GRPose: Learning Graph Relations for Human Image Generation with Pose Priors

Xiangchen Yin[1,5], Donglin Di[2], Lei Fan[3], Hao Li[2], Wei Chen[2*], Gouxiaofei[2], Yang Song[3], Xiao Sun[4], Xun Yang[1*]

[1]University of Science and Technology of China
[2]Space AI, Li Auto
[3]University of New South Wales
[4] Hefei University of Technology
[5] Institute of Artificial Intelligence, Hefei Comprehensive National Science Center
yinxiangchen@mail.ustc.edu.cn {didonglin,lihao43,chenwei10,gouxiaofei}@lixiang.com
{lei.fan1,yang.song1}@unsw.edu.au sunx@hfut.edu.cn xyang21@ustc.edu.cn

Abstract

Recent methods using diffusion models have made significant progress in human image generation with various control signals such as pose priors. However, existing efforts are still struggling to generate high-quality images with consistent pose alignment, resulting in unsatisfactory output. In this paper, we propose a framework that delves into the graph relations of pose priors to provide control information for human image generation. The main idea is to establish a graph topological structure between the pose priors and latent representation of diffusion models to capture the intrinsic associations between different pose parts. A Progressive Graph Integrator (PGI) is designed to learn the spatial relationships of the pose priors with the graph structure, adopting a hierarchical strategy within an Adapter to gradually propagate information across different pose parts. Besides, a pose perception loss is introduced based on a pretrained pose estimation network to minimize the pose differences. Extensive qualitative and quantitative experiments conducted on the HumanArt and LAION-Human datasets clearly demonstrate that our model can achieve significant performance improvement over the latest benchmark models.

Code — https://xiangchenyin.github.io/GRPose

1 Introduction

Human image generation aims to synthesize high-quality images under specific conditions based on a series of prompts, such as canny edge, pose and depth (Yang et al. 2023; Wang et al. 2024a). Its diverse applications range from animation (Corona et al. 2024) and game production (Pan et al. 2024) to other fields. Early methods (Men et al. 2020; Ma et al. 2017) primarily adopted variational autoencoders (VAEs) (Kingma and Welling 2013) or Generative Adversarial Networks (GANs) (Goodfellow et al. 2020), leveraging a source image to synthesize target images with specific human attributes. Although these methods achieve control through the reference appearance, the synthesis process is unstable and the training heavily depends on the distribution of the source images. Recently, Stable Diffusion (SD) (Rombach et al. 2022) and its variants (Podell et al. 2023) have been developed to address these limitations, in which high-fidelity human images are synthesized with the help of a prompt. Considering the data availability and computation costs, controllable diffusion models (Zhang, Rao, and Agrawala 2023; Mou et al. 2024) further introduce a learnable control branch into the frozen SD model, enabling spatial control of the generative results based on the provided conditions, *e.g.,* depth maps and segmentation masks.

In recent years, ControlNet (Zhang et al. 2023) has emerged as a new benchmark by fine-tuning the frozen SD model with trainable copy parameters, enabling spatial control through conditional inputs. HumanSD (Ju et al. 2023b) adopts a heat-guided loss to achieve pose control. While these methods synthesize images with satisfactory semantics and style, they are still struggling to produce high-quality output based on the pose priors, often resulting in unrealistic body alignment (as shown in Figure 1). A key challenge in existing approaches is the poor alignment with the given pose priors. We observed that these methods incorporate human pose information into frozen generative models based on Euclidean space, which inadequately models the nonlinear and higher-order relationships between different parts of pose priors, particularly in terms of joint connections and overall coordination.

In this study, we propose to capture the topological relationships among different pose parts by leveraging a graph structure to extract intrinsic associations. We propose a framework named Graph Relation Pose (GRPose) to guide the process of stable diffusion. We follow the ControlNet framework that utilizes both a text prompt and a pose prior to generate human images, and primarily fine-tune the Adapter which is a set of trainable copy parameters. We aim to establish a graph topological structure on pose priors and latent representations, adopting Graph Convolutional Networks (GCNs) (Kipf et al. 2016; Han et al. 2022) to effectively capture the higher-order relationships between different pose parts. We introduce a graph construction mechanism called the Progressive Graph Integrator (PGI). PGI

[*]Wei Chen and Xun Yang are co-corresponding authors. The work was done within intern at Li Auto.

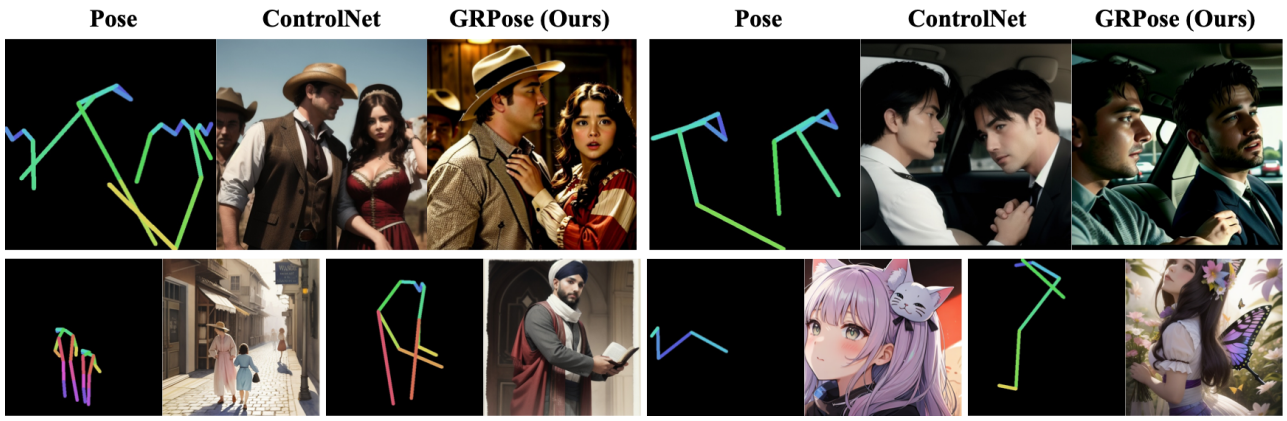

Figure 1: Examples of the pose-guided human image generation task. The first row illustrates the generated results of the ControlNet and our GRPose methods while the second row visualizes the pose alignment across different base models. GRPose generates better results by well aligning with pose prior and scaling outputs to 512×512 pixels.

treats each spatial part of pose priors and latent representations as graph nodes, mapping the spatial relationships of the pose into a graph structure. This mechanism distills pose information into each layer within the Adapter through a hierarchical structure, allowing gradual refinement of the pose while preserving the authenticity of the appearance. Additionally, we propose a novel pose perception loss, which adopts a pretrained pose estimation network to minimize the pose differences between outputs and original images, further encouraging alignment of the synthesized output with pose priors. Our contributions are summarized as follows:

- We propose a novel framework, termed GRPose, that exploits a graph structure to effectively capture the topological and spatial information of pose priors for human image generation, offering a new perspective in the field.
- We design a Progressive Graph Integrator (PGI) to explicitly capture the intrinsic associations between different pose parts, distilling pose information into each layer.
- We introduce a novel pose perception loss that utilizes a pose estimation network to minimize pose differences.
- We conduct extensive experiments on the Human-Art (Ju et al. 2023a) and LAION-Human datasets (Ju et al. 2023c). Our GRPose achieves superior performance compared with advanced methods across multiple metrics, particularly in terms of pose guidance alignment.

2 Related Work

2.1 Pose-Guided Human Image Generation

Traditional pose-guided human generation methods take a specific source image and pose condition as input to generate images that retain the appearance of the source image while presenting a specified pose. Recently, deep learning techniques have achieved significant progress in various domains (Fan et al. 2022; Liang et al. 2024). Some studies (Lv et al. 2021; Ma et al. 2021; Yang et al. 2021) are based on GANs or VAEs to convert the task as a conditional generation. Bhunia *et al.* (Bhunia et al. 2023) leveraged a Person Image Diffusion Model (PIDM) to synthesize images by learning a noise distribution and proposed a cross-attention-based Texture Diffusion Module (TDB) to align the relationship between appearance and pose information. Zhang *et al.* (Zhang et al. 2022) proposed a dual-task pose transformer and introduced an auxiliary task, connecting different branches to obtain correlation between features by building attitude conversion modules. Shen *et al.* (Shen et al. 2023) introduced a Progressive Conditional Diffusion Model (PCDM), which incrementally bridges the gap between character images under target poses and source poses through a three-stage process. HumanSD (Ju et al. 2023b) finetuned the diffusion model with a pose-guided heatmap loss and created a new LAION-Human dataset. Stable-Pose (Wang et al. 2024b) introduced a coarse-to-fine attention masking strategy into ViT to obtain more accurate pose guidance. Despite the significant progress of existing methods, challenges remain in handling complex poses.

2.2 Controllable T2I Diffusion Model

Large Diffusion models (Rombach et al. 2022; Ramesh et al. 2022) generate high-quality images under a set of prompts. Ho *et al.* (Ho and Salimans 2022) proposed classifier-free guidance that combines conditional prediction with unconditional prediction. ControlNet (Zhang, Rao, and Agrawala 2023) introduces a trainable control branch by incorporating a weight copy of the diffusion model, allowing spatial control of generative images using depth maps, segmentation masks and more. T2I-Adapter (Mou et al. 2024) trains a lightweight adapter to guide the frozen SD model. Uni-ControlNet (Zhao et al. 2024) utilizes two additional control modules to achieve control over multiple conditions. Although these methods demonstrate strong control capabilities, they struggle with handling complex pose relationships.

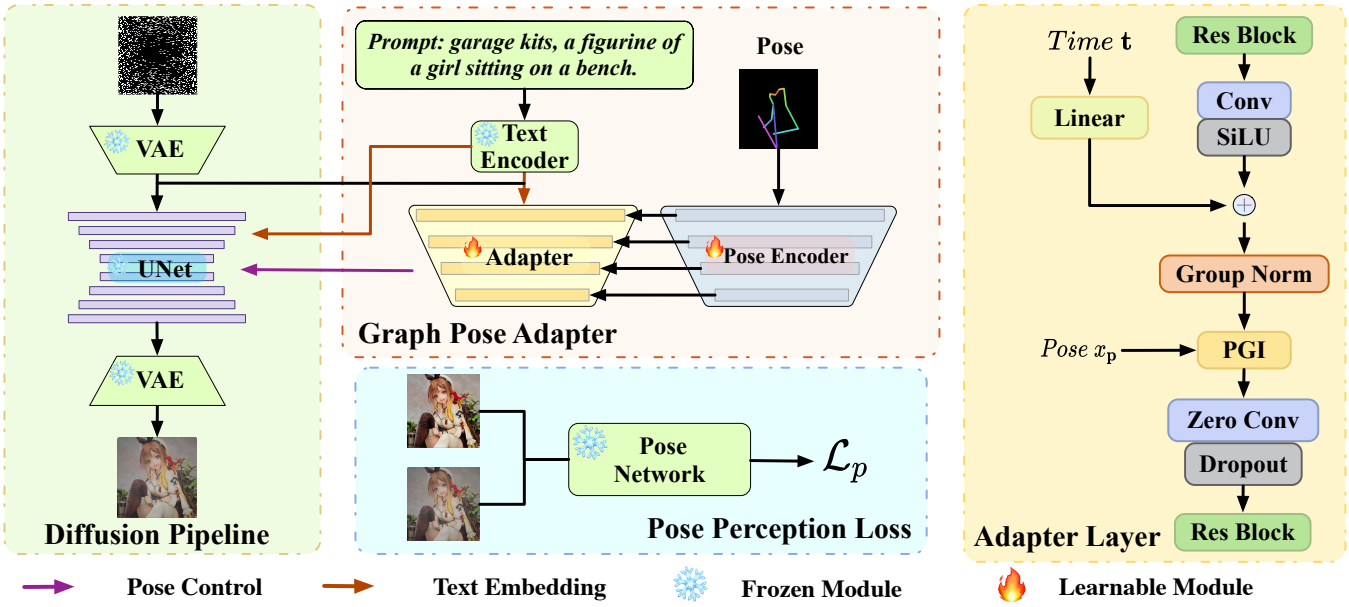

Figure 2: Overview of Graph Relation Pose (GRPose). The Pose Encoder is adopted to capture multi-level scales of pose priors within a hierarchical structure, where a Progressive Graph Integrator is incorporated to capture graph relationships between different pose parts. The Pose Perception Loss adopts a pre-trained pose estimation network to regularize the pose alignment.

3 Methodology

3.1 Overview of Our GRPose

Our aim is to generate high-quality human images conditioned on pose priors. Our proposed GRPose consists of three main components: Diffusion Pipeline, Graph Pose Adapter and Pose Perception Loss, as shown in Figure 2. Given a pose condition $c_p \in \mathbb{R}^{H\times W\times C}$ and a text prompt as inputs, the CLIP text encoder (Radford et al. 2021) converts the text into its embedding c_t. In the Pose-Guided Diffusion Pipeline, we use Stable Diffusion (Rombach et al. 2022) as the base diffusion model, which includes a VAE for image encoding and decoding, and a U-Net for noise estimation. Specifically, an image is fed into the encoder of the VAE to obtain the latent representation z_0. The optimization objective is to maximize the logarithmic likelihood of the data distribution, as defined:

$$\mathcal{L}_d = E_{\mathbf{z}_t,t,\epsilon_t\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[\|\epsilon_t - \epsilon_\theta\left(\mathbf{z}_t,\mathbf{c}_t,\mathbf{c}_p,t\right)\|^2\right], \quad (1)$$

where z_t denotes the latent representation at time step t, $\mathcal{N}$ denotes the standard normal distribution, ϵ_θ denotes the noise prediction network and ϵ_t denotes real noise.

Within the entire framework, the Graph Pose Adapter is a trainable component that encodes the pose condition c_p into a graph structure and integrates it into the Adapter through a hierarchical structure. At the beginning of each encoder layer in the Adapter, the encoded pose and the current latent representation are fed into the Progressive Graph Integrator (PGI) to capture the topological relationships between different pose parts through graph learning. This process fine-tunes the Adapter to transfer the control signals into the SD model, producing the synthesized image $\hat{x}$. Additionally, to further encourage the alignment of synthesized output with pose priors, the pose perception loss is formulated using a pretrained pose estimation network to quantify the pose differences between the outputs and original images.

3.2 Progressive Graph Integrator

We adopt a pose encoder with L layers (L=4, similar to the Adpater) to capture multi-level scales of pose priors within a hierarchical structure, where a PGI is incorporated into each encoding layer and its corresponding layer in the Adapter. In the PGI, each spatial point in the encoded pose prior x_p and the latent representation $x_l \in \mathbb{R}^{H\times W\times C}$ is treated as a node to construct the graph structures, as shown in Figure 3. By doing this, both local features of spatial relationships in the pose prior and latent representation are captured. Then, we proceed to construct the feature graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where the node set is defined as $\mathcal{V} = \{v_1, v_2, \ldots, v_N\}$ where $N = H \times W$ denotes the number of nodes. The graph edges between nodes v_i and v_j are established by applying a K-nearest neighbors (KNN) search algorithm:

$$N_K(v_i) = \{v_j \in V \mid d(v_i, v_j) \leq K\}, \quad (2)$$

where d denotes the L_2 distance between two node vectors. Notably, we incorporate positional encoding p_i into each node ($v_i \leftarrow v_i + p_i$), where p_i denotes the spatial positions of the corresponding patch. A set of edge $\mathcal{E}$ can be obtained as:

$$\mathcal{E} = \{(v_i, v_j) | v_j \in N_K(v_i), i \neq j\}, \quad (3)$$

Then, two separate graph structures, $\mathcal{G}_p$ and $\mathcal{G}_l$, are formed to represent the pose prior x_p and latent representastion x_l respectively. With these two graphs, we employ graph convolution (GC) layers to facilitate the association of features

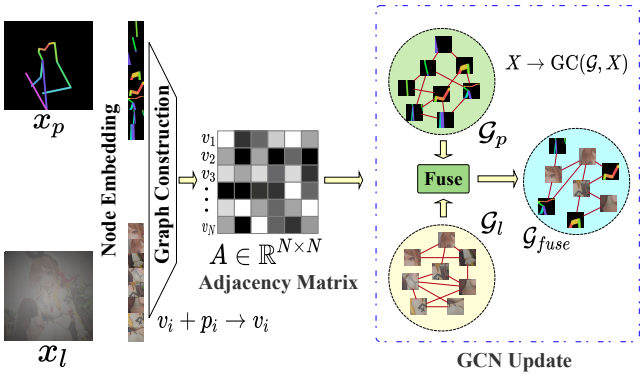

Figure 3: Details of Progressive Graph Integrator (PGI). The pose prior x_p and latent representation x_l are gridded to construct graphs $\mathcal{G}_p$ and $\mathcal{G}_l$ respectively, where GCNs are employed to fuse and update the information.

between nodes within each graph. The GC layer is divided into two steps: aggregation and update. In the aggregation step, features of neighboring nodes are gathered as follows:

$$F_{agg} = \sum_{j \in \mathcal{K}(i)} \left(\mathcal{A}_{ij} \cdot X_j\right), \tag{4}$$

where $\mathcal{K}(i)$ denotes the set of the neighboring nodes of the node i determined by the KNN algorithm, $\mathcal{A} \in \mathbb{R}^{N \times N}$ is an adjacency matrix, and X denotes the features of nodes. In the update step, the current node representation is updated based on the aggregated features, as follows:

$$X_i' = \boldsymbol{\phi}\left(\boldsymbol{\Theta} \cdot X_i \oplus (1 - \boldsymbol{\Theta}) \cdot F_{agg}\right), \tag{5}$$

where $\boldsymbol{\phi}$ denotes the activation function, $\boldsymbol{\Theta}$ is a learnable parameter. Then, the adjacency matrix $\mathcal{A}$ is normalized as:

$$\hat{\mathcal{A}} = D^{-\frac{1}{2}}(\mathcal{A} + I)D^{-\frac{1}{2}}, \tag{6}$$

where D denotes the diagonal degree of $\mathcal{A}$, and I denotes the identity matrix. The GC process can be described as:

$$\hat{X} = \boldsymbol{\phi}\left(\hat{\mathcal{A}}XW\right) + X, \tag{7}$$

where W denotes the weight matrix. The GC layer distills information across different parts for both pose prior and latent representation. Take the pose prior as an example, this process is summarized as:

$$\hat{X_p} = \mathrm{GC}(\mathcal{G}_p, X_p), \tag{8}$$

We further fuse the graph features of pose and latent through a fusion layer, which captures the cross-modal interactions. The fusion layer consists of three convolution blocks. The fused features are then used to construct a graph $\mathcal{G}_{fuse}$, which is fed into an additional GC layer to capture the complex associations to refine the feature representation.

3.3 Pose Perception Loss

We introduce a novel Pose Perception Loss to enhance the pose alignment during the image synthesis process under the guidance of a pose perception network. Our approach employs a pre-trained pose estimation network, specifically designed to accurately identify and comprehend the poses of individuals within images. By encoding both the generated image $\hat{x}$ and the original image x, the discrepancies between them can be captured effectively, defined as follows:

$$\mathcal{L}_p = \frac{1}{hw} \sum_{i=0}^{h} \sum_{j=0}^{w} \left\| \varphi_p\left(\hat{x}_{ij}\right) - \varphi_p\left(x_{ij}\right) \right\|_2^2, \tag{9}$$

where φ_p denotes the encoder of the pose estimation network, which extract the features from the pose backbone, h and w denotes the width and height of both x and $\hat{x}$ respectively and $\|\cdot\|_2$ denotes the Euclidean distance. By directly minimizing the pose differences in the feature space, our method is capable of generating images that are not only visually more realistic but also consistent with pose priors. This loss function can be integrated as a plug-and-play component into existing image generation frameworks, providing additional pose guidance to these models. The optimization objective of the training process is formulated as:

$$\mathcal{L} = \mathcal{L}_d + \alpha * \mathcal{L}_p. \tag{10}$$

where $\mathcal{L}_d$ denotes the diffusion loss, and α denotes the weight of the pose perception loss $\mathcal{L}_p$. The parameters of the Adapter are updated through this loss function.

4 Experiments

4.1 Experimental Settings

Datasets. We evaluated our model on the Human-Art (Ju et al. 2023a) and LAION-Human (Ju et al. 2023c) datasets. The Human-Art dataset comprises 50,000 high-quality images from 5 real-world and 15 virtual scenarios, featuring human bounding boxes, key points and textual descriptions. The LAION-Human dataset consists of approximately 1 million image-caption pairs, filtered by high image quality and human estimation confidence scores, using a diverse range of human activities and more realistic images. In the LAION-Human dataset, we randomly selected 200,000 samples for training and 20,000 samples for testing.

Implementation Details. For a fair comparison, our diffusion pipeline adopts the same Stable Diffusion 1.5[1] as previous methods. We used the Adam optimizer (Kingma and Ba 2014) with an initial learning rate of 1×10^{-5}. We trained models on $8\times$ NVIDIA L40S-48GB, using Pytorch and PytorchLightning[2]. The batch size was set as 6, and the number of epochs was 20. The weight of pose loss α was set to 0.01 and we adopted a gradual constraint strategy, adding pose perception loss in the last 5 epochs. We replaced the text prompts with an empty string with a probability of 0.5. For the PGI, we set the parameter K of the KNN algorithm to 9. During inference, the number of steps T of the DDIM sampler was set to 50. We used the MMpose[3] framework as the pre-trained pose estimation networks. It is worth noting

[1]https://huggingface.co/runwayml/stable-diffusion-v1-5
[2]https://lightning.ai/docs/pytorch/stable/
[3]https://github.com/open-mmlab/mmpose

Dataset	Methods	AP (%) ↑	SAP (%) ↑	PCE ↓	FID ↓	KID ↓	CLIP-Score ↑
Human-Art	SD* (Rombach et al. 2022)	0.24	55.71	2.30	11.53	3.36	33.33
	T2I-Adapter (Mou et al. 2024)	27.22	65.65	1.75	11.92	2.73	33.27
	ControlNet (Zhang et al. 2023)	39.52	69.19	1.54	11.01	2.23	32.65
	Uni-ControlNet (Zhao et al. 2024)	41.94	69.32	1.48	14.63	2.30	32.51
	HumanSD (Ju et al. 2023b)	44.57	69.68	1.37	10.03	2.70	32.24
	GRPose (Ours)	49.50	70.84	1.43	13.76	2.53	32.31
LAION-Human	SD* (Rombach et al. 2022)	0.73	44.47	2.45	4.53	4.80	32.32
	T2I-Adapter* (Mou et al. 2024)	36.65	63.64	1.62	6.77	5.44	32.30
	ControlNet* (Zhang et al. 2023)	44.90	66.74	1.55	7.53	6.53	32.31
	Uni-ControlNet (Zhao et al. 2024)	50.83	66.16	1.41	6.82	4.52	32.39
	HumanSD (Ju et al. 2023b)	50.95	65.84	1.25	5.62	7.48	30.85
	GRPose (Ours)	57.01	67.20	1.29	6.52	4.65	32.12

Table 1: Results on the Human-Art and LAION-Human datasets. The best results and the second best results are marked in green and blue, respectively. Results marked with asterisk (*) are reimplemented based on the released models.

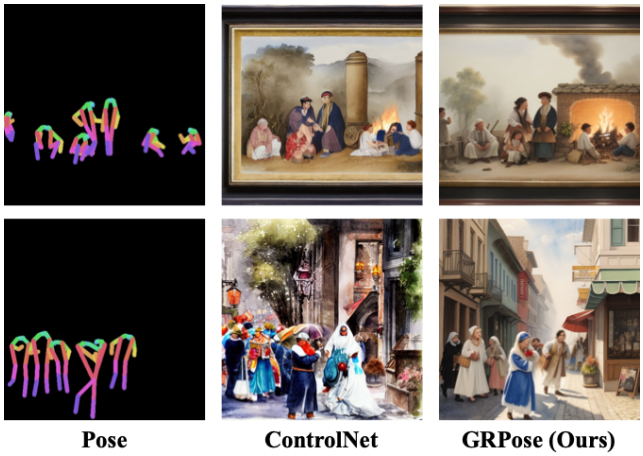

Figure 4: Two cases of Multi-Pose Generation. Our model outperforms ControlNet in generating multiple poses.

that in the evaluation phase we used a different pre-trained pose estimation network during the inference process to assess the accuracy of the pose in the generated images. This ensures the independence and reliability of the evaluation.

Metrics. We adopted Average Precision (AP), Similarity Average Precision (SAP) and Person Count Error (PCE) (Cheong et al. 2022). Higher scores in AP and SAP indicate better pose alignment. Furthermore, we used the Fréchet Inception Distance (FID) (Heusel et al. 2017) and the Kernel Inception Distance (KID) (Bińkowski et al. 2018) to assess the quality of the generated images and used the CLIP-Score (Radford et al. 2021) to assess the alignment between the image and text.

4.2 Comparison with SOTA Methods

Quantitative results comparing our method with other state-of-the-art (SOTA) approaches are shown in Table 1. The experimental results indicate that our GRPose achieved the highest AP and SAP. On the Human-Art dataset, our model attained AP and SAP of 49.50% and 70.84% respectively, bringing an improvement of 9.98% in AP, compared with ControlNet. Similarly, on the LAION-Human dataset, our model showed a consistent improvement of 6.06% in AP compared to HumanSD (Ju et al. 2023b). Notably, our method significantly improved pose alignment accuracy, as evidenced by the AP, SAP, and PCE metrics. KID is multiplied by 100 for Human-Art and 1000 for LAION-Human for readability. We observed that SD1.5 using only a text prompt without pose conditions lacked the ability for pose alignment. Although controllable diffusion models such as ControlNet and T2I-Adapter used pose conditions for spatial control, their capabilities of pose alignment were limited due to incorporating it without considering the intrinsic high-order associations between different pose parts. While retaining pose alignment, our model also maintained good performance in image quality and text alignment with minor decline in scores.

We further compared the visualized qualitative results with other methods (Zhang et al. 2023; Mou et al. 2024; Ju et al. 2023b; Mou et al. 2024; Rombach et al. 2022) across several samples, as shown in Figure 7. We found that SD1.5 lacked effective pose alignment capability since it relies solely on text prompts. Although T2I-Adapter and ControlNet exhibited a good ability to comprehend textual semantics, their performance of pose alignment was poor, and HumanSD also performs poorly in pose alignment. In contrast, our GRPose demonstrated good pose alignment, significantly enhancing the details in human image generation. We also verified the model's ability in handling multiple poses in a single condition, as shown in Figure 4. Compared with ControlNet, our model can achieve accurate alignment with multiple poses, proving the feasibility of the graph learning approach. Our model addressed issues such as unnatural body positions and misaligned pose, sig-

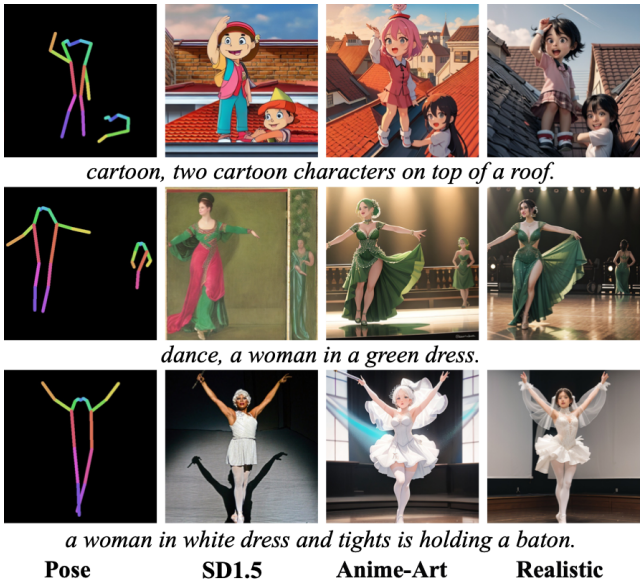

Figure 5: Qualitative Results of our GRPose with different base diffusion models. We compared SD1.5, Anime Art and Realistic models of different styles.

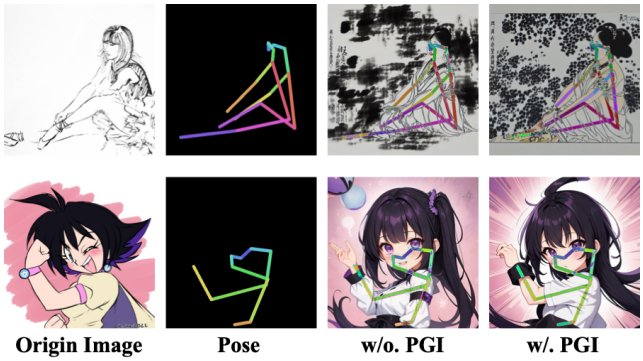

Figure 6: **Effect of PGI**. Our model with PGI demonstrates better pose guidance due to graph associations.

nificantly enhancing the accuracy of pose estimation. We perform transfer verification of the base model across different variants of SD-1.5 from the open-source community Civitai, as shown in Figure 5. Our GRPose, when integrated with different base models, can effectively improve the performance of pose-guided alignment, demonstrating that GRPose is a plug-and-play and effective solution.

4.3 Ablation Studies

All Components. We conducted experiments on the Human-Art dataset to validate the effectiveness of each component, as shown in Table 2. Compared with ControlNet, the usage of the Pose Perception Loss $\mathcal{L}_p$ led to an increase of 4.29% in AP and 1.18% in SAP, demonstrating that the pose perception loss effectively constrained the pose differences in the feature space from an alternative perspective. Incorporating the PGI further improved the performance of our model, with AP and SAP increasing by 8.28% and 1.44% respectively. This improvement indicated that PGI, benefited from its modeling of complex graph topological relationships, accurately captured high-order relationships between pose parts. It is evident that after PGI associates the information of different parts of the pose, our model produced significant improvements in both AP and SAP, addressing previous issues such as blurred limbs and unrealistic poses, as shown in Figure 4.

Components	AP (%)↑	SAP (%)↑	PCE↓
ControlNet	39.52	69.19	1.54
+PGI	47.80	70.63	1.51
+$\mathcal{L}_p$	43.81	70.37	1.48
+PGI+$\mathcal{L}_p$	**49.50**	**70.84**	**1.43**

Table 2: **Results of each component.** $\mathcal{L}_p$ denotes the Pose Perception Loss. The results verify the effectiveness of each component in our framework.

Scenes	Methods	AP (%)↑	SAP (%)↑	PCE↓
Upper Body	ControlNet	15.51	66.45	0.63
	Ours	**31.87**	**70.05**	**0.57**
Full Body	ControlNet	30.78	70.12	1.29
	Ours	**39.19**	**71.03**	**1.17**

Table 3: Results of pose guidance in upper body and full body on the cartoon subset from Human-Art dataset.

To verify the ability of portrait generation, we conducted additional experiments on *upper body* and *full body* samples from the cartoon subset in Human-Art, as shown in Table 3. In the upper body scenario, our model outperformed ControlNet by 16.36% in the AP. In the full body scenario it was 8.41% higher in the AP. The results showed that compared with ControlNet, our model performs much better in upper body generation than in full body generation, further verifying the effectiveness of our PGI associations. The performance in the upper body scenario indicated that our model achieved good results in portrait generation, which demonstrates the high potential in many downstream applications.

Graph Stages. We conducted empirical exploration to assess the performance of our model with varying numbers of graph structures, as shown in Table 4. The results indicated that we can observe a significant performance increase when the number of graph layers increased from 1 to 4. We did not report more results due to computational resource limitations. In addition, we visualize our results of pose alignment with and without PGI in Figure 6. In the first row of samples, the leg pose showed a noticeable error and was partially missing without PGI. In the second row, the arm was in an incorrect pose. With PGI, pose guidance significantly improved image quality, bringing the generated pose of the image closer to the origin. This shows that capturing high-order relationships through the graph structure effectively facilitates information propagation between different

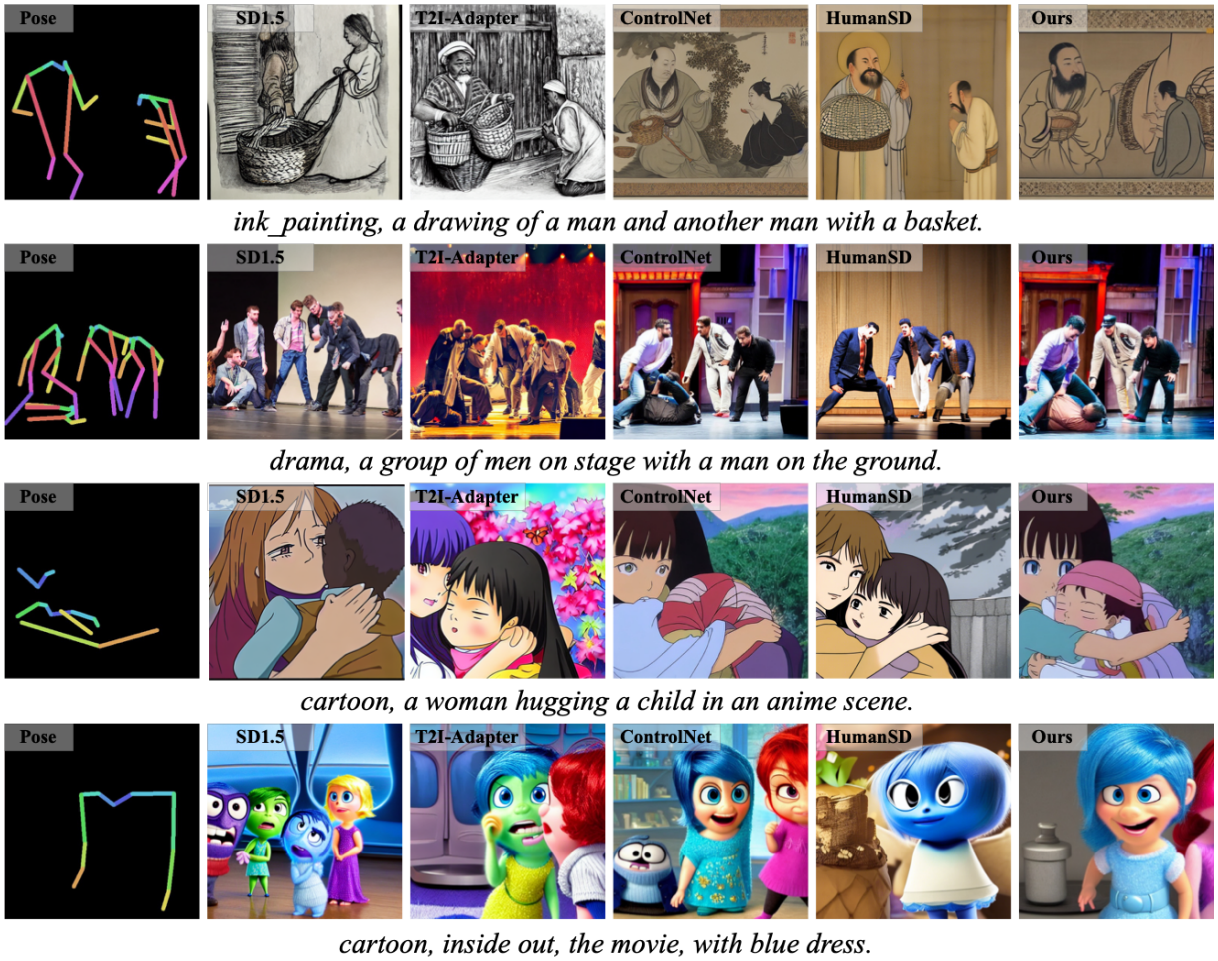

Figure 7: **Qualitative comparison between ours and other methods.** The samples are from the Human-Art dataset, where each row illustrates a sample along with its corresponding pose and prompt.

#. Graph	AP (%)↑	SAP (%)↑	PCE↓
0	41.85	69.96	1.57
1	42.75	70.56	1.52
2	46.59	70.59	1.58
3	**47.80**	**70.63**	**1.51**

Table 4: **Effects of different graph number.** The use of graph convolution layers at different stages significantly improves the quality of pose alignment.

parts of the pose.

Pose Perception Loss. We investigated the impact of different loss weights α on Pose Perception Loss $\mathcal{L}_p$, as shown in Table 5. It can be found that the optimal loss weight is 0.01. However, when the weight was set to 0.1, there was a significant performance decrease in AP, suggesting that it imposed over-high constraints on the diffusion loss. A higher weight for pose perception loss may lead to a failure of pose perception.

α	AP (%)↑	SAP (%)↑	PCE↓
0.1	45.58	70.36	1.52
0.01	**49.50**	**70.84**	**1.43**
0.05	48.44	70.68	1.45

Table 5: **Results of different loss weights α.** When α was set to 0.01, our model yielded the best performance.

5 Conclusion

In this paper, we proposed a framework named Graph Relation Pose (GRPose), for pose-guided human image generation. We designed a Progressive Graph Integrator (PGI), which adopts a hierarchical structure to ensure higher-order associations. In the future, we will try to extend our solution for 3D generation tasks (Di et al. 2024; Luo et al. 2025).

Acknowledgments

We acknowledge the support of the advanced computing resources provided by the Supercomputing Center of the USTC.

References

Bhunia, A. K.; Khan, S.; Cholakkal, H.; et al. 2023. Person image synthesis via denoising diffusion model. In *CVPR*, 5968–5976.

Bińkowski, M.; Sutherland, D. J.; Arbel, M.; and Gretton, A. 2018. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*.

Cheong, S. Y.; Mustafa, A.; and Gilbert, A. 2022. Kpe: Keypoint pose encoding for transformer-based image generation. *arXiv preprint arXiv:2203.04907*.

Corona, E.; Zanfir, A.; Bazavan, E. G.; Kolotouros, N.; Alldieck, T.; and Sminchisescu, C. 2024. VLOGGER: Multimodal diffusion for embodied avatar synthesis. *arXiv preprint arXiv:2403.08764*.

Di, D.; Yang, J.; Luo, C.; Xue, Z.; Chen, W.; Yang, X.; and Gao, Y. 2024. Hyper-3DG: Text-to-3D Gaussian Generation via Hypergraph. *IJCV*.

Fan, L.; Sowmya, A.; Meijering, E.; and Song, Y. 2022. Fast FF-to-FFPE whole slide image translation via Laplacian pyramid and contrastive learning. In *MICCAI*, 409–419. Springer.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2020. Generative adversarial networks. *Communications of the ACM*, 139–144.

Han, K.; Wang, Y.; Guo, J.; Tang, Y.; and Wu, E. 2022. Vision gnn: An image is worth graph of nodes. *NeurIPS*, 35: 8291–8303.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. *NeurIPS*.

Ho, J.; and Salimans, T. 2022. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*.

Ju, X.; Zeng, A.; Wang, J.; Xu, Q.; and Zhang, L. 2023a. Human-art: A versatile human-centric dataset bridging natural and artificial scenes. In *CVPR*, 618–629.

Ju, X.; Zeng, A.; Zhao, C.; Wang, J.; Zhang, L.; and Xu, Q. 2023b. Humansd: A native skeleton-guided diffusion model for human image generation. In *ICCV*, 15988–15998.

Ju, X.; Zeng, A.; Zhao, C.; Wang, J.; Zhang, L.; and Xu, Q. 2023c. Humansd: A native skeleton-guided diffusion model for human image generation. In *ICCV*, 15988–15998.

Kingma, D.; and Ba, J. 2014. Adam: A Method for Stochastic Optimization. *arXiv: Learning*.

Kingma, D. P.; and Welling, M. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Kipf, T. N.; and Welling, M. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Liang, Y.; Hu, Z.; Huang, J.; Di, D.; Su, A.; and Fan, L. 2024. ToCoAD: Two-Stage Contrastive Learning for Industrial Anomaly Detection. *IEEE TIM*.

Luo, C.; Di, D.; Yang, X.; Ma, Y.; Xue, Z.; Wei, C.; and Liu, Y. 2025. TrAME: Trajectory-Anchored Multi-View Editing for Text-Guided 3D Gaussian Splatting Manipulation. *IEEE TMM*.

Lv, Z.; Li, X.; Li, X.; Li, F.; Lin, T.; He, D.; and Zuo, W. 2021. Learning semantic person image generation by region-adaptive normalization. In *CVPR*, 10806–10815.

Ma, L.; Jia, X.; Sun, Q.; Schiele, B.; Tuytelaars, T.; and Van Gool, L. 2017. Pose guided person image generation. *NeurIPS*, 30.

Ma, T.; Peng, B.; Wang, W.; and Dong, J. 2021. Must-gan: Multi-level statistics transfer for self-driven person image generation. In *CVPR*, 13622–13631.

Men, Y.; Mao, Y.; Jiang, Y.; Ma, W.-Y.; and Lian, Z. 2020. Controllable person image synthesis with attribute-decomposed gan. In *CVPR*, 5084–5093.

Mou, C.; Wang, X.; Xie, L.; Wu, Y.; Zhang, J.; Qi, Z.; and Shan, Y. 2024. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *AAAI*, 4296–4304.

Pan, X.; Qin, P.; Li, Y.; Xue, H.; and Chen, W. 2024. Synthesizing coherent story with auto-regressive latent diffusion models. In *WACV*, 2920–2930.

Podell, D.; English, Z.; Lacey, K.; et al. 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*.

Radford, A.; Kim, J. W.; Hallacy, C.; et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*, 8748–8763. PMLR.

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

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *CVPR*, 10684–10695.

Shen, F.; Ye, H.; Zhang, J.; Wang, C.; Han, X.; and Yang, W. 2023. Advancing pose-guided image synthesis with progressive conditional diffusion models. *arXiv preprint arXiv:2310.06313*.

Wang, F.; Guo, D.; Li, K.; and Wang, M. 2024a. Eulermormer: Robust eulerian motion magnification via dynamic filtering within transformer. In *AAAI*, volume 38, 5345–5353.

Wang, J.; Ghahremani, M.; Li, Y.; Ommer, B.; and Wachinger, C. 2024b. Stable-Pose: Leveraging Transformers for Pose-Guided Text-to-Image Generation. *arXiv preprint arXiv:2406.02485*.

Yang, L.; Wang, P.; Liu, C.; Gao, Z.; Ren, P.; Zhang, X.; Wang, S.; Ma, S.; Hua, X.; and Gao, W. 2021. Towards fine-grained human pose transfer with detail replenishing network. *IEEE TIP*, 2422–2435.

Yang, L.; Zhang, Z.; Song, Y.; et al. 2023. Diffusion models: A comprehensive survey of methods and applications. *ACM Computing Surveys*, 1–39.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. In *ICCV*, 3836–3847.

Zhang, P.; Yang, L.; Lai, J.-H.; and Xie, X. 2022. Exploring dual-task correlation for pose guided person image generation. In *CVPR*, 7713–7722.

Zhao, S.; Chen, D.; Chen, Y.-C.; et al. 2024. Uni-controlnet: All-in-one control to text-to-image diffusion models. *NeurIPS*, 36.