—— Supplementary Materials ——
HumanSD: A Native Skeleton-Guided Diffusion Model for Human Image Generation

Xuan Ju[1,2*‡], Ailing Zeng[1*], Chenchen Zhao[2*], Jianan Wang[1], Lei Zhang[1], Qiang Xu[2†]
[1]International Digital Economy Academy, [2]The Chinese University of Hong Kong
{xju22, cczhao, qxu}@cse.cuhk.edu.hk, {zengailing, wangjianan, leizhang}@idea.edu.cn

Overview

This supplementary material presents more details and additional results not included in the main paper due to page limitation[1]. The list of items included are:

- Experimental details in Sec. A.
- More Quantitative results in Sec. B.
- More Qualitative results in Sec. C.
- Future Work in Sec. D.

A. Experimental Details

We first provide details of evaluation metrics in Sec. A.1, training details of *HumanSD* in Sec. A.3, and detailed explanations of *heatmap-guided denoising loss* in Sec. A.4.

A.1. Evaluation Metrics

Details of the evaluation metrics are as follows:

- **Image Quality:**

  The calculation of Fréchet Inception Distance (FID [11]) is given by:

$$FID = |\mu - \mu_w| + tr\left(\Sigma + \Sigma_w - 2\left(\Sigma\Sigma_w\right)^{\frac{1}{2}}\right) \quad (1)$$

  where $\mathcal{N}(\mu, \Sigma)$ is the multivariate normal distribution estimated from Inception v3 features calculated on *Human-Art* and $\mathcal{N}(\mu_w, \Sigma_w)$ is the multivariate normal distribution estimated from Inception v3 features calculated on generated (fake) images. We use Inception v3 with a feature layer of 64 by default.

  The calculation of Kernel Inception Distance (KID [3]) is given by:

$$KID = MMD(f_{real}, f_{fake})^2 \quad (2)$$

  where MMD is the maximum mean discrepancy and I_{real}, I_{fake} are extracted features from *Human-Art* and generated images.

  The KID and FID are calculated on each scenario in *Human-Art*'s validation set. Then, we average the results of all scenarios to get the final results. Although the validation set of *Human-Art* is relatively small (3,750 images) for FID/KID calculation, given constraints on pose and scenario, we believe they can reflect the quality of generated images to some extent.

- **Pose Accuracy:**

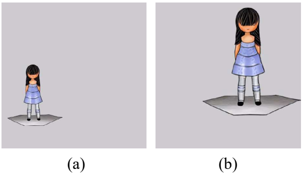


Figure 1: Example of calculating the AP and CAP between (a) and (b) images and obtaining an AP of 0 but a CAP of 1.

  A thorough explanation of the pose accuracy metrics is provided in the main paper. We further explain the difference between distance-based average precision (AP) and pose cosine similarity-based AP (CAP). As shown in Figure 1, humans in (a) and (b) have the same pose but non-overlapping positions. Take the pose of (a) as the target pose. The AP of (b) is 0, but the CAP is 1. Thus, CAP can eliminate the influence of position and focus on the similarity between the two poses. A combination of CAP and AP can better demonstrate the pose controllability over the generated person's pose and position.

- **Text-image Consistency:**

  We employ CLIP-ViT-base-path16 to extract text and image features, using a ViT-B/16 Transformer as the image encoder and a masked self-attention Transformer as the text encoder.

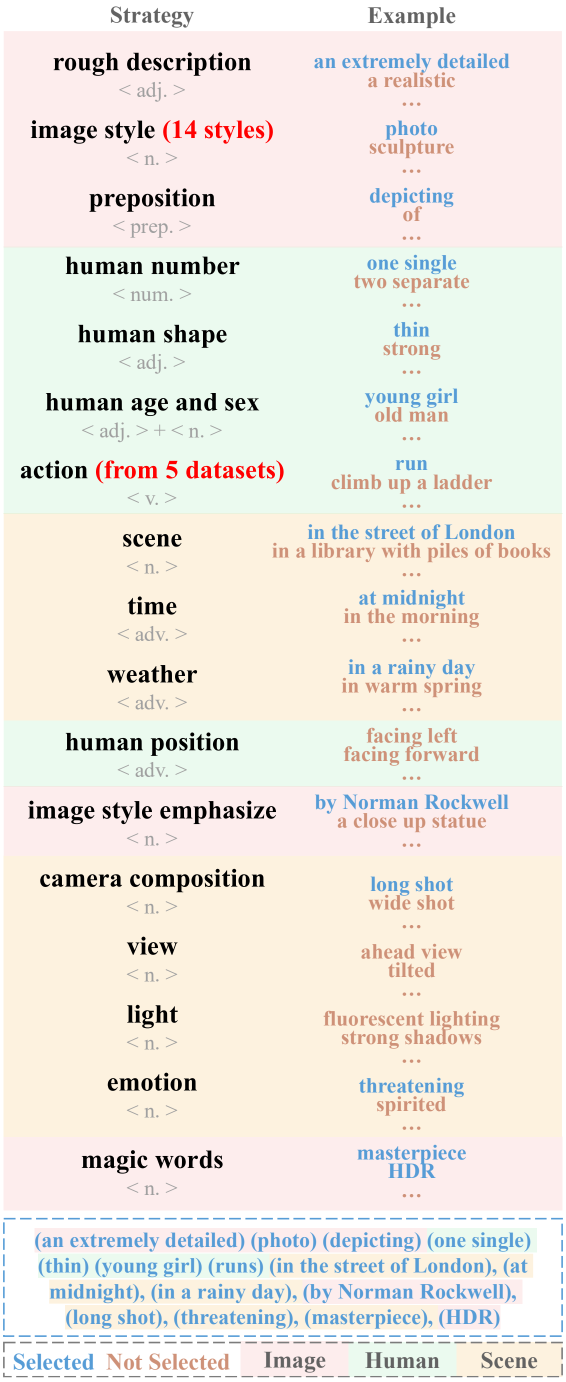
[*] Equal contribution. [‡] Work done during an internship at IDEA.
[†] Corresponding author.
[1] Code and datasets are publicly available for further research: https://idea-research.github.io/HumanSD/

A.2. Prompt Engineering of the GHI Dataset

This work emphasizes the essence of the multi-scenario human-centric image generation task with precise pose control. The quality and scale of the training datasets are crucial. To better control the quality and content of image generation, we design a set of systematic rules for generating prompts. Prompt engineering [21, 28] is one method for increasing the quality of text-to-image models.

In generating the GHI dataset with high-quality, diverse, and human-centric text and the corresponding images, we take advantage of prompt engineering to design large quantities of unrepeatable prompts with a high guarantee of image quality. Specifically, the prompt is composed of 18 parts to describe three main components, like image, human, and scene. Figure 2 includes a comprehensive description of the whole image (e.g., the image style), human features (e.g., the human number, shape, and action), and the background scene (e.g., time, weather, and camera settings). Different parts have distinct selection probabilities and numbers, culminating in a variety of rich and diverse prompts. To ensure the diversity of image styles, we adopt 14 different styles referring Human-Art [13] to cover as many image styles as possible, including photo, garage kits, relief, statue, kids drawing, mural, oil painting, sketch, stained glass, ukiyoe, cartoon, digital art, ink painting, and watercolor. To ensure the diversity of human action, we collect 6826 different human actions referring recent popular human action datasets BABEL [30], NTU RGB+D 120 [20], HuMMan [4], HAA500 [8], and HAKE-HICO [16].

Based on our designed prompts, the used stable diffusion model can generate images of great diversity and quality. However, they may still fail to faithfully respect human structures and generate missing, redundant, replaced body parts or wrong human numbers, which are key to the human-centric image generation task. KPE [7] clarifies the validity of using pose estimators to judge the correctness of both generated body structure and human number. When a pose estimator is fed images with an unreasonable body structure, it typically assigns the incorrect body component to an additional human, resulting in an inconsistent human count. Accordingly, we use the pre-trained pose estimator HigherHRNet to determine whether the estimated human number equals the given number (1-3 humans with a proportion of 7:2:1). We finally reserve 4 images with a correct human number for each prompt in GHI.



Strategy	Example
rough description < adj. >	an extremely detailed a realistic ...
image style (14 styles) < n. >	photo sculpture ...
preposition < prep. >	depicting of ...
human number < num. >	one single two separate ...
human shape < adj. >	thin strong ...
human age and sex < adj. > + < n. >	young girl old man ...
action (from 5 datasets) < v. >	run climb up a ladder ...
scene < n. >	in the street of London in a library with piles of books ...
time < adv. >	at midnight in the morning ...
weather < adv. >	in a rainy day in warm spring ...
human position < adv. >	facing left facing forward ...
image style emphasize < n. >	by Norman Rockwell a close up statue ...
camera composition < n. >	long shot wide shot ...
view < n. >	ahead view tilted ...
light < n. >	fluorescent lighting strong shadows ...
emotion < n. >	threatening spirited ...
magic words < n. >	masterpiece HDR ...

(an extremely detailed) (photo) (depicting) (one single) (thin) (young girl) (runs) (in the street of London), (at midnight), (in a rainy day), (by Norman Rockwell), (long shot), (threatening), (masterpiece), (HDR)

Figure 2: Components of prompts in GHI. Note that the order of prompts has an impact on the generation results.

A.3. The Training Details of HumanSD

We train *HumanSD* with 4 NVIDIA A100 Tensor Core GPUs. By default, we train each model with a batch size of 4 for about 3 days, around 95, 000 iterations and 300 GPU hours. Different from ControlNet, which receives a sudden convergence around the long-lasting 6100_{th} iteration, *HumanSD* shows a fast but smooth convergence from the 0_{th} iteration to the 600_{th} iteration. Although *HumanSD* can generate humans with corresponding pose conditions after only around 600 iterations, training with more iterations leads to better performance. However, training with more than $95,000$ iterations does not improve metrics in our experiments. The results are also shown in Table 4.

We initialize the weights of the first layer in the UNet by copying a portion of the weights (4 channels) from the pre-trained model, while randomly initializing the remaining parts, which is similar to Stable Diffusion's2 fine-tuning for depth2image generation. For weights beyond the first layer, we directly copy them from the pre-trained model. The benefit of initializing only a portion of the weights in the first layer is that it enhances the model's ability to learn from the conditioning information and speeds up convergence during training. While this practice may cause a slight loss of pre-trained information, empirical evidence shows minimal impact on the overall results.

A.4. Detailed Implementation of the Heatmap-Guided Denoising Loss

As explained in the main paper, the vanilla LDM [39] has a loss function:

$$L_{\text{LDM}} = \mathbb{E}_{t,z,\epsilon}\left[\left\|\epsilon - \epsilon_\theta\left(\sqrt{\bar{\alpha}_t}z_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, c, t\right)\right\|^2\right] \quad (3)$$

To obtain a difference map to be recognized by the pose estimator, we feed $\epsilon - \epsilon_\theta$ into the VAE decoder of Stable Diffusion (SD) and get:

$$M = VAE_{decoder}\left(\epsilon - \epsilon_\theta\right) \quad (4)$$

where M is the difference map of noise difference.

We use a bottom-up pose estimator HigherHRNet [6] pre-trained on MSCOCO [18] and *Human-Art* for heatmap estimation of $Difference\ Map$. The bottom-up pose estimator shows better performance on blurred difference maps. Moreover, the estimation can become more inclusive by combining MSCOCO and *Human-Art* in training. We determine the heatmap by:

$$H = F\left(M\right) \quad (5)$$

where $H \in \mathbb{R}^{\mathbf{h}\times\mathbf{w}\times\mathbf{k}}$ is the heatmap matrix with height $\mathbf{h}$ and width $\mathbf{w}$. $\mathbf{k}$ is the human joint number. F is the heatmap estimator. For ease of calculation, we then sum H across the joint dimension to generate a single heatmap.

The larger the difference between the output noise of the UNet ϵ_θ and ground-truth noise ϵ is, the more noticeable the human figure will be in M, and the larger value H will have at the corresponding joint positions. Therefore, to get the heatmap mask, we set 0.1 as the empirical threshold on H to form heatmap mask H_M. Then, we pass H_M back to the VAE encoder to get the heatmap embedding.

$$H_E = VAE_{encoder}\left(H_M\right) \quad (6)$$

where H_E is the heatmap embedding.

Finally, the weighted loss is calculated as follows:

$$L_{\text{h}} = \mathbb{E}_{t,z,\epsilon}\left[\left\|W_a \cdot \left(\epsilon - \epsilon_\theta\left(\sqrt{\bar{\alpha}_t}z_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, c, t\right)\right)\right\|^2\right] \quad (7)$$

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

B. Quantitative results

This section reports more quantitative results to comprehensively assess pose controllability in different image scenarios and when inputting pose conditions with different human numbers. We also provide more ablation study in thie section.

Comparisons on different image scenarios. As shown in Figure 3, all methods exhibit comparable patterns in Pose AP throughout different scenarios, with garage kits having the highest AP and shadow play having the lowest AP. This is partially due to the uneven distribution of the training and fine-tuning datasets, where garage kits have a large amount of data with multi-view and shadow play has only a small number of images. Natural scenarios such as cosplay also have a comparatively higher AP, which again reflects the differential distribution of the dataset. However, due to the complexity of pose conditions, natural scenes such as acrobatics and dance do not show a high AP.

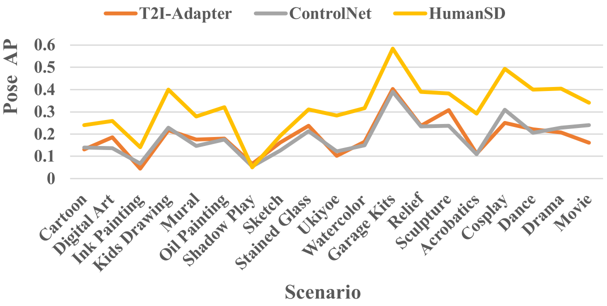


Figure 3: Comparisons of the pose average precision (Y-axis) on different scenarios (X-axis).

HumanSD consistently displays higher AP in all but the shadow play scenario. The lack of shadow play images in the dataset used to train T2I-Adapter, and ControlNet is most likely to blame for this. As unaware of the shadow play scenario, these two methods are more likely to produce images with real people (as shown in the first figure of Figure 1 II in the main paper) as a replacement for shadow play. By cheating the pose accuracy criteria, a falsely high AP score is obtained. From the comparison of different image scenarios, we can lead to two conclusions: (1) The wide variation of APs across different scenarios indicates that the potential challenges in various scenarios are different, and that evaluating and generating multiple scenario images is still challenging. (2) *HumanSD* outperforms ControlNet and T2I-Adapter in all scenarios, especially cosplay, with a boost of 18.2% to ControlNet and 24.1% to T2I-Adapter.

Comparisons on human numbers per image. As shown in Figure 4, the Pose AP value tends to decline with the increase of human number in a single image, which shows the difficulty in generating images with multiple per-

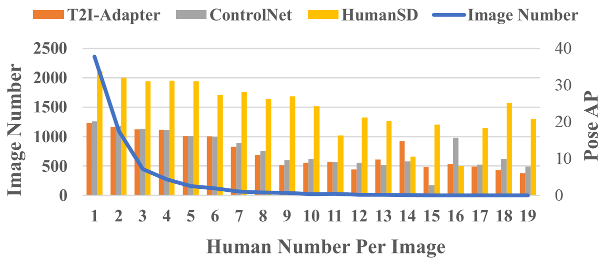


Figure 4: Comparisons of pose average precision (AP) (right Y-axis with the histogram) with 1 to 19 human number per image (X-axis) in the validation set of *Human-Art*. The image number statistic is shown in the left Y-axis with the blue curve to state the total sample size.

sons. *HumanSD* can still retain a high pose AP as the number of humans increases, demonstrating its ability to generate the multi-human image. Moreover, *HumanSD* shows a unified better AP score among images with 1-13 humans than ControlNet and T2I-Adapter. The limited number of images with 14-19 humans lead to fluctuation of Pose AP, but *HumanSD* still shows a relatively better result.

Ablation of Different Pose Estimator Since we focus on multi-scene human image generation, we train pose estimators on the joint of Human-Art and MSCOCO. In Tab.1, we provide various combinations of estimators used for *heatmap-guided denoising loss* and evaluation, results show that AP and PCE are not sensitive to estimators.

Loss Estimator	Evaluation Estimator	AP ↑	PCE ↓
HRNet	HRNet*	31.97	1.62
HRNet*	HRNet*	32.30	1.62
HRNet	HigherHRNet*	32.03	1.64
HRNet*	HigherHRNet*	32.41	**1.55**
ViTPose	HigherHRNet*	30.28	1.57
ViTPose*	HigherHRNet*	31.11	1.61
HigherHRNet*	HigherHRNet*	**32.66**	1.56

[1] We use Faster-RCNN trained on MSCOCO and Human-Art datasets as the human detector for the top-down pose estimator ViTPose.
[2] With / without * means trained on MSCOCO&*Human-Art* / MSCOCO.
[3] Blue means bottom-up estimator. Yellow means top-down estimator.

Table 1: AP and PCE results of *HumanSD* with different pose estimators. For comprehensive comparisons, we use HRNet as the bottom-up estimator and ViTPose as the top-down estimator.

Results of Unbiased FID Since the FID metric has been shown to be biased, we provide results of unbiased FID[2] in Table 2.

Metric	T2I-Adapter	ControlNet	*HumanSD*
FID Infinity ↓	129.55	125.28	**120.89**

Table 2: Comparisons of unbiased FID.

C. Qualitative Results

More qualitative comparison of ControlNet, T2I-Adapter, and *HumanSD* is shown from Figure 5 to 9, which correspondingly shows the natural human scene with only half body, hard poses / small human in the natural scene, text and human orientation controllability in sketch scene, rare scenes such as shadow play and kids drawing, and human detail in oil painting and digital art. We generate three groups of images with different seeds for each text and pose condition to avoid randomness and show diversity.

D. Future Work

Although *HumanSD* has reached a high performance, there are still many issues waiting for exploration. Specifically, future directions include but are not limited to: (1) We notice a significant trade-off between whole-body generation and local body part generation. For example, as shown in Figure 9, the left image can generate high-fidelity human faces. But when we force the model to generate whole body images in the right image, the facial detail retention shows a huge decline, which is extremely obvious in ControlNet. We leave it to future work for solutions. (2) *HumanSD* still fails in extremely crowded scenes and complex/rare actions, as shown in Figure 11. Generation models with higher accuracy and faster speed are still in need. (3) Similar to other generation tasks, the text and pose-guided image generation evaluation system are not yet comprehensive and complete, which entails a lot of randomnesses. (4) augmentations for complex poses and different orientations of humans. We have noticed that human poses that do not frequently appear (e.g., stand upside down) tend to fail more frequently. Appropriate augmentations may alleviate this problem in the future. (5) extending the *heatmap-guided denoising loss* to other conditions. Since our paper primarily focuses on human-centric image generation and relies on the intermediate representation of heatmaps in human pose estimation, the proposed heatmap-guided denoising loss is particularly suitable. However, this concept is not limited to humans or heatmap representations. The primary purpose of the loss is to direct the fine-tuning process of diffusion models to focus on specific regions. For other conditions or objects, we can also enhance the diffusion models by using segmentation maps or sketches to guide the fine-tuning process[3].

[2] https://github.com/mchong6/FID_IS_infinity

[3] https://omriavrahami.com/break-a-scene/

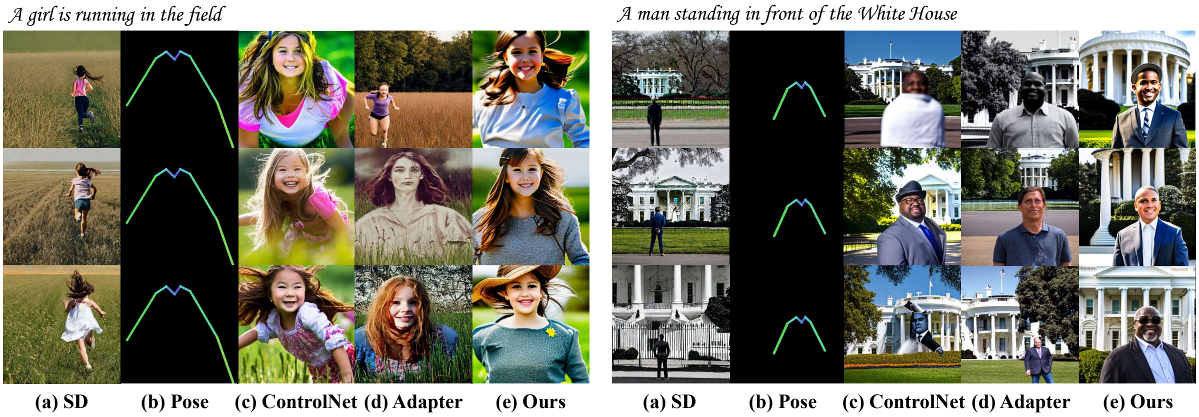


Figure 5: **Natural Human Scene - Half Body**. (a) a generation by the pre-trained text-guided stable diffusion (SD) [35], (b) pose skeleton images as the condition to ControlNet, T2I-Adapter and our proposed *HumanSD*, (c) a generation by ControlNet [52], (d) a generation by T2I-Adapter [27], and (e) a generation by *HumanSD* (ours). ControlNet, T2I-Adapter, and *HumanSD* receive both text and pose conditions. We use three different seeds (the three rows) to generate diverse images.

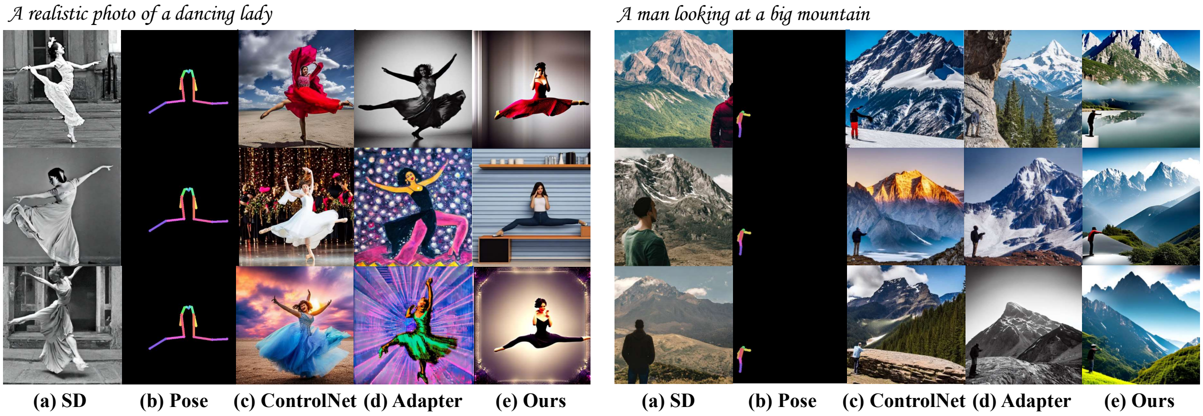


Figure 6: **Natural Human Scene - Hard Poses/Small Human**. The explanation of (a)-(e) can be found in Figure 5's caption.

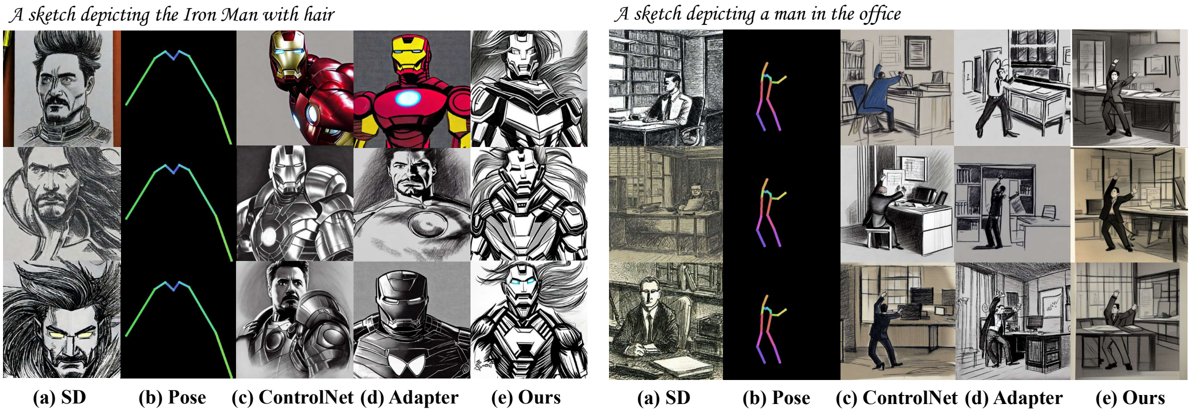


Figure 7: **Sketch Scene - Text Control/Human Orientation**. The explanation of (a)-(e) can be found in Figure 5's caption.

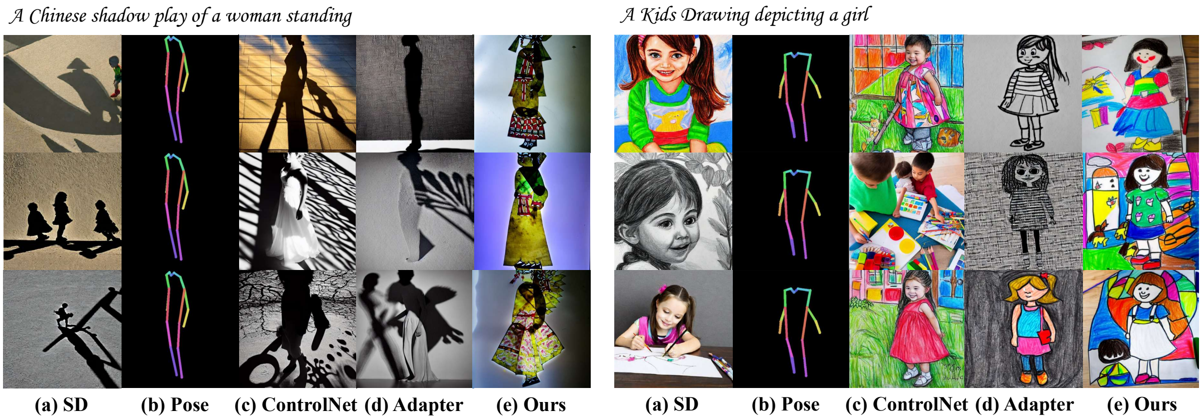


Figure 8: **Shadow Play / Kids Drawing Scene - Rare Scenes**. The explanation of (a)-(e) can be found in Figure 5's caption.

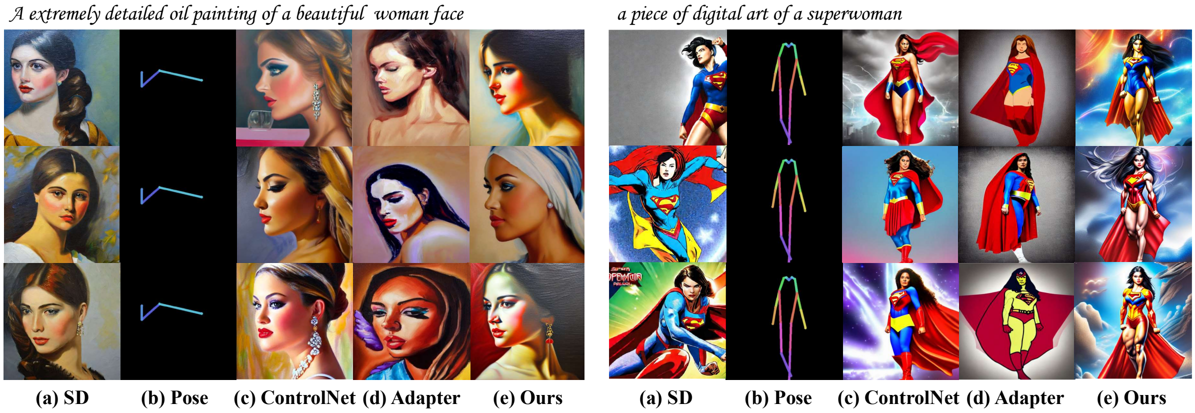

A extremely detailed oil painting of a beautiful woman face
a piece of digital art of a superwoman
(a) SD
(b) Pose
(c) ControlNet
(d) Adapter
(e) Ours
(a) SD
(b) Pose
(c) ControlNet
(d) Adapter
(e) Ours

Figure 9: **Oil Painting / Digital Art Scene - Human Detail**. The explanation of (a)-(e) can be found in Figure 5's caption.

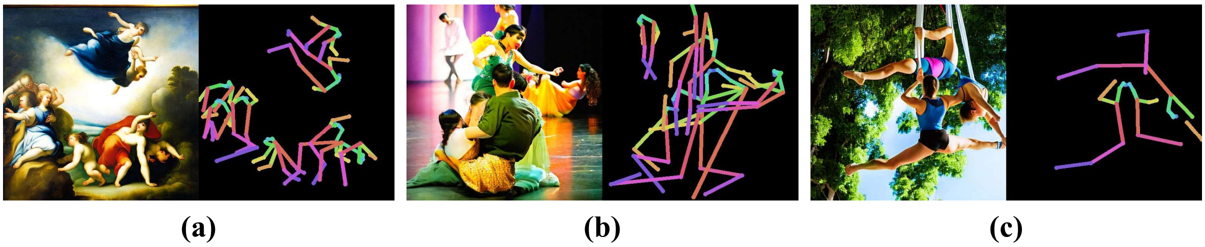


Figure 10: **Failure cases** on (a)/(b) extremely crowded scenes and (c) complex/rare actions.

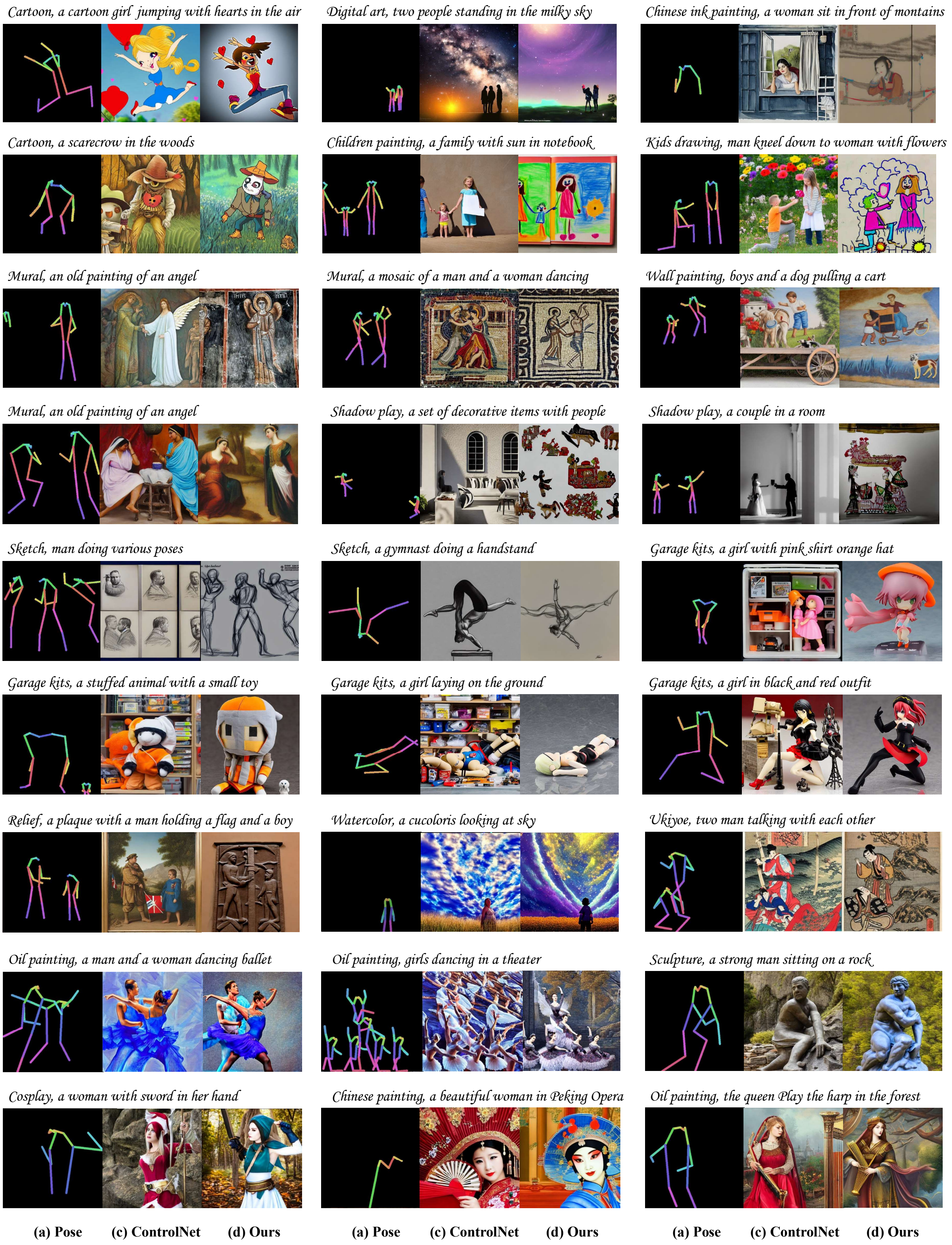

Cartoon, a cartoon girl jumping with hearts in the air
Digital art, two people standing in the milky sky
Chinese ink painting, a woman sit in front of montains
Cartoon, a scarecrow in the woods
Children painting, a family with sun in notebook
Kids drawing, man kneel down to woman with flowers
Mural, an old painting of an angel
Mural, a mosaic of a man and a woman dancing
Wall painting, boys and a dog pulling a cart
Mural, an old painting of an angel
Shadow play, a set of decorative items with people
Shadow play, a couple in a room
Sketch, man doing various poses
Sketch, a gymnast doing a handstand
Garage kits, a girl with pink shirt orange hat
Garage kits, a stuffed animal with a small toy
Garage kits, a girl laying on the ground
Garage kits, a girl in black and red outfit
Relief, a plaque with a man holding a flag and a boy
Watercolor, a cucoloris looking at sky
Ukiyoe, two man talking with each other
Oil painting, a man and a woman dancing ballet
Oil painting, girls dancing in a theater
Sculpture, a strong man sitting on a rock
Cosplay, a woman with sword in her hand
Chinese painting, a beautiful woman in Peking Opera
Oil painting, the queen Play the harp in the forest
(a) Pose
(c) ControlNet
(d) Ours
(a) Pose
(c) ControlNet
(d) Ours
(a) Pose
(c) ControlNet
(d) Ours

Figure 11: **More comparision** between ControlNet and *HumanSD*. (a) pose skeleton images as the condition to ControlNet, T2I-Adapter and our proposed *HumanSD*, (b) a generation by ControlNet [52], and (c) a generation by T2I-Adapter [27].

References

[1] Anything-v4.0. https://huggingface.co/andite/anything-v4.0.

[2] PoseNet similarity. https://github.com/freshsomebody/posenet-similarity.

[3] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. *arXiv preprint arXiv:1801.01401*, 2018. 1

[4] Zhongang Cai, Daxuan Ren, Ailing Zeng, Zhengyu Lin, Tao Yu, Wenjia Wang, Xiangyu Fan, Yang Gao, Yifan Yu, Liang Pan, Fangzhou Hong, Mingyuan Zhang, Chen Change Loy, Lei Yang, and Ziwei Liu. HuMMan: Multi-modal 4D human dataset for versatile sensing and modeling. In *European Conference on Computer Vision (ECCV)*, pages 557–577. Springer, 2022. 2

[5] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2D pose estimation using part affinity fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7291–7299, 2017.

[6] Bowen Cheng, Bin Xiao, Jingdong Wang, Honghui Shi, Thomas S Huang, and Lei 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)*, pages 5386–5395, 2020. 3

[7] Soon Yau Cheong, Armin Mustafa, and Andrew Gilbert. KPE: Keypoint pose encoding for transformer-based image generation. In *British Machine Vision Conference (BMVC)*, 2022. 2

[8] Jihoon Chung, Cheng-hsin Wuu, Hsuan-ru Yang, Yu-Wing Tai, and Chi-Keung Tang. HAA500: Human-centric atomic action dataset with curated videos. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13465–13474, 2021. 2

[9] Nicolas Dufour, David Picard, and Vicky Kalogeiton. SCAM! transferring humans between images with semantic cross attention modulation. In *European Conference on Computer Vision (ECCV)*, pages 713–729. Springer, 2022.

[10] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[11] 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. *Advances in Neural Information Processing Systems (NeurIPS)*, 30, 2017. 1

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

[13] Xuan Ju, Ailing Zeng, Jianan Wang, Qiang Xu, and Lei Zhang. Human-Art: A versatile human-centric dataset bridging natural and artificial scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[14] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[15] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. GLIGEN: Open-set grounded text-to-image generation. *arXiv preprint arXiv:2301.07093*, 2023.

[16] Yong-Lu Li, Liang Xu, Xinpeng Liu, Xijie Huang, Yue Xu, Mingyang Chen, Ze Ma, Shiyi Wang, Hao-Shu Fang, and Cewu Lu. HAKE: Human activity knowledge engine. *arXiv preprint arXiv:1904.06539*, 2019. 2

[17] Jing Lin, Ailing Zeng, Haoqian Wang, Lei Zhang, and Yu Li. One-stage 3d whole-body mesh recovery with component aware transformer. 2023.

[18] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision (ECCV)*, pages 740–755. Springer, 2014. 3

[19] Deyin Liu, Lin Wu, Feng Zheng, Lingqiao Liu, and Meng Wang. Verbal-person nets: Pose-guided multi-granularity language-to-person generation. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[20] 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 (PAMI)*, 42(10):2684–2701, 2019. 2

[21] Vivian Liu and Lydia B Chilton. Design guidelines for prompt engineering text-to-image generative models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–23, 2022. 2

[22] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. DeepFashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1096–1104, 2016.

[23] Zhengyao Lv, Xiaoming Li, Xin Li, Fu Li, Tianwei Lin, Dongliang He, and Wangmeng Zuo. Learning semantic person image generation by region-adaptive normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10806–10815, 2021.

[24] Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool. Pose guided person image generation. volume 30, 2017.

[25] Tianxiang Ma, Bo Peng, Wei Wang, and Jing Dong. MUST-GAN: Multi-level statistics transfer for self-driven person image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13622–13631, 2021.

[26] Yifang Men, Yiming Mao, Yuning Jiang, Wei-Ying Ma, and Zhouhui Lian. Controllable person image synthesis with attribute-decomposed GAN. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5084–5093, 2020.

[27] Chong Mou, Xintao Wang, Liangbin Xie, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2I-Adapter: Learning

adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453*, 2023. 5, 7

[28] Jonas Oppenlaender. Prompt engineering for text-based generative art. *arXiv preprint arXiv:2204.13988*, 2022. 2

[29] Xichen Pan, Pengda Qin, Yuhong Li, Hui Xue, and Wenhu Chen. Synthesizing coherent story with auto-regressive latent diffusion models. *arXiv preprint arXiv:2211.10950*, 2022.

[30] Abhinanda R. Punnakkal, Arjun Chandrasekaran, Nikos Athanasiou, Alejandra Quiros-Ramirez, and Michael J. Black. BABEL: Bodies, action and behavior with English labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 722–731, June 2021. 2

[31] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, pages 8748–8763. PMLR, 2021.

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

[33] Yurui Ren, Xiaoqing Fan, Ge Li, Shan Liu, and Thomas H Li. Neural texture extraction and distribution for controllable person image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13535–13544, 2022.

[34] Yurui Ren, Xiaoming Yu, Junming Chen, Thomas H Li, and Ge Li. Deep image spatial transformation for person image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7690–7699, 2020.

[35] 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 (CVPR)*, pages 10684–10695, 2022. 5

[36] Prasun Roy, Subhankar Ghosh, Saumik Bhattacharya, Umapada Pal, and Michael Blumenstein. TIPS: Text-induced pose synthesis. In *European Conference on Computer Vision (ECCV)*, pages 161–178. Springer, 2022.

[37] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5B: An open large-scale dataset for training next generation image-text models. *arXiv preprint arXiv:2210.08402*, 2022.

[38] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. LAION-400M: Open dataset of CLIP-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021.

[39] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning (ICML)*, pages 2256–2265. PMLR, 2015. 3

[40] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[41] Chenfei Wu, Lun Huang, Qianxi Zhang, Binyang Li, Lei Ji, Fan Yang, Guillermo Sapiro, and Nan Duan. GODIVA: Generating open-domain videos from natural descriptions. *arXiv preprint arXiv:2104.14806*, 2021.

[42] Xiaogang Xu, Ying-Cong Chen, Xin Tao, and Jiaya Jia. Text-guided human image manipulation via image-text shared space. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 44(10):6486–6500, 2021.

[43] Fan Yang and Guosheng Lin. CT-Net: Complementary transferring network for garment transfer with arbitrary geometric changes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9899–9908, 2021.

[44] Jie Yang, Ailing Zeng, Shilong Liu, Feng Li, Ruimao Zhang, and Lei Zhang. Explicit box detection unifies end-to-end multi-person pose estimation. In *International Conference on Learning Representations (ICLR)*, 2023.

[45] Lingbo Yang, Pan Wang, Chang Liu, Zhanning Gao, Peiran Ren, Xinfeng Zhang, Shanshe Wang, Siwei Ma, Xiansheng Hua, and Wen Gao. Towards fine-grained human pose transfer with detail replenishing network. *IEEE Transactions on Image Processing*, 30:2422–2435, 2021.

[46] Ruiyun Yu, Xiaoqi Wang, and Xiaohui Xie. VTNFP: An image-based virtual try-on network with body and clothing feature preservation. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10511–10520, 2019.

[47] Ailing Zeng, Xuan Ju, Lei Yang, Ruiyuan Gao, Xizhou Zhu, Bo Dai, and Qiang Xu. Deciwatch: A simple baseline for 10x efficient 2d and 3d pose estimation. In *ECCV*, 2022.

[48] Ailing Zeng, Xiao Sun, Lei Yang, Nanxuan Zhao, Minhao Liu, and Qiang Xu. Learning skeletal graph neural networks for hard 3d pose estimation. In *ICCV*, 2021.

[49] Ailing Zeng, Lei Yang, Xuan Ju, Jiefeng Li, Jianyi Wang, and Qiang Xu. Smoothnet: A plug-and-play network for refining human poses in videos. In *ECCV*, 2022.

[50] Jinsong Zhang, Kun Li, Yu-Kun Lai, and Jingyu Yang. PISE: Person image synthesis and editing with decoupled GAN. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7982–7990, 2021.

[51] Kaiduo Zhang, Muyi Sun, Jianxin Sun, Binghao Zhao, Kunbo Zhang, Zhenan Sun, and Tieniu Tan. HumanDiffusion: A coarse-to-fine alignment diffusion framework for controllable text-driven person image generation. *arXiv preprint arXiv:2211.06235*, 2022.

[52] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*, 2023. 5, 7

[53] Pengze Zhang, Lingxiao Yang, Jian-Huang Lai, and Xiaohua Xie. Exploring dual-task correlation for pose guided person

image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7713–7722, 2022.

[54] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1116–1124, 2015.

[55] Xinyue Zhou, Mingyu Yin, Xinyuan Chen, Li Sun, Changxin Gao, and Qingli Li. Cross attention based style distribution for controllable person image synthesis. In *European Conference on Computer Vision (ECCV)*, pages 161–178. Springer, 2022.

[56] Shizhan Zhu, Raquel Urtasun, Sanja Fidler, Dahua Lin, and Chen Change Loy. Be your own prada: Fashion synthesis with structural coherence. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1680–1688, 2017.