^{\pm.071}$	$3.703^{\pm.008}$	$0.368^{\pm.018}$
MotionDiffuse [39]	$0.630^{\pm.001}$	**0.782**$^{\pm.001}$	$9.410^{\pm.049}$	**3.113**$^{\pm.001}$	$1.553^{\pm.042}$
T2M [11]	$1.067^{\pm.002}$	$0.740^{\pm.003}$	$9.188^{\pm.002}$	$3.340^{\pm.008}$	$2.090^{\pm.083}$
MDM [33]	$0.544^{\pm.044}$	$0.611^{\pm.007}$	$9.559^{\pm.086}$	$5.566^{\pm.027}$	$2.799^{\pm.072}$
MLD [4]	$0.473^{\pm.013}$	$0.772^{\pm.002}$	$9.724^{\pm.082}$	$3.196^{\pm.010}$	$2.413^{\pm.079}$
STGMD(Ours)	**0.329**$^{\pm.034}$	$0.612^{\pm.006}$	**9.480**$^{\pm.105}$	$5.527^{\pm.028}$	**2.917**$^{\pm.106}$

Table 1: Quantitative results on the HumanML3D test set.

Methods	FID ↓	R Precision ↑ Top 3	Diversity →	Multimodal Dist ↓
With STG-UNet	**0.329**$^{\pm.034}$	**0.612**$^{\pm.006}$	**9.480**$^{\pm.105}$	**5.527**$^{\pm.028}$
Without STG-UNet	$0.564^{\pm.034}$	$0.607^{\pm.017}$	$9.334^{\pm.085}$	$5.617^{\pm.028}$

Table 2: Ablation study: Impact of excluding STG-UNet on FID.

rics, especially, a significant degradation in FID. These results show that utilizing graph neural networks to capture fine-grained local details can benefit the quality of generated samples.

The next question we are interested in is the most efficient STG-UNet architecture. There are two important hyper-parameters in the proposed STG-UNet: the depth and the number of nodes used at each scale. Table 3 presents the performance of STGMD across several variants. Among them, the depth has a more significant impact on the results, and STGMD achieves optimal performance when the depth is set to 3. While the number of the nodes at different scales has a relatively small impact.

Depths	# nodes in each scale	FID ↓
4	21, 16, 11, 6	0.608
4	21, 16, 8, 4	0.553
3	21, 16, 8	0.367
3	21, 10, 5	**0.329**
2	21, 10	0.536
1	21	0.427

Table 3: Ablation study: impact of different depths and number of nodes in each scale on FID.

5 Conclusion

In this work, we investigate the graph neural network based diffusion model in text-driven human motion generation. Compared to existing diffusion-based methods, STGMD can generate more realistic and diverse human motion sequences, highlighting the continued importance of graph network models in human modelling. Moreover, STGMD provides a potential research direction for enabling graph neural networks to handle multimodal data.

References

[1] 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.

[2] Defu Cao, Yujing Wang, Juanyong Duan, Ce Zhang, Xia Zhu, Congrui Huang, Yunhai Tong, Bixiong Xu, Jing Bai, Jie Tong, et al. Spectral temporal graph neural network for multivariate time-series forecasting. *Advances in neural information processing systems*, 33:17766–17778, 2020.

[3] Deli Chen, Yankai Lin, Wei Li, Peng Li, Jie Zhou, and Xu Sun. Measuring and relieving the over-smoothing problem for graph neural networks from the topological view. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 3438–3445, 2020.

[4] 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. *arXiv preprint arXiv:2212.04048*, 2022.

[5] Lingwei Dang, Yongwei Nie, Chengjiang Long, Qing Zhang, and Guiqing Li. Msr-gcn: Multi-scale residual graph convolution networks for human motion prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11467–11476, 2021.

[6] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021.

[7] Joao P Ferreira, Thiago M Coutinho, Thiago L Gomes, José F Neto, Rafael Azevedo, Renato Martins, and Erickson R Nascimento. Learning to dance: A graph convolutional adversarial network to generate realistic dance motions from audio. *Computers & Graphics*, 94:11–21, 2021.

[8] Hongyang Gao and Shuiwang Ji. Graph u-nets. In *international conference on machine learning*, pages 2083–2092. PMLR, 2019.

[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.

[10] Chuan Guo, Xinxin Zuo, Sen Wang, Shihao Zou, Qingyao Sun, Annan Deng, Minglun Gong, and Li Cheng. Action2motion: Conditioned generation of 3d human motions. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2021–2029, 2020.

[11] 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.

[12] Chuan Guo, Xinxin Zuo, Sen Wang, and Li Cheng. Tm2t: Stochastic and tokenized modeling for the reciprocal generation of 3d human motions and texts. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXV*, pages 580–597. Springer, 2022.

[13] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

[14] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

[15] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[16] Junhyun Lee, Inyeop Lee, and Jaewoo Kang. Self-attention graph pooling. In *International conference on machine learning*, pages 3734–3743. PMLR, 2019.

[17] Maosen Li, Siheng Chen, Yangheng Zhao, Ya Zhang, Yanfeng Wang, and Qi Tian. Dynamic multiscale graph neural networks for 3d skeleton based human motion prediction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 214–223, 2020.

[18] Qin Li, Georgia Chalvatzaki, Jan Peters, and Yong Wang. Directed acyclic graph neural network for human motion prediction. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3197–3204. IEEE, 2021.

[19] Yujia Li, Daniel Tarlow, Marc Brockschmidt, and Richard Zemel. Gated graph sequence neural networks. *arXiv preprint arXiv:1511.05493*, 2015.

[20] Dening Lu, Qian Xie, Kyle Gao, Linlin Xu, and Jonathan Li. 3dctn: 3d convolution-transformer network for point cloud classification. *IEEE Transactions on Intelligent Transportation Systems*, 23(12):24854–24865, 2022.

[21] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. Amass: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5442–5451, 2019.

[22] Wei Mao, Miaomiao Liu, Mathieu Salzmann, and Hongdong Li. Learning trajectory dependencies for human motion prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9489–9497, 2019.

[23] Riktim Mondal, Debadyuti Mukherjee, Pawan Kumar Singh, Vikrant Bhateja, and Ram Sarkar. A new framework for smartphone sensor-based human activity recognition using graph neural network. *IEEE Sensors Journal*, 21(10):11461–11468, 2020.

[24] Boris N Oreshkin, Arezou Amini, Lucy Coyle, and Mark Coates. Fc-gaga: Fully connected gated graph architecture for spatio-temporal traffic forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 9233–9241, 2021.

[25] Vardan Papyan, Yaniv Romano, and Michael Elad. Convolutional neural networks analyzed via convolutional sparse coding. *The Journal of Machine Learning Research*, 18(1):2887–2938, 2017.

[26] Mathis Petrovich, Michael J Black, and Gül Varol. Action-conditioned 3d human motion synthesis with transformer vae. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10985–10995, 2021.

[27] Mathis Petrovich, Michael J Black, and Gül Varol. Temos: Generating diverse human motions from textual descriptions. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXII*, pages 480–497. Springer, 2022.

[28] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

[29] Xiangbo Shu, Liyan Zhang, Guo-Jun Qi, Wei Liu, and Jinhui Tang. Spatiotemporal co-attention recurrent neural networks for human-skeleton motion prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(6):3300–3315, 2021.

[30] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015.

[31] Xuxiang Ta, Zihan Liu, Xiao Hu, Le Yu, Leilei Sun, and Bowen Du. Adaptive spatio-temporal graph neural network for traffic forecasting. *Knowledge-Based Systems*, 242: 108199, 2022.

[32] Guy Tevet, Brian Gordon, Amir Hertz, Amit H Bermano, and Daniel Cohen-Or. Motionclip: Exposing human motion generation to clip space. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXII*, pages 358–374. Springer, 2022.

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

[34] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[35] Haomin Wen, Youfang Lin, Yutong Xia, Huaiyu Wan, Roger Zimmermann, and Yuxuan Liang. Diffstg: Probabilistic spatio-temporal graph forecasting with denoising diffusion models. *arXiv preprint arXiv:2301.13629*, 2023.

[36] Zhanghao Wu, Zhijian Liu, Ji Lin, Yujun Lin, and Song Han. Lite transformer with long-short range attention. *arXiv preprint arXiv:2004.11886*, 2020.

[37] Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V Le. Qanet: Combining local convolution with global self-attention for reading comprehension. *arXiv preprint arXiv:1804.09541*, 2018.

[38] Bing Yu, Haoteng Yin, and Zhanxing Zhu. Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. *arXiv preprint arXiv:1709.04875*, 2017.

[39] 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.

[40] Zhen Zhang, Jiajun Bu, Martin Ester, Jianfeng Zhang, Chengwei Yao, Zhi Yu, and Can Wang. Hierarchical graph pooling with structure learning. *arXiv preprint arXiv:1911.05954*, 2019.

[41] Mengyi Zhao, Mengyuan Liu, Bin Ren, Shuling Dai, and Nicu Sebe. Modiff: Action-conditioned 3d motion generation with denoising diffusion probabilistic models. *arXiv preprint arXiv:2301.03949*, 2023.