修士論文概要書

Master's Thesis Summary

Date of submission: 07/21/2025 (MM/DD/YYYY)

専攻名（専門分野） Department	Computer Science and Communications Engineering	氏 名 Name	Jiu Yi	指 導 教 員 Advisor	Hiroshi Watanabe 印 Seal
研究指導名 Research guidance	Research on Audiovisual Information Processing	学籍番号 Student ID number	CD 5123FG15-9		
研究題目 Title	Pose-Guided Human Image Generation via Multi-Stage ControlNet Fine-Tuning				

1. Introduction

Pose-Guided Human Image generation is a task which generates new human-centric images based on conditional inputs, such as text prompt and pose image. The new image should follow the text description and pose guidance. Early methods mainly relied on GANs [1] and VAEs [2], in which the generated human images suffer from poor quality and weak pose alignment. With the emergence of Stable diffusion model (SD) [3], the quality of generated images has improved significantly. Current research mainly follows two approaches. One is directly fine-tuning SD on pose conditions. The other way is ControlNet [4], which introduces an additional learnable branch to the frozen SD for conditional generation.

However, directly fine-tuning SD requires high computational resources and a large custom dataset and carries a risk of training collapse due to distribution shift. As a result, it's difficult to reproduce. ControlNet is widely used, but the original method still struggles to achieve precise pose control, meaning the generated image may not align well with the pose condition. Therefore, we propose a multi-stage fine-tuning method for ControlNet to improve pose accuracy in human image generation based on pose input. In the first stage, we train the ControlNet from scratch using original latent denoising loss until convergence. In the second stage, we apply a heatmap-guided denoising loss. We conduct extensive evaluations of our method, which show that it significantly improves pose fidelity while keeps the original generation quality compared with the baseline.

2. Related Work

2.1 ControlNet

ControlNet is a neural network architecture that allows diffusion models to integrate additional conditioning inputs for more precise structural control during image generation. It introduces conditional inputs such as human pose, depth map, sketch and so on. It copies a partial backbone of SD and attaches it as an extra branch to the original model. The parameters of the original SD remain frozen, while only the extra branch is trained. This design preserves the capability of pre-trained SD model, while enabling the model to learn new conditions with relatively low computational cost. However, it still has problems generating fine details of human body or handling complex poses.

2.2 Heatmap-guided denoising loss

The heatmap-guided denoising loss is originally from method HumanSD [5], which is used for fine-tuning SD model on pose conditions. The purpose of this loss is to create a heatmap mask in the latent space that assigns higher weights to the pose-relevant regions, guiding SD to focus on the pose areas instead of the background. The loss is calculated as follows:

$$\boldsymbol{L_h = E_{t,z,\varepsilon}\left[\left\|W_a \cdot \left(\varepsilon - \varepsilon_\theta\left(\sqrt{(\bar{\alpha}_t)}z_0 + \sqrt{(1-\bar{\alpha}_t)}\varepsilon, c, t\right)\right)\right\|^2\right], (1)}$$

where $W_a = w \cdot H_E + 1$, H_E is heatmap mask.

Our method is inspired by their loss design and attempts to apply this heatmap-guided denoising to the fine-tuning of ControlNet, which further improves the pose accuracy of generated human images.

3. Proposed Method

We propose a multi-stage fine-tuning method for ControlNet to improve the pose alignment between the generated image and pose condition. As shown in Fig.1, during fine-tuning, we freeze the parameters of VAE encoder and SD model, only update the parameters of ControlNet. We apply two different loss functions at separate stages. In the first stage, we use the original denoising loss. The goal is to obtain a converged ControlNet that enables the generated human image to roughly follow the input pose. The objective of this stage is to make the model responsive to diverse pose conditions. In the second stage, we apply the heatmap-guided denoising loss. The training objective is to refine the model to accurately follow the input pose. This loss encourages model to focus more on the human structure. Through this second stage of continued fine-tuning, ControlNet is further optimized to improve pose alignment accuracy. As a result, the keypoints of the generated human images more closely match the input pose.

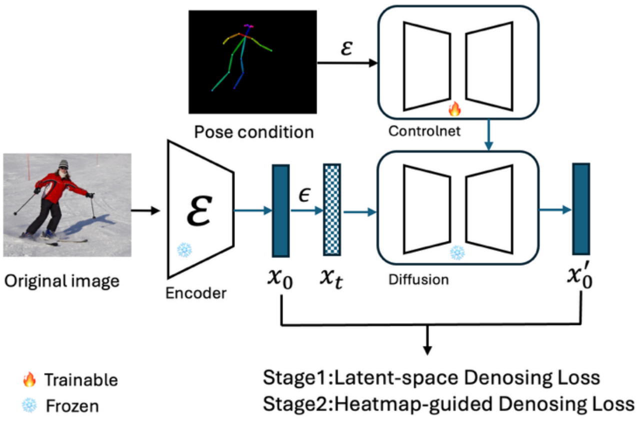

Fig. 1. Overview of multi-stage ControlNet Fine-tuning

4. Experiment

4.1 Training Details

We conduct our experiments on the Captioned COCO-Pose[6] dataset. The dataset includes 61.4k image-pose-caption triplets in the training split and 2.69k pairs for evaluation. We use Stable Diffusion 1.5[7] as the base model for fine-tuning of ControlNet. During training, we adopt different configurations for the two stages. In the common setup, we set batch size to 1, the learning rate to 5×10^{-7},and the gradient accumulation steps to 1. The primary differences lie in the number of training epochs and the loss function used. We apply 4 training epochs in stage 1 and 2 epochs in stage 2. In the first stage, we employ the original denoising loss from SD. In the second stage, we apply the heatmap-guided denoising loss.

4.2 Evaluation

We use checkpoint 255000 as stage 1 model and checkpoint 380000 as stage 2 model. To evaluate the effectiveness of our proposed multi-stage fine-tuning method, we compare both stage models against baseline model, ControlNet-OpenPose. The evaluation is based on three main criteria: OKS for pose accuracy, LIPIS for accessing image quality, and CLIP score for measuring text-image alignment. We conduct both qualitative and quantitative evaluations of our method. Table 1 presents the quantitative comparison with the baseline, showing that our stage 2 model significantly improves pose accuracy while preserving the original model's generative capabilities.

Table 1: Quantitative comparison with the baseline in terms of pose accuracy, image quality, and text-image alignment

Model	CLIP Score ↑	LPIPS ↓	OKS ↑
Stage1	32.3476	0.7762	0.6853
Controlnet-Openpose	31.3786	0.7956	0.7186
Stage2	**31.9787**	**0.7657**	**0.7857**

For the qualitative evaluation, we select serval generated images from our stage 1, 2 and baseline using the same text prompt and pose condition. As shown in Fig. 2, the stage 2 model achieves more accurate pose alignment. Fig.3 highlights the superiority of our stage 2 model in terms of text-image alignment. In the example, the prompt explicitly mentions a bicycle, and only the image generated by our stage 2 model successfully reflects this detail.

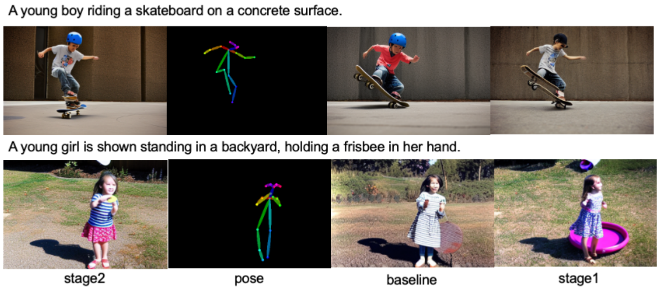
A young boy riding a skateboard on a concrete surface.
A young girl is shown standing in a backyard, holding a frisbee in her hand.
stage2
pose
baseline
stage1

Fig. 2. Qualitative results illustrating pose accuracy

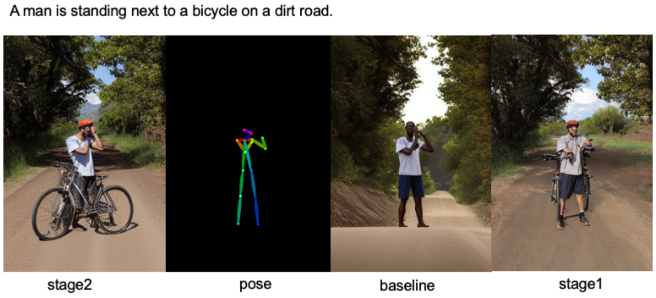

Fig. 3. Qualitative results illustrating text-image alignment

5. Conclusion

We propose a multi-stage fine-tuning method for ControlNet to enhance the pose accuracy of generated human images.In the first stage,we adopt the original latent denoising loss.In the second stage,we contine fine-tuning the model using a heatmap-guided denoising loss, which encourges the model's generation to better align with the input pose condition.Our extensive evalution shows that our method siginificanlty improve pose fidelity compared to the baseline.

Reference

[1] Goodfellow et al., "Generative adversarial networks," Commun. ACM, vol. 63, no. 11, pp. 139–144, 2020.

[2] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," arXiv:1312.6114, 2013.

[3] Rombach et al., "High-resolution image synthesis with latent diffusion models," CVPR, pp. 10684–10695, 2022.

[4] L. Zhang and M. Agrawala, "Adding conditional control to text-to-image diffusion models," arXiv:2302.05543, 2023.

[5] X. Ju et al., "HumanSD: A native skeleton-guided diffusion model for human image generation," ICCV, pp. 15988–15998, 2023.

[6]Liming CV, "Captioned COCOPose dataset," Hugging Face.Available:https://huggingface.co/datasets/limingcv/Captioned_COCOPose, 2023.

[7] Stability AI, "Stable Diffusion v1-5 model card," HuggingFace.Available:https://huggingface.co/stable-diffusion-v1-5/stable-diffusion-v1-5.

Pose-Guided Human Image Generation via Multi-Stage ControlNet Fine-Tuning

A Thesis Submitted to the Department of Computer Science and Communications Engineering, the Graduate School of Fundamental Science and Engineering of Waseda University in Partial Fulfillment of the Requirements for the Degree of Master of Engineering

Submission Date: July 21st, 2025

Jiu Yi

(5123FG15-9)

Advisor: Prof. Hiroshi Watanabe

Research guidance: Research on Audiovisual Information Processing

Acknowledgements

First of all, I would like to thank Professor Hiroshi Watanabe for giving me the opportunity to join the lab and pursue my master's degree. Over the past two years, he has provided many detailed and helpful suggestions during lectures, seminars, conferences, and research discussions. He has always been supportive and responsive whenever I sought guidance, and we've had many meaningful conversations that I will remember for a long time.

I would also like to express my gratitude to the students in the Advanced Multimedia Lab. By participating in seminars, I met many talented and open-minded students from whom I learned a lot.I enjoyed sharing my research interests and gaining new sights from others' presentation.I also received a great deal of help and support,which made my master's journey smoother and enjoyable.I am very thankful for all the encouragement and advice I received from my lab mates.

Finally, I hope I can continue believing in myself and pursue the things I truly want to do in the future.

Contents

1 Introduction

1.1 Research Background

Pose-guided human image generation is a task that aims to synthesize human-centric images based on given conditions such as human pose and text. It has found applications in various real-world scenarios, including avatar generation and human motion transfer, attracting significant attention from both academia and industry.

Early approaches[5, 6] to pose-guided human image generation were based on generative adversarial networks[1] (GANs) and variational autoencoders[2] (VAEs). However, these methods often suffered from poor image generation quality and weak pose alignment. With the emergence of diffusion models[3], the quality of generated human images has improved significantly. Recent research mainly follows two approaches:

(1) Direct fine-tuning of Stable Diffusion model (SD)[3] on pose conditions.

(2) ControlNet[4], which introduces an additional learnable diffusion branch on top of a frozen pre-trained SD model. This extra branch allows for the enforcement of various conditions, such as skeleton poses, during image generation.

However, directly fine-tuning the SD model requires large datasets, significant computational resources, and training time. If the distribution of the fine-tuning dataset deviates too much from that of the original SD training data, it may lead to training collapse or performance degradation. As for ControlNet, the original method still struggles to achieve precise positional control, often producing images that are misaligned with complex poses and lack detail in areas such as hands and fingers.

Therefore, there remains substantial room for improvement in enhancing pose accuracy and achieving better alignment in pose-guided human image generation.

1.2 Research Objectives

Our research aims to improve pose accuracy in human image generation conditioned on pose input. Current ControlNet models conditioned on pose images still struggle to generate fine details of the human body and often fail to reproduce complex poses accurately. To address these limitations, we propose a two-stage fine-tuning method for ControlNet. In the Stage 1, ControlNet is trained from scratch until the model starts to converge. In the Stage 2, we apply a heatmap-guided denoising loss and continue fine-tuning the model. This approach significantly enhances the pose accuracy of the generated human images. We conduct a series of qualitative and quantitative evaluations on image quality, pose accuracy, and text-image consistency, comparing our fine-tuned model with the original ControlNet baseline, demonstrating the efficiency and effectiveness of our method.

1.3 Thesis Outline

The outline of this thesis is as follows:

Chapter 1: This chapter introduces the research background and objectives of the thesis.

Chapter 2: We discuss related work, including previous methods used in pose-guided human image generation and their limitations. We also introduce the foundational research on ControlNet and heatmap-guided loss, which form the basis for our proposed multi-stage fine-tuning method.

Chapter 3: We explain the details of our multi-stage fine-tuning approach, including the inspiration behind it and the different objectives targeted at each training stage.

Chapter 4: We present the experiments conducted to demonstrate the effectiveness and efficiency of our method. This includes a description of the dataset used, training details and setups for each stage, and both quantitative and qualitative evaluations compared to the baseline ControlNet-OpenPose checkpoint.

Chapter 5: We conclude by summarizing our contributions. Our method preserves the original model's generation capability while significantly improving the pose accuracy of the generated human images.

2 Related Work

2.1 Pose-guided human image generation

During the past few years, pose-guided human image generation has become a popular research topic due to the pose's validity in motion description[7]. With source images and pose conditions, a generative model can output a realistic image with the source image's appearance and the desired pose. These methods[5, 6] are mainly based on GANs or VAEs. These methods do not generate a completely new person but transform the person in the source image into a desired pose. Because these methods focus on natural scene manipulation, they usually fail under arbitrary pose conditions and diverse cross-modality feature alignment, and the generated image quality strongly depends on the source image. However, compared with an image, text can be a more flexible and informative condition. With the emergence of Stable Diffusion[3], text-to-image models can generate high-quality human images with text descriptions. However, if we want to precisely control the pose of the generated image, text is not accurate and may require creating a complex prompt to guide the generation process.

Among the very recent works, there are two main directions that introduce precise pose control for human image generation. The first one is HumanSD[8], they directly fine-tune the SD model on the pose condition image, which first inputs the pose image into the VAE encoder the same way as SD, obtains the pose latent embedding, and then concatenates it with the noisy latent embedding generated by diffusion and inputs it into the UNet. The other is called ControlNet[4]. ControlNet is a method that introduces arbitrary conditions to the diffusion model, such as Canny edge, human pose, and depth. ControlNet freezes the parameters of the original SD and only trains the conditional branch, aiming to reuse the image generation capability of the original SD. It is widely adopted due to its modular design, which allows integration of various conditional inputs. Compared with the first method, training Controlnet has the advantage of being trainable with fewer computing resources on a smaller dataset than directly fine-tuning Stable diffusion. However, the current Controlnet model[9], which is conditioned on the OpenPose image, still failed to replicate pose conditions accurately. To improve the model's accuracy regarding pose condition, in this paper, we propose a method to fine-tune Controlnet with multi-stage training.

2.1.1 Adding Conditional Control to Text-to-Image Diffusion Models (ControlNet)

ControlNet[4] is a neural network architecture that allows diffusion models to integrate additional conditioning inputs for more precise structural control during image generation. It works by duplicating the backbone network of Stable Diffusion (SD) and attaching it as an auxiliary branch to the original model. The parameters of the original model remain frozen, while only the newly added components are trained. This design preserves the capabilities of the pre-trained large model, while enabling the model to learn new conditional controls with relatively low computational cost and a smaller training dataset compared to full fine-tuning of Stable Diffusion.

ControlNet has been applied to various tasks, including pose-guided image generation, where it offers more precise control compared to the original text-to-image Stable Diffusion model. However, it still struggles with accurately generating fine details of the human body or handling complex poses. To address this, our method proposes a multi-stage fine-tuning strategy to enhance the pose accuracy of the ControlNet baseline when conditioned on pose inputs.

2.1.2 Heatmap-guided denoising loss

The Heatmap-guided denoising loss was originally invented from the method HumanSD[8], which is used to fine-tune the SD model. The purpose of this custom loss is to make sure the diffusion model can learn better with greater concentration on the condition(human pose) processing. It is realized by assigning a bigger priority factor W_a for feature pixels that are more related to the condition, for example, the human pose area. The heatmap-guided loss is revised from the Vallina LDM[10] loss function:

$$L_{\mathrm{LDM}} = \mathbb{E}_{t,z,\varepsilon}\left[\left\|\varepsilon - \varepsilon_\theta\left(\sqrt{\bar{\alpha}_t} z_0 + \sqrt{1-\bar{\alpha}_t}\varepsilon, c, t\right)\right\|^2\right]. \tag{2.1}$$

The first step is to obtain a difference map that is recognized by the pose estimator. To achieve this, $\varepsilon - \varepsilon_\theta$ is input into the VAE decoder and get :

$$M = \mathrm{VAE}_{decoder}\left(\varepsilon - \varepsilon_\theta\right), \tag{2.2}$$

where M is the resulting difference map.

Next, a state-of-the-art human pose estimation model[11] is applied to the difference map to generate a pose heatmap, defined as:

$$H = F(M), \tag{2.3}$$

where F is the pose estimation model.

To obtain the heatmap mask, a threshold is applied to the heatmap H, resulting in a binary mask H_M. This mask is then passed through the VAE encoder to produce the heatmap embedding.

$$H_E = VAE_{encoder}(H_M), \tag{2.4}$$

where H_E is the heatmap embedding.

Finally,the weighted loss is calculated as follows:

$$L_h = \mathbb{E}_{t,z,\varepsilon}\left[\left\|W_a \cdot \left(\varepsilon - \varepsilon_\theta\left(\sqrt{\bar{\alpha}_t}z_0 + \sqrt{1-\bar{\alpha}_t}\varepsilon, c, t\right)\right)\right\|^2\right], \tag{2.5}$$

where $W_a = w \cdot H_E + 1$,w is set 0.05 by default.

Our method is inspired by their loss design and attempts to apply this heatmap-guided loss to fine-tune ControlNet. We aim to enable ControlNet to focus more on the pose-conditioned regions during the generation process.

3 Proposed Method

3.1 Multi-Stage ControlNet Fine-Tuning

The concept of multi-stage training or fine-tuning is inspired by the training paradigm of large language models (LLMs). Models such as ChatGPT are typically trained in multiple stages, including pre-training, post-training, and supervised fine-tuning. Each stage progressively enhances the model's capabilities, evolving from basic text completion to more complex tasks such as zero-shot question answering. Notably, each stage often employs distinct training strategies and datasets tailored to specific objectives[12, 13].

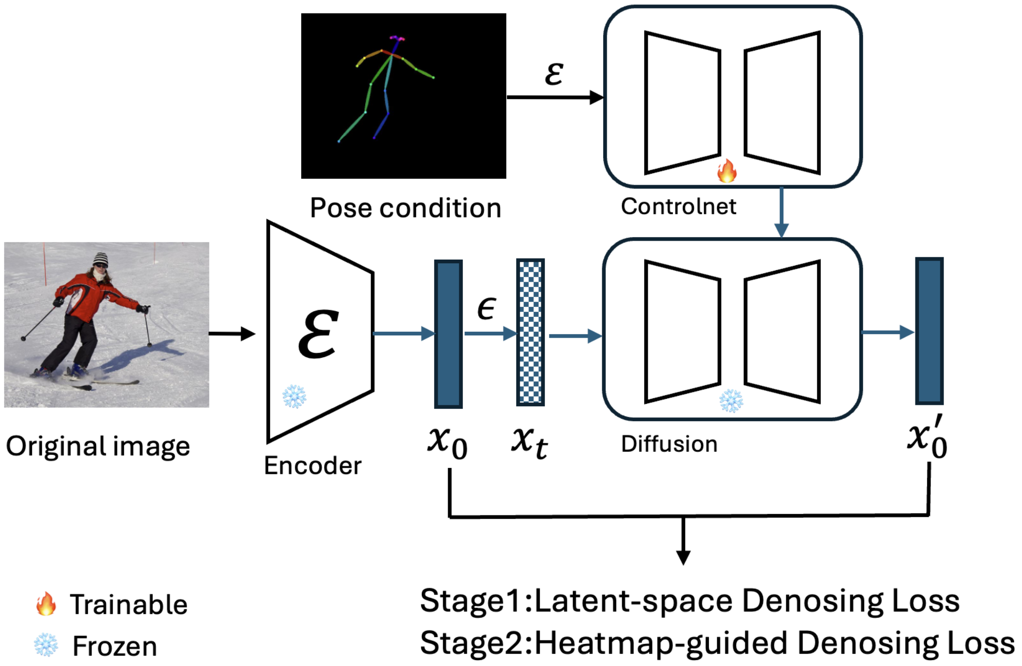

Fig. 3.1. Overview of multi-stage Controlnet Fine-tuning.
Note: Original and pose condition images are from the Captioned COCO-Pose dataset

Following this paradigm, our method, illustrated in Fig. 3.1, adopts a two-stage fine-tuning strategy for ControlNet[4]. We freeze the parameters of the VAE encoder and the Stable Diffusion model[3], and update only the parameters of ControlNet[4] during fine-tuning. In the Stage 1, the goal is to obtain a converged ControlNet that enables the generated human image to roughly follow the input pose. Precise pose fidelity is not required at this stage; instead, the objective is to make the model responsive to diverse pose conditions. We use the original denoising loss[10] from Stable Diffusion[3] during this phase.

In the Stage 2, the training objective shifts to refining the model to accurately follow the input pose. Specifically, the goal is for the keypoints in the generated human image to align closely with those in the input pose. To achieve this, we introduce a heatmap-guided denoising loss[8], which applies a spatial mask in the latent space to assign higher weights to pose-relevant regions. This encourages the model to focus more on human structure rather than the background. Through this fine-tuning stage, the ControlNet model is further optimized to improve alignment accuracy between the generated image and the pose condition.

4 Experiment

4.1 Dataset

We conduct training and evaluation using the Captioned COCO-Pose[15] dataset, which consists of three data modalities: images, control poses, and corresponding textual captions. The dataset includes 61.4k image-pose-caption triplets in the training split and 2.69k pairs for evaluation. All images are resized to 512×512 resolution to meet the input requirements of Stable Diffusion 1.5[14]. During inference, the generated images are resized back to their original resolution for evaluation purposes.

4.2 Training Details

We use Stable Diffusion 1.5[14] (SD 1.5) as the base model in our fine-tuning of ControlNet. A consistent training setup is maintained across the two fine-tuning stages, with the primary differences lying in the number of epochs and the loss functions used.

In the Stage 1, we train the ControlNet from scratch for 4 epochs with a batch size of 1, a gradient accumulation step of 1, and a learning rate of 5×10^{-7}. The loss function employed is the original denoising loss from Stable Diffusion[3]. In the Stage 2, we retain the same batch size, learning rate, and gradient accumulation setting, but reduce the number of training epochs to 2. In this phase, we replace the standard denoising loss with a heatmap-guided denoising loss[8] to better emphasize human pose conditions.The training loss and learning rate schedules for both stages are illustrated in Fig. 4.1.

As an example, we take a single image from the training batch at timestep 859 of the forward diffusion process to illustrate how the heatmap-guided loss is calculated during training, as shown in Fig 4.2. After obtaining H_M, we feed it into the VAE encoder to produce H_E. We inspect the minimum and maximum values of H_E and find that they range from -3.14 to 3.34. Given that the weight w is set to 0.05, the computation of W_a ensures that the mask has a tangible effect on the final loss calculation.

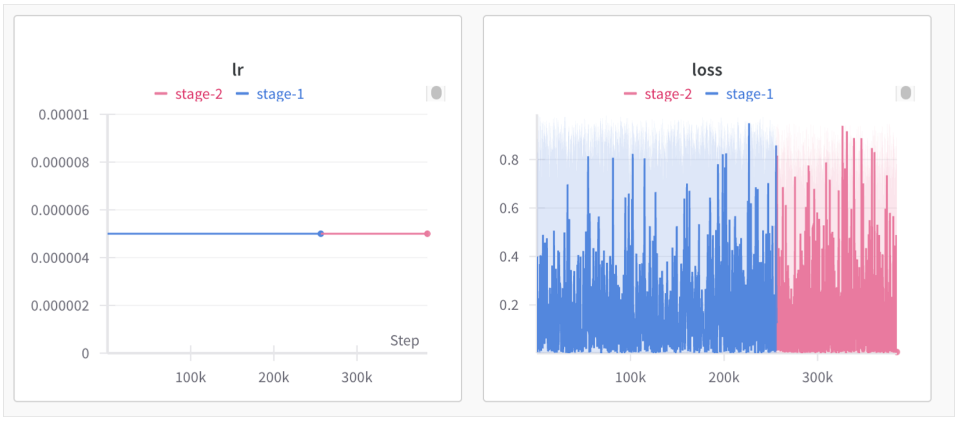

Fig. 4.1. Training loss and learning rate curves for the two-stage fine-tuning process.

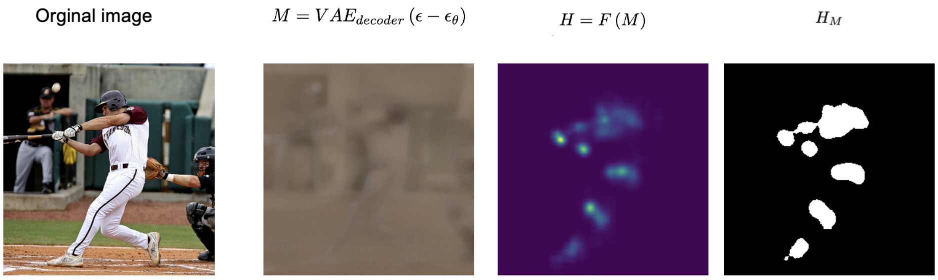

Fig. 4.2. An illustration explaining the calculation of heatmap-guided loss.
Note: Original image is from the Captioned COCO-Pose dataset

4.3 Evaluation Strategy

We use `checkpoint_255000` as our Stage 1 model and `checkpoint_380000` as our Stage 2 model. To evaluate the effectiveness of our proposed multi-stage fine-tuning method, we compare both Stage 1 and Stage 2 models against the baseline model, ControlNet-OpenPose[9]. The evaluation is based on three main criteria: pose accuracy, image quality, and text-image alignment. While the primary goal of our method is to improve the pose accuracy of the generated images, it is also crucial to maintain the overall generative capabilities of the original ControlNet. Therefore, we include image quality and text-image alignment as additional evaluation metrics to ensure that improvements in pose fidelity do not compromise general generation performance.

Pose Accuracy: We adopt Object Keypoint Similarity[16] (OKS) as the primary metric for evaluating pose accuracy. To extract keypoint positions from the generated images, we utilize the YOLOv8x[17] pose estimation model. The OKS score is then computed by comparing the predicted keypoints with the ground-truth annotations. Our objective is to achieve higher OKS scores relative to the baseline models, indicating more accurate pose alignment in the generated outputs.

Image quality: To assess image quality, we use LPIPS [18](Learned Perceptual Image Patch Similarity), which measures perceptual similarity between generated and original images. Lower scores indicate better quality. Our goal is to maintain LPIPS performance comparable to the baseline to ensure pose improvements do not degrade image quality.

Text Alignment: We use the CLIP score[19] to evaluate text-image alignment. This metric measures the similarity between generated images and their corresponding text, ranging from 1 to 100, with higher scores indicating better alignment. Our goal is to maintain CLIP scores comparable to the baseline models.

4.4 Quantitative Results

The quantitative results shown in Table 4.1 demonstrate that our Stage 2 model achieves superior performance in both image quality and text-image alignment. Most notably, compared with the baseline[9], our Stage 2 model significantly improves pose accuracy, highlighting the effectiveness and success of our proposed method.

Table 4.1: Quantitative comparison with the baseline in terms of pose accuracy, image quality, and text-image alignment

Model	**CLIP Score ↑**	**LPIPS ↓**	**OKS ↑**
Stage 1	32.3476	0.7762	0.6853
ControlNet-OpenPose	31.3786	0.7956	0.7186
Stage 2 (Ours)	**31.9787**	**0.7657**	**0.7857**

Additionally, regarding training time and resource efficiency, we used an RTX 3060 GPU to train Stage 1 (31h) and Stage 2 (28h), totaling 59 hours to obtain the final model. In contrast, the original ControlNet-OpenPose[9] checkpoint was trained on an A100 GPU for 300 hours. Our method demonstrates significantly better efficiency. Moreover, our training dataset (64.1k image-pose-caption pairs) is also smaller than the dataset used to train the original ControlNet.

4.5 Qualitative Results

As shown in Fig. 4.3 and Fig. 4.4 ,the Stage 2 model achieves more accurate pose alignment, with the generated images' keypoints aligning more closely to the input pose compared to the baseline. In Fig. 4.3, the first row shows that the position of the boy's right foot is more precisely aligned with the pose condition in the results generated by the Stage 2 model. In the second row, the positions of the girl's hands are also better aligned in the Stage 2 results compared to earlier stages.

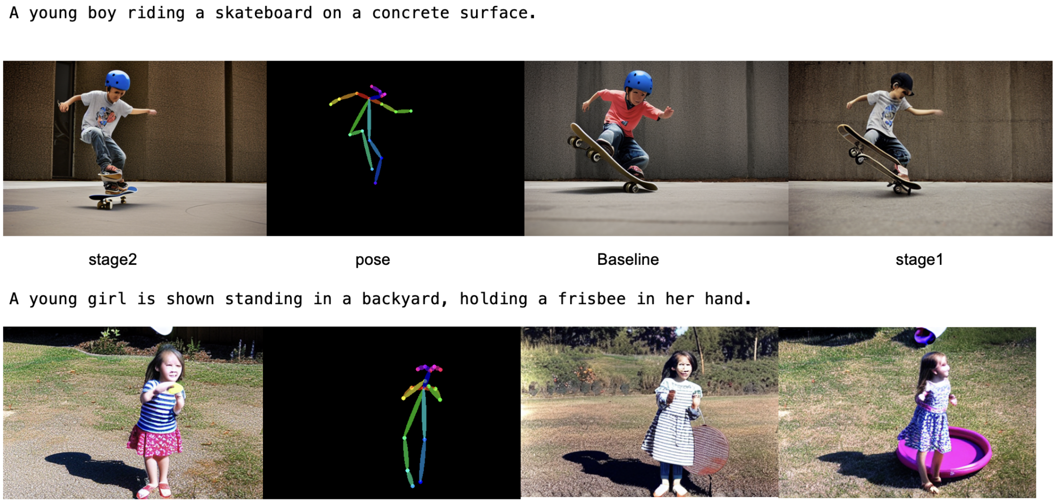
A young boy riding a skateboard on a concrete surface.
stage2
pose
Baseline
stage1
A young girl is shown standing in a backyard, holding a frisbee in her hand.

Fig. 4.3. Qualitative results illustrating pose accuracy.
Note: Pose condition image is from the Captioned COCO-Pose dataset

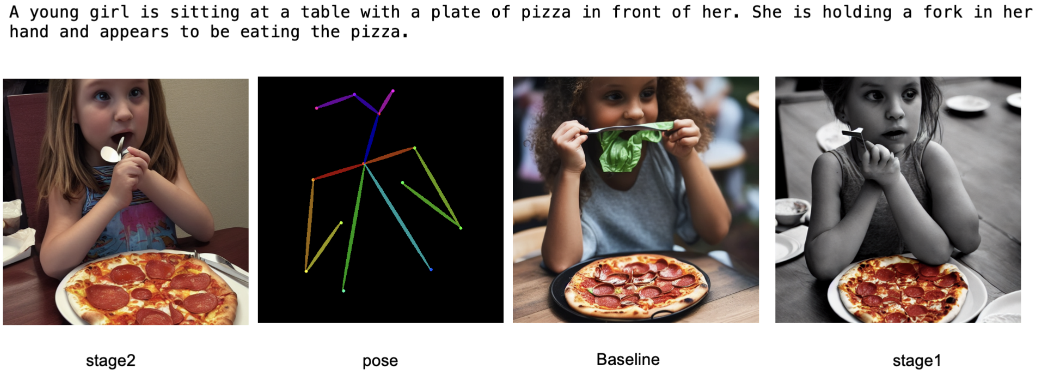
A young girl is sitting at a table with a plate of pizza in front of her. She is holding a fork in her hand and appears to be eating the pizza.
stage2
pose
Baseline
stage1

Fig. 4.4. Qualitative results illustrating pose accuracy.
Note: Pose condition image is from the Captioned COCO-Pose dataset

Fig. 4.5 highlights the superiority of our Stage 2 model in terms of text-image alignment. In this example, the prompt explicitly mentions a bicycle, and only the image generated by our Stage 2 model successfully reflects this detail. Fig. 4.6 further shows that our Stage 2 model generates a more coherent composition in which the woman is holding an umbrella and sitting in a pose that closely follows the guided condition.

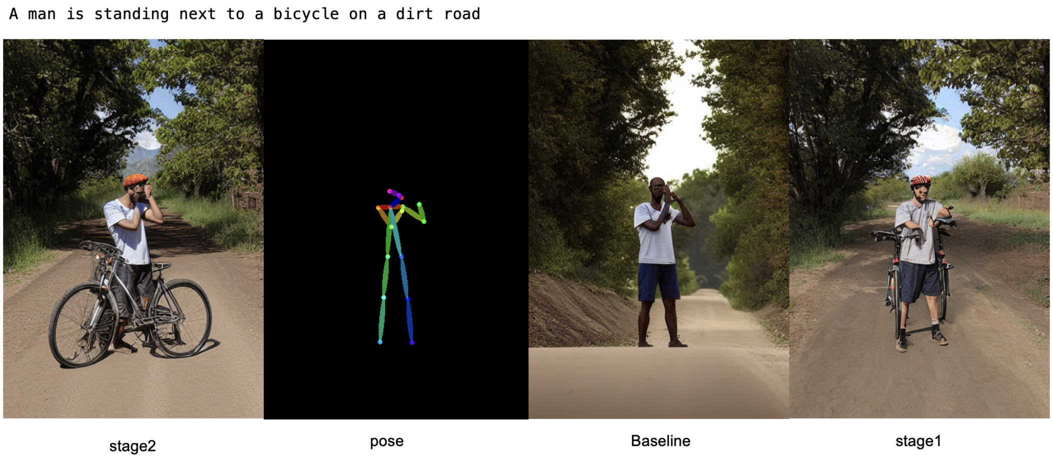

Fig. 4.5. Qualitative results illustrating text alignment.
Note: Pose condition image is from the Captioned COCO-Pose dataset

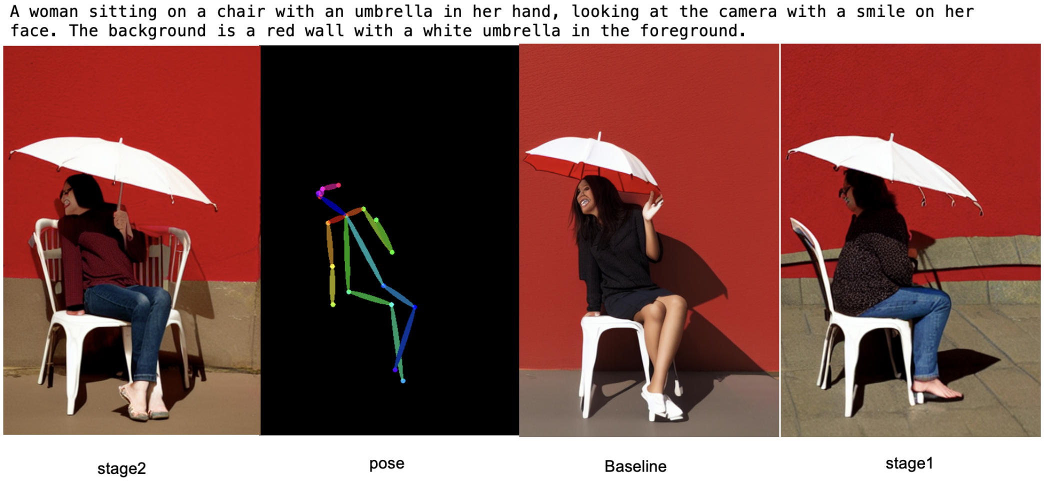
A woman sitting on a chair with an umbrella in her hand, looking at the camera with a smile on her face. The background is a red wall with a white umbrella in the foreground.
stage2
pose
Baseline
stage1

Fig. 4.6. Qualitative results illustrating both pose accuracy and text alignment.
Note: Pose condition image is from the Captioned COCO-Pose dataset

5 Conclusion

In this paper, we propose a multi-stage fine-tuning strategy for ControlNet to enhance the pose accuracy of generated human images. Our approach introduces two distinct denoising losses applied across separate training stages. In the Stage 1, we adopt the original latent denoising loss to train the ControlNet model until convergence. In the Stage 2, we continue fine-tuning the model using a heatmap-guided denoising loss, which incorporates a heatmap mask in the latent space to emphasize pose-relevant regions. This encourages the model to better preserve and align with the input pose. We conduct extensive qualitative and quantitative evaluations in terms of pose accuracy, text-image alignment, and overall image quality. Experimental results show that our method significantly improves pose fidelity compared to the baseline.

Furthermore, our method is both efficient and practical, requiring only low-computation GPUs and small datasets for fine-tuning, which leads to reduced training time. This demonstrates the method's suitability for resource-constrained environments.

List of Publication

1. **Jiu Yi**, Haoyuan Liu, Hiroshi Watanabe: “Distilled RSN: Lightweight Pose Estimation Using Knowledge Distillation,” IEEE Global Conference on Consumer Electronics (GCCE2024), OS-AIP(3), pp.882-884, Nov. 2024.

Bibliography

[1] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial networks,” *Commun. ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[2] D. P. Kingma and M. Welling, “Auto-encoding variational Bayes,” *arXiv preprint arXiv:1312.6114*, 2013.

[3] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, “High-resolution image synthesis with latent diffusion models,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, 2022.

[4] L. Zhang and M. Agrawala, “Adding conditional control to text-to-image diffusion models,” *arXiv preprint arXiv:2302.05543*, 2023.

[5] J. Zhang, K. Li, Y.-K. Lai, and J. Yang, “PISE: Person image synthesis and editing with decoupled GAN,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7982–7990, 2021.

[6] Y. Ren, X. Yu, J. Chen, T. H. Li, and G. Li, “Deep image spatial transformation for person image generation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7690–7699, 2020.

[7] A. Zeng, X. Ju, L. Yang, R. Gao, X. Zhu, B. Dai, and Q. Xu, “DeciWatch: A simple baseline for 10× efficient 2D and 3D pose estimation,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022.

[8] X. Ju, A. Zeng, C. Zhao, J. Wang, L. Zhang, and Q. Xu, “HumanSD: A native skeleton-guided diffusion model for human image generation,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15988–15998, 2023.

[9] L. Zhang, “ControlNet with OpenPose,” Hugging Face. Available: `https://huggingface.co/lllyasviel/sd-controlnet-openpose`.

[10] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 2256–2265, 2015.

[11] B. Cheng, B. Xiao, J. Wang, H. Shi, T. S. Huang, and L. Zhang, "HigherHRNet: Scale-aware representation learning for bottom-up human pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5386–5395, 2020.

[12] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al., "Training language models to follow instructions with human feedback," in *Advances in Neural Information Processing Systems*, pp. 27730–27744, 2022.

[13] OpenAI, "GPT-4 technical report," arXiv preprint arXiv:2303.08774, 2023.

[14] Stability AI, "Stable Diffusion v1-5 model card," Hugging Face. Available: `https://huggingface.co/stable-diffusion-v1-5/stable-diffusion-v1-5`.

[15] Liming CV, "Captioned COCOPose dataset," Hugging Face. Available:`https://huggingface.co/datasets/limingcv/Captioned_COCOPose`, 2023.

[16] D. Maji, S. Nagori, M. Mathew, and D. Poddar, "YOLO-Pose: Enhancing YOLO for multi person pose estimation using object keypoint similarity loss," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, pp. 2637–2646, 2022.

[17] Ultralytics, "YOLOv8 model card," Available: `https://huggingface.co/Ultralytics/YOLOv8`

[18] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, pp. 586–595, 2018.

[19] J. Hessel, A. Holtzman, M. Forbes, R. Le Bras, and Y. Choi, "CLIPScore: A reference-free evaluation metric for image captioning," arXiv preprint arXiv:2104.08718, 2021.