

Spatiotemporal Graph Latent Diffusion Models for Human Action Synthesis

Bruno Degardin^{1,3,4}, Vasco Lopes^{2,3,4}, João Neves^{2,3}, *Member, IEEE*, Hugo Proença^{1,3}, *Senior Member, IEEE*

¹Instituto de Telecomunicações, ²NOVA LINS, ³University of Beira Interior, Portugal, ⁴DeepNeuronic

bruno.degardin@ubi.pt, vasco.lopes@ubi.pt, jcneves@di.ubi.pt, hugomcp@di.ubi.pt

Abstract—Generating natural human actions and behaviours remains a complex task, encompassing first, the refinement of generated forms; second, the enrichment of their diversity; and third, achieving temporal flexibility. This complexity is especially evident in the task of synthesising lifelike body movements tailored to specific actions, often referred to as action conditioning. This paper introduces GLDM, a new architecture that builds on Latent Diffusion Models (DMs) and Graph Convolutional Networks (GCNs) to synthesise human motions effectively. By decomposing the process of motion formation into a sequence of denoising graph autoencoders (GAEs) within a graph latent space, it is possible to generate unconstrained body movements while improving sample quality, diversity, conditioning and temporal flexibility. The experiments carried out in three well-known datasets show that GLDM notably outperforms state-of-the-art approaches in terms of distribution quality metrics while having the ability to generate over 120 distinct actions and behaviours, including text prompt conditioning. Furthermore, it is demonstrated that the improvements attained in the generation of synthetic behaviours allow the increase of the accuracy of action recognition methods when trained on these data.

Index Terms—Human behaviour analysis, Human action synthesis, Behaviour modelling, Graph convolutional networks, Latent diffusion models, Skeleton-based action recognition.

I. INTRODUCTION

Generating realistic and controllable human motion is a challenging open problem which has broad applications in fields such as motion synthesis, character animation, and human-robot interaction. In the context of biometrics, behaviour, and identity science, these applications extend to areas like gait analysis, activity recognition, and the development of more natural human-computer interfaces. Skeleton-based representations provide a robust characterization of body movement by encoding joint positions over time, making them particularly useful for these applications. However, existing generative models for skeleton-based motion often fail to capture the full complexity and diversity of natural human actions, limiting their effectiveness in these fields.

Existing approaches can be categorized into autoregressive, generative, and diffusion approaches. Autoregressive methods [17], [30], [68] often employ Recurrent Neural Networks (RNNs), treating skeleton data as sequential vectors to generate actions based on observed frames. While individual samples from autoregressive methods exhibit decent quality, these approaches fail to extract optimal intrinsic structural body details through vectorized skeleton sequences and the

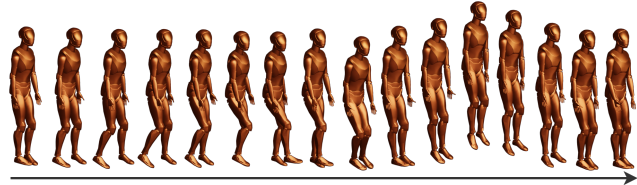
"A person walking"



"A person jumping"



"A person walking and then jumping"



Time

Fig. 1: Text-to-Action Generation. The proposed graph latent diffusion model. GLDM is able to generate over 120 different actions, including text prompt conditioning for action combination (last row).

use of LSTMs [24], GRUs [1], and Seq2Seq [51] potentially restricting scalability concerning temporal flexibility. The use of Generative Adversarial Networks (GANs) [18] for skeleton action generation [2], [3], [13], [60], [62] from a latent space addresses this concern. Yet, many methods still rely on manually structured vector sequences, autoregressive techniques, and Gaussian processes, limiting their adaptability for action conditioning. Additionally, relying on GANs can lead to mode collapse and reduced sample quality. The remarkable achievements of Diffusion Generative Models [47] in the realm of image synthesis [41], [42], [44] fostered the creation of diffusion-based strategies for human motion synthesis [4], [28], [53], [64]. Nevertheless, their reliance on one-dimensional latent representations and conventional convolutions limits generation quality, temporal diversity, and action conditioning. Moreover, most recent approaches operate on SMPL body models [5], [29], [35], generating parametric meshes rather than skeleton graphs. This representation mismatch necessitates additional fitting or retargeting procedures

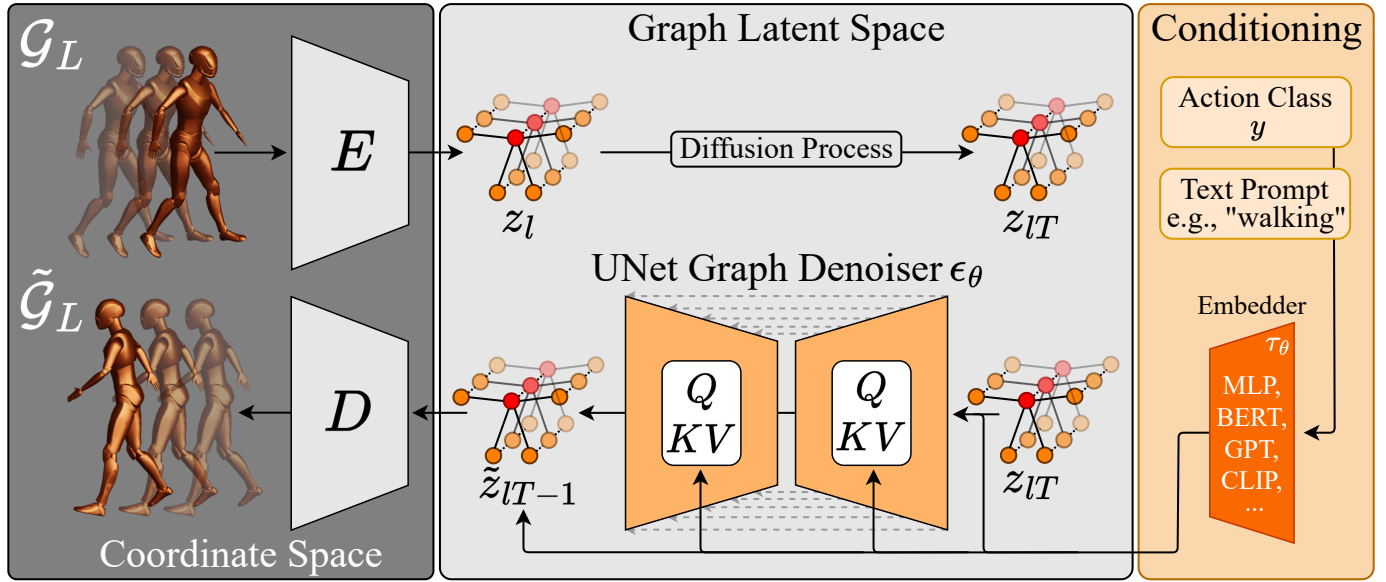


Fig. 2: **Architecture of our proposed GLDM.** A skeleton action $\mathcal{G}_L$ is processed by the graph encoder E , which applies spatial and temporal downsampling from graph resolution level L to l , converting the action into a 2D graph latent representation z_l . During training, T diffusion steps are applied to z_l , where the graph denoiser ϵ_θ learns to predict the denoised latent representation $\tilde{z}_l$. This output is then passed to the graph decoder D to reconstruct the action at its original resolution L . Additionally, GLDM supports conditioning through embedded inputs from a domain-specific encoder τ_θ , using a cross-attention mechanism within ϵ_θ to enhance control over generation.

for comparison, potentially introducing external modeling bias and altering the evaluation protocol [67].

a) : In this paper, we propose Graph Latent Diffusion Model (GLDM) to address the above-mentioned limitations. Our architecture leverages the benefits of DMs and GCNs, such that we generate conditioned human action sequences from a graph latent space to better model spatial and temporal relationships through the diffusion and denoising process. This approach addresses the limitations of existing models and enhances their applicability in fields related to biometrics, behaviour analysis.

b) : Motivated by the extension of convolutions from images to graphs and the application of diffusion processes in the latent space, we combine spatiotemporal graph convolutions across the encoder, decoder, and UNet denoiser to better model skeleton data and exploit the skeleton's inherent graph structure through a 2D graph latent representation of a human skeleton's downsampling resolution. Also, the proposed approach improves the state-of-the-art in action conditioning and text-to-action generation (examples in Fig. 1 and Fig. 6) through a cross-attention mechanism over our graph denoiser layers, whereas previous methods had limited control over desired actions [13], [21], [59], [60], [62]. Such graph latent representation application may also inspire future directions in human behaviour analysis, such as poorly explored areas like human-to-human and human-to-scene interaction generation [57], [58]. Fig. 2 presents an overview of the proposed architecture. In summary, our main contributions are four-fold: (1) A new scalable graph latent diffusion model to generate human actions. (2) A framework that can be extended into a conditional model enabling action and text prompt

conditioning. (3) A graph latent projection based on graph-aware upsample and downsample modules, which increases temporal flexibility. (4) We extensively perform experiments on three datasets: NTU RGB+D [45], NTU-120 RGB+D [31], and Human3.6M [25], demonstrating that GLDM significantly outperforms the state-of-the-art approaches.

II. RELATED WORK

Skeleton-based Behaviour Analysis. Body pose estimation holds great potential for human behaviour analysis. Representing human dynamics as body poses yields semantically rich and descriptive information, minimizing the influence of appearance noise often found in RGB and depth data. This approach enables the learning process to concentrate solely on human behaviour. In the last decade, this analysis progressed from using pseudo-images with CNNs [10], [27], [34], [50] and sequence coordinate vectors with RNNs [15], [32], [33], [48], [65] to the recent adoption of GCNs [6], [7], [9], [11], [12], [37], [46], [61], [66]. The employment of GCNs models skeleton data as spatiotemporal graphs enhances the capacity to capture underlying structural information more effectively.

Autoregressive Models. Inspired by action prediction models, some works [17], [68] rely on autoregressive methods to synthesise human actions based on observed frames. Fragkiadaki *et al.* [17] introduced an encoder-decoder network with LSTM units, capturing temporal dependencies from the low-dimensional representation of the input skeleton. Zhou *et al.* [68] proposed a conditioned LSTM network, generating data that is conditioned at periodic intervals within the sequence.

Generative Models. Generative adversarial networks [18] have inspired human action synthesis methods [30], [59],

[60], [62]. While some approaches still rely on autoregressive techniques [30], [59], [62] or Gaussian processes [60], this choice constrains their temporal flexibility, stochastic variation, and ability for action conditioning. Yan *et al.* [60] proposed a Gaussian process prior to the generator to analyze latent space dimensions, but its heavy computation on latent points correlation limits their diversity. Degardin *et al.* [13] have recently introduced a mapping network to disentangle the latent space. Despite incorporating GCNs, the reliance on GANs is prone to mode collapse and diminished sample quality.

Diffusion Models. The significant success of Diffusion Generative Models [47] in image synthesis [41], [42], [44] inspired diffusion models for human motion synthesis [4], [28], [53], [64]. MotionDiffuse [64] is the first text-based motion DM with detailed instructions for SMPL body parts. MDM [53] introduces DM for raw motion data to capture motion-condition relations. MLD [4] advances motion conditioning by adopting motion latent-based diffusion, aiming to reduce computational demands. Nonetheless, their reliance on 1D latent representations and traditional convolutions hampers generation quality, temporal diversity, and action conditioning.

III. PROPOSED METHOD

GLDM seeks to enhance the quality of generated controllable action samples while retaining their diversity and temporal flexibility via a diffusion process within a graph latent space. This section outlines the proposed approach, starting with an introduction to the adopted GCN, followed by network descriptions and the conditioning mechanisms incorporated into our graph denoiser.

A. Graph-based Autoencoders

Similar to image modelling with DMs, the proposed solution employs an encoder E to map the input from the coordinate space into a graph latent representation, and a decoder D to reconstruct it back to the coordinate space (as depicted in the dark grey section of Fig. 2). Nonetheless, utilizing standard convolutional neural network (CNN) operations is not optimal within the framework of modelling skeletal human actions, since it would lead to spatial and temporal distortions [13], [60]. More precisely, traditional convolutional kernels would struggle to encapsulate crucial structural information inherent in the skeletal data. This challenge arises primarily due to the intricate interdependencies among neighboring joints in the pseudo-image representation.

Thus, we adopt a perceptual graph compression model similar to the concepts employed in perceptual image compression studies [16], [42]. Accordingly, we define $\mathcal{G}_l = (\mathcal{V}_l, \mathcal{E}_l)$ to be a spatiotemporal graph with F_l frames and N_l joints, where $l = \{1, \dots, L\}$, and L represents the skeleton graph's resolution levels. Consequently, the feature map of the skeletal sequence is denoted as $\mathbf{X}_l \in \mathbb{R}^{N_l \times F_l \times C}$, where C denotes the number of channels, which in turn conveys the joint coordinates corresponding to resolution level L . Since a GCN consists of both spatial and temporal graph convolutions, we utilize an adjacency matrix $\mathbf{A}_l \in 0, 1^{N_l \times N_l}$ along with

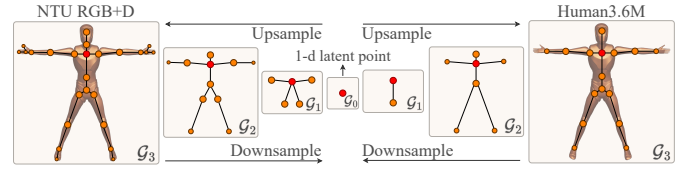


Fig. 3: **Graph upsampling and downsampling paths.** Graph pyramid applied by GLDM over the NTU RGB+D and NTU-120 RGB+D (left side) and Human3.6M (right side) datasets.

the corresponding identity matrix $\mathbf{I}_l$. These matrices help establish connections among intra-body joints, a crucial factor in regulating convolutional receptive fields through spatial neighborhood approaches [61]. To achieve this, a partitioning strategy is employed to define the neighbor set of each joint for constructing convolution operations over the skeleton graph. In this approach, $\mathbf{A}_l$ and $\mathbf{I}_l$ are decomposed into three partitions p , following the spatial configuration proposed by [61], ensuring that $\mathbf{A}_l + \mathbf{I}_l = \sum_p \mathbf{A}_{l_p}$. This partitioning is designed to mimic the structure of image-based convolutions within a graph framework, where adjacency, learnable weight, and degree matrices play roles similar to kernel connectivity, kernel weights, and normalization in CNNs, while explicit graph coarsening and reconstruction operate as pooling and unpooling stages to control spatial resolution. We then apply a learnable weight matrix $\mathbf{M}_l \in \mathbb{R}^{N_l \times N_l}$ (initialized as an all-one matrix) over the layers of both the encoder and decoder to successfully extract multilevel semantic information [8], [46], [61]. The spatial graph convolution employed in this work to dynamically optimize the spatial weighting configuration of $\mathbf{A}_l$ can be described as follows:

$$\mathcal{S}(\mathbf{X}_l) = \sum_{i=1}^p \Lambda_{l_i}^{-\frac{1}{2}} (\mathbf{A}_{l_i} \odot \mathbf{M}_l) \Lambda_{l_i}^{-\frac{1}{2}} \mathbf{X}_l \mathbf{W}_{l_i}, \quad (1)$$

where the degree matrix $\Lambda_{l_p}^{ii} = \sum_j (\mathbf{A}_{l_p}^{ij})$ is computed to normalize the adjacency matrix $\mathbf{A}_{l_p}$ by considering the number of edges connected to each joint node. The stacked weight vectors for each partition group p at resolution level l are denoted as $\mathbf{W}_{l_p}$. And finally, we perform 1D kernels as the temporal graph convolution after the spatial graph convolution giving $\mathcal{T}(\mathcal{S}(\mathbf{X}_l)) = \mathcal{S}(\mathbf{X}_l) * \mathbf{w}_l$, where $\mathbf{w}_l \in \mathbb{R}^{1 \times f' \times C}$ represents the temporal kernel at resolution l , with f' indicating the number of frames to be involved in the convolutional operation of the kernel.

The graph encoder E encodes $\mathbf{X}_L$ into a graph latent representation $z = E(\mathbf{X}_L)$, and the decoder D reconstructs the skeleton action from the graph latent, giving $\hat{\mathbf{X}} = D(z) = D(E(\mathbf{X}_L))$, where $z \in \mathbb{R}^{n_L \times f'_L \times C}$. In this paper, we solely rely on the Mean Squared Error (MSE) and the Kullback-Leibler (KL) loss over the training procedure of our GAE.

Furthermore, aiming to enhance reconstructions, our subsequent Diffusion Model is tailored to function with a 2D structure of the graph latent space (as illustrated in the light grey section of Fig. 2). This concept is also depicted in the graph $\mathcal{G}_1$ in Fig. 3, representing the lowest level of downsampled resolution selected in this paper (check ablation studies in section IV-B1), denoted as l . This strategy harnesses

the potential of our acquired graph latent space $z = E(\mathbf{X}_L)$, contributing to the generation of a wider range of diverse motions. Unlike prior studies [4], [13], [63] that depend on an arbitrary one-dimensional ordering of the learned space z for autoregressively modelling its distribution, our approach preserves the intrinsic structure of z and does not disregard its inherent characteristics. Hence, our compression model preserves details of $\mathbf{X}_L$ better (see Tab. I).

B. Graph Latent Diffusion Models

Diffusion Models [47] are designed to learn a data distribution $p(z_l)$ by gradually denoising a variable with normal distribution. This corresponds to learning the opposite of a fixed T -length Markov Chain. The most effective models in image synthesis [14], [23] use a modified version of the variational lower bound on $p(z_l)$, similar to denoising score-matching [49]. We can think of these models as a series of equally-weighted denoising autoencoders $\epsilon_\theta(z_{lt}, t); t = 1 \dots T$. They are trained to predict a denoised version of their input z_{lt} , where z_{lt} is a noisy variant of the initial input z_l .

While skeleton data may be less computationally demanding than pixel data, employing diffusion on raw spatiotemporal graphs can still be inefficient and demanding in terms of computational resources. Moreover, raw motion data often persist as high-frequency outliers due to motion capture systems [4], [13], potentially exerting an impact on the ability of the diffusion model to effectively learn the training distribution. Considering this and aiming at enhancing generation quality, we carry out the diffusion process of our graph denoiser ϵ_θ within a concise and representative graph latent space specifically designed for motion data and simultaneously more efficient computationally. The diffusion on the graph latent space is modeled as a Markov noising process $\{z_{lt}^{1:F}\}_{t=0}^T$ as follows:

$$q(z_{lt}^{1:F} | z_{l(t-1)}^{1:F}) = \mathcal{N}(\sqrt{\alpha_t} z_{l(t-1)}^{1:F}, (1 - \alpha_t)I), \quad (2)$$

where $z_{l0}^{1:F}$ is drawn from the data distribution and $\alpha_t \in (0, 1)$ denotes constant hyper-parameters for sampling. When α_t is small enough, we can approximate $z_{lt}^{1:F} \sim \mathcal{N}(0, I)$.

Differing from previous methods that rely on compressed discrete latent spaces and autoregressive, attention-based transformer models [4], [39], [60], we build on the graph-specific inductive biases inherent in our graph denoiser ϵ_θ . Which can be carried out by constructing the underlying UNet architecture predominantly using graph convolutional layers (see Fig. 4, where $\mathcal{G}_l$ denotes the graph latent space z_l for simplicity). Additionally, it involves fine-tuning the objective to emphasize the most perceptually significant aspects through the utilization of the reweighted bound, as:

$$L_{GLDM} = \mathbb{E}_{E(\mathbf{X}_L), \epsilon, t} [\|\epsilon - \epsilon_\theta(z_{lt}, \mathbf{A}_l, t)\|_2^2], \quad (3)$$

where $\epsilon \sim \mathcal{N}(0, 1)$ and t is sampled uniformly from $\{1, \dots, T\}$. Also, our graph denoiser ϵ_θ has a time-conditional UNet [43]. Initially, within the training of ϵ_θ , the encoder E compresses the skeleton action into z_1 . Subsequently, during the reverse diffusion phase, ϵ_θ predicts $\tilde{z}_1$ through T iterative denoising

steps, and samples can be decoded to the action space with a single pass through D .

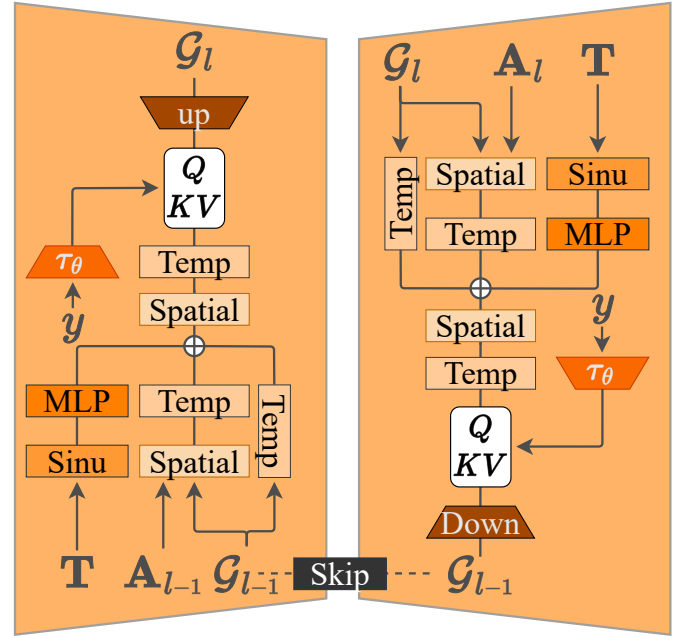


Fig. 4: UNet layers architecture in GLDM. The spatial graph convolution takes as input a graph $\mathcal{G}_l$ latent representation z_l together with its adjacency matrix $\mathbf{A}_l$ to handle inherent structural spatiotemporal information. The timesteps $\mathcal{T}$ of the diffusion process are encoded by a sinusoidal embedder and MLP before being added to the graph convolution.

C. Conditioning Mechanisms

Similar to other diffusion models [42], [44], [53], our GLDM model can also be extended to generate skeleton actions under specific conditions. This is achieved by utilizing the conditional distribution of $p(\mathbf{X}_L|y)$, which can be determined based on class information [13], [20], [38] and language prompts [4], [19], [39]. This can be implemented with a conditional graph denoiser $\epsilon_\theta(z_{lt}, \mathbf{A}_l, t, y)$, which for different conditions, only the learning of ϵ_θ is necessary.

Two conditioning procedures are introduced: class-to-action and text-to-action generation. Instead of only injecting the conditioning embedding information directly to the model [4], [13], [20], we enhance the flexibility of our GLDM as a conditional action generator by integrating the cross-attention mechanism [55] into the underlying UNet backbone at different resolution levels of the skeleton. To achieve this, a domain-specific encoder τ_θ is utilized to convert y into an intermediate representation $\tau_\theta(y) \in \mathbb{R}^{1 \times d_\tau}$. Subsequently, this representation is mapped to the intermediate layers of our graph denoiser ϵ_θ through a cross-attention layer (Fig. 4), to each resolution level l , implementing $\text{Attention}(Q_l, K_l, V_l) = \text{softmax}\left(\frac{Q_l K_l^T}{\sqrt{d}}\right) \cdot V_l$, with:

$$Q_l = W_{Q_l}^{(i)} \cdot \psi_i(z_{lt}), K_l = W_{K_l}^{(i)} \cdot \tau_\theta(y), V_l = W_{V_l}^{(i)} \cdot \tau_\theta(y), \quad (4)$$

where $\psi_i(z_{lt}) \in \mathbb{R}^{N \times d_e^i}$ represents the intermediate representation of the UNet implementing ϵ_θ and $W_{Q_l}^{(i)} \in$

$\mathbb{R}^{d \times d_\tau}$, $W_{K_l}^{(i)} \in \mathbb{R}^{d \times d_\tau}$, $W_{V_l}^{(i)} \in \mathbb{R}^{d \times d_\epsilon^i}$ are learnable projection matrices [26], [55] at resolution level l . Finally, our conditional GLDM is learned with the following objective:

$$L_{GLDM} = \mathbb{E}_{E(\mathbf{x}_L), y, \epsilon, t} [\|\epsilon - \epsilon_\theta(z_{lt}, \mathbf{A}_l, t, \tau_\theta(y))\|_2^2], \quad (5)$$

where τ_θ is freed for the text-to-action task as previously suggested [39] to optimize ϵ_θ via this objective. τ_θ can be parametrized with action domain-specific experts, which also increases its flexibility (see Section IV-B3).

IV. EXPERIMENTS AND DISCUSSION

In this section, we conduct thorough experimental evaluations to validate the proposed approach across four datasets: NTU RGB+D [45], NTU-120 RGB+D [31], Human3.6M [25], and NTU-2D RGB+D [45]. Initially, ablation studies are conducted to assess the effectiveness of different properties and configurations of GLDM. Following that, we quantitatively and qualitatively compare our best-performing model with state-of-the-art approaches. Finally, we also dive into the advantages of utilizing synthesised data to enhance skeleton-based action recognition.

A. Datasets and evaluation metrics

NTU RGB+D [45]. The dataset comprises 56,880 videos that belong to 60 action classes. For each sample, 3D skeleton data captured from $N_l = 25$ joints of 40 volunteers is provided across $F_l = 300$ frames. The authors suggest two benchmarks for evaluating models: 1) cross-subject, where models are tested on other subjects than the ones they were trained on, and 2) cross-view, where models are tested on different camera views than the ones they were trained on. A curated version of this dataset is also used for a fair comparison in Table III, denoted by NTU-2D RGB+D.

NTU-120 RGB+D [31]. The dataset is an expanded iteration of its previous version, encompassing 114,480 video samples that are categorized into 120 action classes. The videos were captured using three camera views in 32 distinct setups, with data collected from 106 volunteers and $N_l = 25$ body joints per skeleton across $F_l = 300$ frames. The authors suggested two benchmarks for evaluation: 1) cross-subject, where models are tested on other subjects than the ones they were trained on, and 2) cross-setup, where models are tested on different setups than the ones they were trained on.

Human3.6M [25]. This dataset is a more simplistic set with 2D human motions and $N_l = 15$ body joints for each skeleton across $F_l = 1024$ frames. For a fair comparison between the state-of-the-art, the same pre-processing and settings as [59], [62] were followed with corresponding 10 action classes.

Evaluation Metrics. Two evaluation metrics are used for estimating the quality of synthetic samples. Similar to image modelling, the Fréchet Inception Distance (FID) [22] is computed using InceptionV3 [52], extracting 2048-dimensional pool3 features. Skeleton sequences are reshaped into image-like multi-channel tensors (without RGB rendering), truncated to 64 frames, and evaluated at the sequence level without per-frame aggregation. FID measures the distance between

the real and synthesised data distributions in this feature space. Additionally, the Maximum Mean Discrepancy (MMD) is exploited between real and synthetic samples based on a two-sample test to measure the discrepancy between both distributions. The MMD over motion dynamics, denoted as MMD_a , corresponds to the average MMD across each frame and reflects the smoothness of motion, where lower values indicate less jitter or unrealistic frame-to-frame changes. The MMD over whole sequences, denoted as MMD_s , captures the overall coherence of movements across the entire sequence. For joints over 2D space, only the MMD is used.

B. Action Synthesis

1) *Ablation study:* Prior to delving into comparisons with state-of-the-art methods, we initially demonstrate through experimental evidence that the properties of GLDM significantly enhance the quality of generated samples.

Method		FID	MMD _a	MMD _s	FID	MMD _a	MMD _s
NTU RGB+D	<i>Cross-Subject</i>						
	c-GAN [36]	27.480	0.919	0.975	31.875	0.993	1.088
	CSGN [60]	6.030	0.873	0.954	7.114	0.910	0.991
	Kinetic-GAN [13]	3.618	0.772	0.871	4.235	0.824	0.913
	A Vanilla LDM	4.053	0.796	0.893	4.730	0.851	0.932
	B + GCNs	2.754	0.734	0.801	3.498	0.781	0.856
	C + Conditional	2.433	0.712	0.790	3.233	0.777	0.879
	D + Cross Attn	2.231	0.693	0.752	3.101	0.734	0.842
		<i>Cross-View</i>					

TABLE I: **Evaluating different architecture designs.** We compare FID and MMD scores (lower is better) between real and synthetic samples under global body movements.

Table I compares FID and MMD metrics across distinct architectural designs in the NTU RGB+D benchmarks. The evaluation assesses each distribution within the specific configurations of each method. Although Kinetic-GAN [13] employs graph convolutions for synthesizing actions, a noteworthy observation is the competitive performance achieved by our Vanilla LDM (without graph convolutions). This underscores the effectiveness of diffusion models in contrast to GAN-based methods for action synthesis. GAN-based approaches often face challenges such as mode dropping or mode collapse when generating human actions.

We start by employing an unconditional Vanilla Latent Diffusion Model (A) with a universal null token $\emptyset$. The rationale is to prevent introducing biased performance via conditioned samples, which have been proven to improve the generation quality significantly [36]. Subsequently, we validate the significance of implementing graph convolutions (B) across our encoder E , decoder D , and denoiser ϵ_θ , which clearly improves sample quality, mainly by capturing essential structural information during both the latent projection and diffusion process. Furthermore, we also introduce a conditioned configuration (C), where the action class is incorporated into the noise generation process. This adaptation enables the generation of desired actions while concurrently enhancing sample quality in comparison to configuration (B). Finally, we also include the cross-attention module into our graph denoiser

ϵ_θ (D), where we distinctly overcome previous baselines and configurations due to its contextual information across the graph structure.

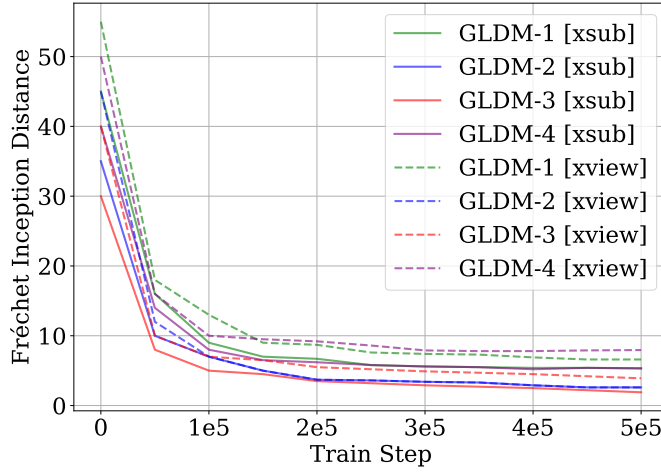


Fig. 5: **Class-conditional GLDM training with different downsampling factors f** over 500k train steps on the NTU RGB+D [45] dataset.

The proposed GLDM approach is founded on the principle of mapping skeletal actions onto a 2-dimensional latent space through a coarsening representation of the initial graph structure. In this section, we investigate the performance of our GLDM under varying downsampling factors $f \in \{1, 2, 3, 4\}$, denoted as GLDM- f , where GLDM-1 represents coordinate-based diffusion models and GLDM-4 denotes 1-dimension latent representation. To ensure a consistent evaluation, we train all models using an identical number of steps and parameters. Fig. 5 illustrates the evolution of sample quality throughout 500 thousand training steps for class-conditional models across the NTU RGB+D dataset [45]. It is observed that a lower downsampling factor (GLDM-1) and a higher downsampling factor (GLDM-4) lead to a slower advancement in training progress, and both models eventually reach a plateau in terms of fidelity. This phenomenon can be explained by the fact that, with a minor downsampling factor, a significant portion of the compression is handled by the diffusion model. Conversely, with a larger downsampling factor, excessive compression occurs, leading to loss of information and consequently imposing restrictions on the attainable quality. GLDM- $\{2, 3\}$ achieve a favorable equilibrium between convergence and perceptual accuracy. This is evident from the substantial difference of 5.5 units separating the models GLDM-1 and GLDM-3 over the cross-subject (xsub) benchmark.

2) *Action Conditioning*: Within this section, the experiments are categorized into local and global movement settings based on whether the 3D locations were normalized into 2D space or retained in their original form, respectively. In addition to the ablation studies presented in Section IV-B1, which were conducted under the context of global movement settings, Table II reports the results achieved by GLDM on the NTU-120 RGB+D dataset, comprising 120 distinct actions. While successfully attaining notably improved sample quality compared to preceding works, our approach not only

enables the generation of 120 different actions but also enables the combination of these class-specific actions (refer to Fig. 1) through our text-to-action conditioning mechanism. It is important to note that, to the best of our knowledge, this capability addresses a challenge that previous GAN-based methods for skeleton-based action synthesis have encountered. Here, we use the original textual action descriptions from the dataset, with the text encoding performed by a frozen CLIP model (denoted as τ_θ) as detailed in Section IV-B3, to ensure a fair comparison with fixed action-conditioned models.

Method		FID	MMD _a	MMD _s	FID	MMD _a	MMD _s
NTU-120	<i>Cross-Subject</i>			<i>Cross-Setup</i>			
	c-GAN [36]	54.403	1.037	1.104	58.531	1.082	1.141
	Kinetic-GAN [13]	5.967	0.819	0.906	6.751	0.847	0.934
	GLDM	3.729	0.701	0.830	4.182	0.791	0.852

TABLE II: **Global movement generation results** on NTU-120 RGB+D [31] with 120 different actions.

Since most state-of-the-art methods are still limited to local movement settings, we also present the results under the same conditions for a fair comparison, where 3D locations are projected into 2D space and normalized positions. We follow the same settings as previous methods [21], [59], [62] and report the results on Table III. On both datasets, we verify the superiority of our model by achieving state-of-the-art performance with a large margin.

Method		MMD _a		MMD _s	
Human3.6M	E2E [56]	0.991		0.805	
	EPVA [56]	0.996		0.806	
	adv-EPVA [56]	0.977		0.792	
	SkeletonVAE [21]	0.452		0.467	
	SkeletonGAN [3]	0.419		0.419	
	c-SkeletonGAN [59]	0.195		0.218	
	c-GAN [36]	0.161		0.187	
	SA-GCN [62]	0.146		0.134	
	Kinetic-GAN [13]	0.071		0.082	
	GLDM	0.034		0.062	
		MMD _a	MMD _s	MMD _a	MMD _s
NTU-2D RGB+D	<i>Cross-Subject</i>		<i>Cross-View</i>		
	SkeletonVAE [21]	0.992	1.136	1.079	1.205
	SkeletonGAN [3]	0.698	0.788	0.999	1.311
	c-SkeletonGAN [59]	0.338	0.402	0.371	0.398
	c-GAN [36]	0.334	0.354	0.365	0.373
	SA-GCN [62]	0.285	0.299	0.316	0.335
	Kinetic-GAN [13]	0.256	0.273	0.295	0.310
	GLDM	0.131	0.149	0.157	0.178

TABLE III: **Local movement generation results**. We compare FID and MMD scores (lower is better) between real and synthetic samples under local body movements.

Additionally, Table III provides a comparison of GLDM with previous methods under local movement settings while maintaining uniform conditions to ensure fairness in the assessment. In these comparisons, the 3D locations were projected into 2D space and subjected to normalized positions. We follow identical settings as utilized in prior methods [13],

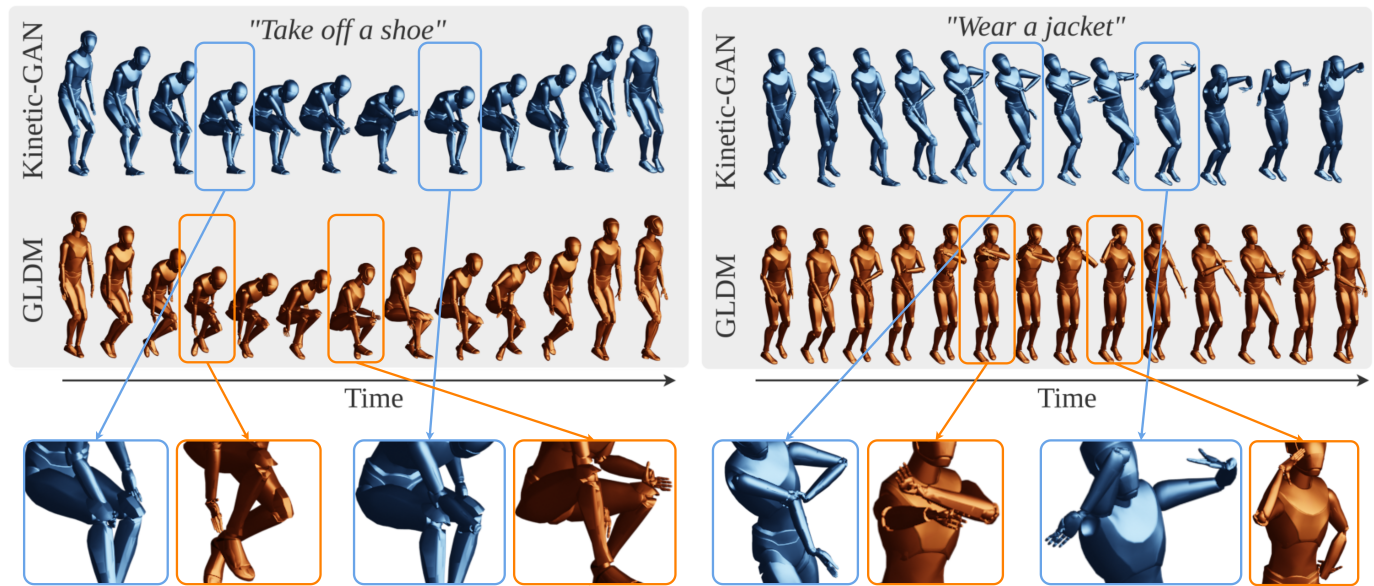


Fig. 6: **Qualitative comparison** of GLDM (orange) with a GAN-based baseline (blue) across detailed and sophisticated synthetic movements.

[21], [59], [62]. Across both the Human3.6M and NTU-2D RGB+D datasets, we verify the superiority of GLDM over existing approaches by a substantial margin.

3) *Text-to-Motion Synthesis*: The introduction of graph latent diffusion models guided by language prompts enables us to explore class action-unconstrained conditional generation. Unless specified otherwise, all experiments in this paper were conducted utilizing a frozen CLIP [40] model as τ_θ , which encodes the text prompt into an intermediate representation denoted as $\tau_\theta(y) \in \mathbb{R}^{1 \times 768}$. This allows us to not only combine different actions but also to achieve high text prompt matching, as illustrated in Fig. 1. For a fair comparison in Fig. 6, considering our baseline [13] is solely class-conditioned, we utilize the very same input text prompt that GLDM was trained with. Our generated results clearly depict the detailed and sophisticated movements achieved by our GLDM, outperforming the baseline. This distinction is evident even in complex actions like “taking off a shoe”, where generating both arms and legs in a precise motion is necessary.

Furthermore, the adoption of a 2-dimensional graph latent representation also enables the reduction of physically implausible body movements, such as limbs’ articulation or their passage through the body, as illustrated in the “wear a jacket” action. This beneficial effect primarily stems from the preservation of structural information throughout the latent space and denoising process.

4) *Unconstrained Temporal Action Length*: Acquiring skeleton-based data is a demanding task, which leads to the current challenging datasets derived from real-world scenarios containing a limited amount of data and short action sequences (10 seconds or less). Consequently, modelling long-term human actions remains a substantial challenge in the domain of action synthesis. The temporal scope of the dataset still restricts current models, such as GAN-based [13], [60] and even recent DM-based approaches [4], [63]. Many of these models face challenges in generating longer actions and effec-

tively establishing concise spatiotemporal connections within the motion.

For these reasons, GLDM introduces graph latent representations where joint relationships are explicitly modeled using a learnable spatial adjacency matrix A_l at each resolution level l . This structured representation allows the graph denoiser ϵ_θ to extract spatial context information effectively, capturing meaningful dependencies between joints.

A key innovation of our approach is the use of 1D temporal kernel convolutions, which dynamically adjust the temporal resolution of the graph latent representations. Since these convolutions operate within a frozen graph denoiser, temporal extension is achieved by directly sampling Gaussian noise of shape $[C, V, L]$ and denoising the full L -frame tensor in a single diffusion trajectory. No frames are repeated, interpolated, or autoregressively stitched. This allows GLDM to generate longer action sequences beyond the original dataset constraints while maintaining motion consistency. During training, sequences shorter than 64 frames are zero-padded, and temporal convolutions use standard zero-padding. The same padding scheme is preserved at inference, ensuring consistent numerical behavior when generating longer sequences.

Moreover, the same process can be reversed to reduce the temporal length of an action while preserving its structural integrity. By leveraging this adaptive temporal modelling, GLDM achieves unconstrained temporal flexibility, making it capable of generating both extended and compressed versions of human motions without altering the core learned model.

As shown in Fig. 7, we evaluate our GLDM trained over only 64 frames of NTU RGB+D and NTU-120 RGB+D, where we observe that the generation quality stabilizes for lengths exceeding 128 frames. Despite the slightly reduced sample quality observed between 64 and 128 frames, this trend is justified by considering the average action execution length of 64 frames in both the NTU RGB+D and NTU-120 RGB+D datasets. Frames beyond the evaluated temporal

window are masked during normalization to ensure consistent metric computation across different sequence lengths. Nevertheless, during the training phase, GLDM exclusively analyzed complete actions comprising 64 frames. Consequently, the model is capable of generating extended synthetic actions, which can also be positively regarded. This ability highlights GLDM's capacity to manage instances of "collapsed" actions without significantly compromising quality, a limitation that previous works struggled with.

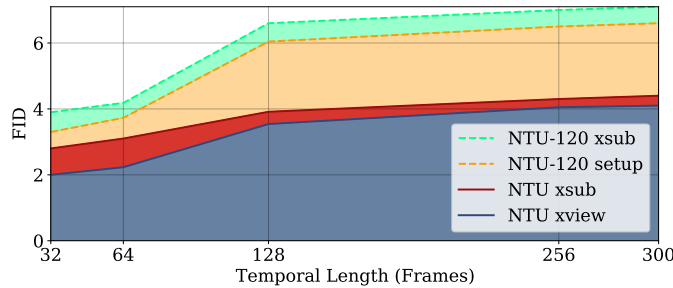


Fig. 7: **Temporal action length flexibility** on NTU RGB+D [45] and NTU-120 RGB+D [31], where GLDM was only trained with 64 frames.

C. Action Recognition

One of the most promising applications of skeleton-based action synthesis is improving action recognition models to mitigate the heavy annotation process. Nonetheless, widely-used benchmarks like those presented in [31], [45] have been acknowledged to contain major annotation errors, as discussed in [54]. These inaccuracies have influenced prior endeavors in skeleton-based action synthesis, leading to the presence of artifacts and improperly generated samples.

To further corroborate the generation quality of GLDM, we opted to assess the performance on the top-5 most challenging actions and the top-5 actions that are comparatively easier to recognize within the NTU RGB+D. In each individual class, we eliminate 90% of the samples, retaining the remaining samples of the other action classes (59 classes). Subsequently, we focus on generating data for the reduced selected class and generate an equivalent quantity to the original training set (approximately 600 samples for each class), referred as "Same synthetic quantity". In our experimental setup, we employ the ST-GCN [61] model for training, utilising the new dataset to enhance action recognition. Additionally, we conduct an experiment involving the generation of 5000 synthetic samples for the respective class. The performance obtained by the action recognition model in each experiment is illustrated in Fig. 8. This evaluation under contrasting difficulty levels allows us to assess augmentation robustness beyond class-specific bias. A noteworthy observation is that, even when using the same number of synthetic samples as the original samples, we achieve competitive performance when evaluating over the test set, particularly in challenging action classes (the last five on the right). While it is expected that generating such a limited number of samples may lead to some loss in motion diversity, the results demonstrate that GLDM effectively captures key action dynamics, enabling

robust recognition. Furthermore, when increasing the number of synthetic samples to 5000 per class, we observe a consistent improvement, in some cases even matching or surpassing the original dataset's performance. This highlights the high quality and generalization ability of the generated motion, reinforcing GLDM's potential for augmenting action recognition models.

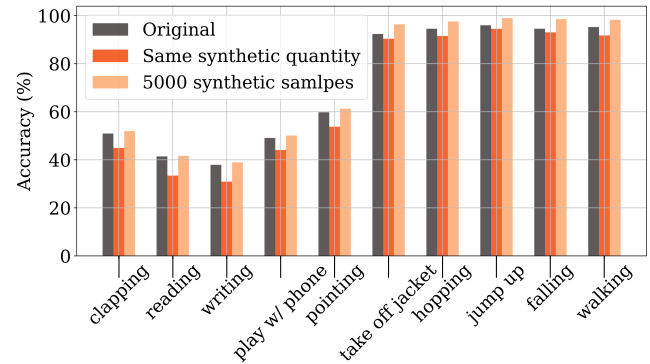


Fig. 8: **Improving action recognition** of ST-GCN [61] on NTU RGBD [45]. The first and the last five classes are the most and less challenging classes, respectively, according to original performances.

V. CONCLUSIONS AND FURTHER WORK

This paper presents a novel Graph Latent Diffusion Model for human action synthesis. By generating human behaviours from a 2D graph latent space, we effectively capture the spatiotemporal connections of motions while preserving their intrinsic structure and characteristics, enabling temporal flexibility. Moreover, the incorporation of graph convolutional networks along with a cross-attention mechanism enables conditional generation while maintaining the integrity of the skeleton structure spatially and temporally. Therefore, the proposed approach enables the generation of a diverse range of 120 distinct actions and even combine them using text-to-action conditioning. This achievement, particularly in the context of global movement settings, posed a challenge for previous methods as far as we know. The proposed approach was assessed on NTU RGB+D, NTU-120 RGB+D, and Human3.6M datasets, achieving a significant quantitative and qualitative improvement over the state-of-the-art performance.

ACKNOWLEDGMENTS

This work was partially supported by the FCT/MEC through National Funds and by the FEDER-PT2020 Partnership Agreement under the Project CENTRO-01-0247-FEDER-113023 - DeepNeuronic. Also, it was supported by FCT/MCTES through national funds and co-funded by EU funds under the project UIDB/50008/2025, Instituto de Telecomunicações. This research was also supported by 'FCT - Fundação para a Ciência e Tecnologia' through the research grant 'UI/BD/150765/2020' and '2020.04588.BD'. This work was also supported by UID/04516/NOVA Laboratory for Computer Science and Informatics (NOVA LINCIS) with the financial support of FCT/IP. The authors acknowledge the use of AI tools to enhance the manuscript for clarity and readability.

REFERENCES

- [1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *ICLR*, 2015. 1
- [2] Emad Barsoum, John Kender, and Zicheng Liu. Hp-gan: Probabilistic 3d human motion prediction via gan. In *CVPR workshops*, 2018. 1
- [3] Haoye Cai, Chunyan Bai, Yu-Wing Tai, and Chi-Keung Tang. Deep video generation, prediction and completion of human action sequences. In *ECCV*, 2018. 1, 6
- [4] Xin Chen, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, and Gang Yu. Executing your commands via motion diffusion in latent space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18000–18010, 2023. 1, 3, 4, 7
- [5] Xin Chen, Anqi Pang, Wei Yang, Yuexin Ma, Lan Xu, and Jingyi Yu. Sportscap: Monocular 3d human motion capture and fine-grained understanding in challenging sports videos. *International Journal of Computer Vision*, 129:2846–2864, 2021. 1
- [6] Yuxin Chen, Ziqi Zhang, Chunfeng Yuan, Bing Li, Ying Deng, and Weiming Hu. Channel-wise topology refinement graph convolution for skeleton-based action recognition. In *ICCV*, 2021. 2
- [7] Zhan Chen, Sicheng Li, Bing Yang, Qinghan Li, and Hong Liu. Multi-scale spatial temporal graph convolutional network for skeleton-based action recognition. In *AAAI*, 2021. 2
- [8] Ke Cheng, Yifan Zhang, Xiangyu He, Weihan Chen, Jian Cheng, and Hanqing Lu. Skeleton-based action recognition with shift graph convolutional network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 183–192, 2020. 3
- [9] Hyung-gun Chi, Myoung Hoon Ha, Seunggeun Chi, Sang Wan Lee, Qixing Huang, and Karthik Ramani. Infogcn: Representation learning for human skeleton-based action recognition. In *CVPR*, 2022. 2
- [10] Seokeon Choi, Jonghee Kim, Wonjun Kim, and Changick Kim. Skeleton-based gait recognition via robust frame-level matching. *IEEE Transactions on Information Forensics and Security*, 14(10):2577–2592, 2019. 2
- [11] Bruno Degardin, Vasco Lopes, and Hugo Proença. Regina—reasoning graph convolutional networks in human action recognition. *IEEE Transactions on Information Forensics and Security*, 2021. 2
- [12] Bruno Degardin, Vasco Lopes, and Hugo Proença. Atom: Self-supervised human action recognition using atomic motion representation learning. *Image and Vision Computing*, page 104750, 2023. 2
- [13] Bruno Degardin, Joao Neves, Vasco Lopes, João Brito, Ehsan Yaghoobi, and Hugo Proença. Generative adversarial graph convolutional networks for human action synthesis. In *WACV*, pages 1150–1159, 2022. 1, 2, 3, 4, 5, 6, 7
- [14] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 4
- [15] Yong Du, Wei Wang, and Liang Wang. Hierarchical recurrent neural network for skeleton based action recognition. In *CVPR*, 2015. 2
- [16] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021. 3
- [17] Katerina Fragkiadaki, Sergey Levine, Panna Felsen, and Jitendra Malik. Recurrent network models for human dynamics. In *CVPR*, 2015. 1, 2
- [18] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *arXiv:1406.2661*, 2014. 1, 2
- [19] 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. 4
- [20] Chuan Guo, Xinxin Zuo, Sen Wang, Shihao Zou, Qingyao Sun, Annan Deng, Minglong 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. 4
- [21] Ikhsanul Habibie, Daniel Holden, Jonathan Schwarz, Joe Yearsley, and Taku Komura. A recurrent variational autoencoder for human motion synthesis. In *BMVC*, 2017. 2, 6
- [22] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *NeurIPS*, 2017. 5
- [23] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 4
- [24] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 1
- [25] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, 2013. 2, 5
- [26] Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *International conference on machine learning*, pages 4651–4664. PMLR, 2021. 5
- [27] Qihong Ke, Mohammed Bannamoun, Senjan An, Ferdous Sohel, and Farid Boussaid. A new representation of skeleton sequences for 3d action recognition. In *CVPR*, 2017. 2
- [28] Jihoon Kim, Jiseob Kim, and Sungjoon Choi. Flame: Free-form language-based motion synthesis & editing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 8255–8263, 2023. 1, 3
- [29] Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. Vibe: Video inference for human body pose and shape estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5253–5263, 2020. 1
- [30] Jogendra Nath Kundu, Maharshi Gor, Phani Krishna Uppala, and Venkatesh Babu Radhakrishnan. Unsupervised feature learning of human actions as trajectories in pose embedding manifold. In *WACV*, 2019. 1, 2, 3
- [31] Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C Kot. Ntu rgb+d 120: A large-scale benchmark for 3d human activity understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(10):2684–2701, 2019. 2, 5, 6, 8
- [32] Jun Liu, Amir Shahroudy, Dong Xu, and Gang Wang. Spatio-temporal lstm with trust gates for 3d human action recognition. In *European conference on computer vision*, pages 816–833. Springer, 2016. 2
- [33] Jun Liu, Gang Wang, Ping Hu, Ling-Yu Duan, and Alex C Kot. Global context-aware attention lstm networks for 3d action recognition. In *CVPR*, 2017. 2
- [34] Mengyuan Liu, Hong Liu, and Chen Chen. Enhanced skeleton visualization for view invariant human action recognition. *Pattern Recognition*, 68:346–362, 2017. 2
- [35] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 851–866, 2023. 1
- [36] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014. 5, 6
- [37] Honghu Pan, Yongyong Chen, Tingyang Xu, Yunqi He, and Zhenyu He. Toward complete-view and high-level pose-based gait recognition. *IEEE Transactions on Information Forensics and Security*, 18:2104–2118, 2023. 2
- [38] Xue Bin Peng, Ze Ma, Pieter Abbeel, Sergey Levine, and Angjoo Kanazawa. Amp: Adversarial motion priors for stylized physics-based character control. *ACM Transactions on Graphics (ToG)*, 40(4):1–20, 2021. 4
- [39] Mathis Petrovich, Michael J Black, and Gül Varol. Temos: Generating diverse human motions from textual descriptions. In *European Conference on Computer Vision*, pages 480–497. Springer, 2022. 4, 5
- [40] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 7
- [41] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022. 1, 3
- [42] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1, 3, 4
- [43] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015: 18th International Conference, Munich, Germany, October 5–9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 4
- [44] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 1, 3, 4
- [45] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+d: A large scale dataset for 3d human activity analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1010–1019, 2016. 2, 5, 6, 8

- [46] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12026–12035, 2019. [2](#), [3](#)
- [47] 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. [1](#), [3](#), [4](#)
- [48] Sijie Song, Cuiling Lan, Junliang Xing, Wenjun Zeng, and Jiaying Liu. An end-to-end spatio-temporal attention model for human action recognition from skeleton data. In *AAAI*, 2017. [2](#)
- [49] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. [4](#)
- [50] Tae Soo Kim and Austin Reiter. Interpretable 3d human action analysis with temporal convolutional networks. In *CVPR Workshops*, 2017. [2](#)
- [51] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *NeurIPS*, 2014. [1](#)
- [52] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, 2016. [5](#)
- [53] 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. [1](#), [3](#), [4](#)
- [54] Neel Trivedi, Anirudh Thatipelli, and Ravi Kiran Sarvadevabhatla. Ntu-x: an enhanced large-scale dataset for improving pose-based recognition of subtle human actions. In *Proceedings of the Twelfth Indian Conference on Computer Vision, Graphics and Image Processing*, pages 1–9, 2021. [8](#)
- [55] 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. [4](#), [5](#)
- [56] Ruben Villegas, Dumitru Erhan, Honglak Lee, et al. Hierarchical long-term video prediction without supervision. In *ICML*, 2018. [6](#)
- [57] Jingbo Wang, Sijie Yan, Bo Dai, and Dahua Lin. Scene-aware generative network for human motion synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12206–12215, 2021. [2](#)
- [58] Zan Wang, Yixin Chen, Tengyu Liu, Yixin Zhu, Wei Liang, and Siyuan Huang. Humanise: Language-conditioned human motion generation in 3d scenes. *Advances in Neural Information Processing Systems*, 35:14959–14971, 2022. [2](#)
- [59] Zhenyi Wang, Ping Yu, Yang Zhao, Ruiyi Zhang, Yufan Zhou, Junsong Yuan, and Changyou Chen. Learning diverse stochastic human-action generators by learning smooth latent transitions. In *AAAI*, 2020. [2](#), [3](#), [5](#), [6](#)
- [60] Sijie Yan, Zhizhong Li, Yuanjun Xiong, Huahan Yan, and Dahua Lin. Convolutional sequence generation for skeleton-based action synthesis. In *CVPR*, 2019. [1](#), [2](#), [3](#), [4](#), [5](#), [7](#)
- [61] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018. [2](#), [3](#), [8](#)
- [62] Ping Yu, Yang Zhao, Chunyuan Li, Junsong Yuan, and Changyou Chen. Structure-aware human-action generation. In *ECCV*, 2020. [1](#), [2](#), [3](#), [5](#), [6](#)
- [63] Ye Yuan, Jiaming Song, Umar Iqbal, Arash Vahdat, and Jan Kautz. Physdiff: Physics-guided human motion diffusion model. *arXiv preprint arXiv:2212.02500*, 2022. [4](#), [7](#)
- [64] 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](#), [3](#)
- [65] Pengfei Zhang, Cuiling Lan, Junliang Xing, Wenjun Zeng, Jianru Xue, and Nanning Zheng. View adaptive recurrent neural networks for high performance human action recognition from skeleton data. In *ICCV*, 2017. [2](#)
- [66] Pengfei Zhang, Cuiling Lan, Wenjun Zeng, Junliang Xing, Jianru Xue, and Nanning Zheng. Semantics-guided neural networks for efficient skeleton-based human action recognition. In *CVPR*, 2020. [2](#)
- [67] Wanrong Zheng, Haidong Zhu, Zhaoheng Zheng, and Ram Nevatia. Gaitstr: Gait recognition with sequential two-stream refinement. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2024. [2](#)
- [68] Yi Zhou, Zimo Li, Shuangjiu Xiao, Chong He, Zeng Huang, and Hao Li. Auto-conditioned recurrent networks for extended complex human motion synthesis. In *ICLR*, 2018. [1](#), [2](#)



Bruno Degardin received the B.Sc. and M.Sc. degrees in 2018 and 2020, respectively, from Universidade da Beira Interior (UBI). He is currently a Ph.D. candidate at UBI with his research interest focused on Computer Vision, emphasising biometrics, human behaviour analysis and video understanding. In the past, Bruno has worked on research projects, such as BIODI, CovidSight and PAPSE-UBI. Since 2020, he has been a Teaching Assistant at UBI and Co-founder & CTO of DeepNeuronic, a company that develops Vision AI products to automate daily processes. Receiving several awards, such as Best M.Sc. Thesis Fraunhofer Portugal Challenge, research of practical utility (Winner at Biomedical World Hackathon, UBIMedical Health Cup, honourable mentions from APDC and winner of WSA Portugal Awards).



as a researcher. Vasco has worked on research projects, such as uPATO, RobotChain, INDTECH 4.0 and CovidSight. Vasco received several awards, such as the APRP Best PhD Thesis in Pattern Recognition 2024/2025 Award and the ICANN21 "1st Springer & ENNS Best Paper Award".

Vasco Lopes received the B.Sc., M.Sc. and Ph.D. degrees in Computer Science and Engineering in 2017, 2019 and 2023 respectively. He is currently an Invited Professor at University of Beira Interior and since 2020, he is the CEO and Co-founder of DeepNeuronic, a company that develops Computer Vision solutions to automate daily processes. In 2021, he collaborated with Huawei R&D Center in Paris as a Research Assistant, where he developed Neural Architecture Search methods for network reliability, and in 2022 in Google Research in Zurich



conferences. He is an Associate Editor of IEEE Access journal, and he has been distinguished with the Best Reviewer Award of the IJCB 2023.

João Neves received the MSc and PhD in Computer Science from the University of Beira Interior, in 2013 and 2018, respectively. Since 2020, he has been affiliated with NOVA Laboratory for Computer Science and Informatics (NOVA LINCIS), University of Beira Interior, where he is an assistant professor. His research interests are focused on computer vision, pattern recognition, and machine learning, particularly in the areas of Video Surveillance and Biometrics. He is the author of more than 40 scientific articles published in international journals and conferences. He is an Associate Editor of IEEE Access journal, and he has been distinguished with the Best Reviewer Award of the IJCB 2023.



Vision Computing, IEEE ACCESS, and International Journal of Biometrics. Also, he served as a Guest Editor of Special Issue of the Pattern Recognition Letters, Image and Vision Computing, and Signal, and Image and Video Processing journals.

Hugo Proença Hugo Proença (Senior Member, IEEE) received the B.Sc., M.Sc., and Ph.D. degrees in 2001, 2004, and 2007, respectively. He is currently a Full Professor at the Department of Computer Science, University of Beira Interior, Portugal, that has been researching mainly about biometrics and visual surveillance. He was the Coordinating Editor of the IEEE Biometrics Council Newsletter and the Area Editor (ocular biometrics) of the IEEE BIOMETRICS COMPENDIUM journal. He is a member of the Editorial Board of the Image and Vision Computing, IEEE ACCESS, and International Journal of Biometrics. Also, he served as a Guest Editor of Special Issue of the Pattern Recognition Letters, Image and Vision Computing, and Signal, and Image and Video Processing journals.