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

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

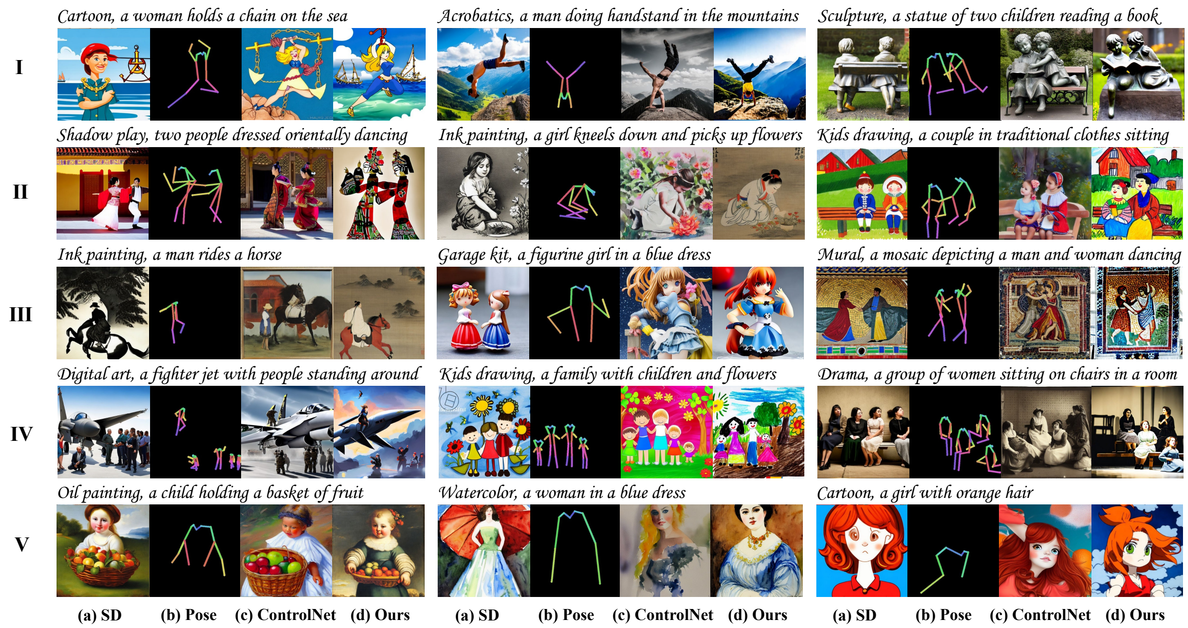


Figure 1: This paper highlights multi-scenario human-centric image generation with precise pose control. Each group of displayed images includes: (a) a generation by the pre-trained pose-less text-guided stable diffusion (SD) [35], (b) pose skeleton images as the condition to ControlNet and our proposed *HumanSD*, (c) a generation by ControlNet [52], and (d) a generation by *HumanSD* (ours). ControlNet and *HumanSD* receive both text and pose conditions. *HumanSD* shows its superiorities in terms of **(I)** challenging poses, **(II)** accurate painting styles, **(III)** pose control capability, **(IV)** multi-person scenarios, and **(V)** delicate details. **Best viewed with zoom-in**.

Abstract

Controllable human image generation (HIG) has numerous real-life applications. State-of-the-art solutions, such as ControlNet and T2I-Adapter, introduce an additional learnable branch on top of the frozen pre-trained stable diffusion (SD) model, which can enforce various conditions, including skeleton guidance of HIG. While such a plug-and-play approach is appealing, the inevitable and uncertain conflicts between the original images produced from the frozen SD branch and the given condition incur significant challenges for the learnable branch, which essentially conducts image feature editing for condition enforcement.

In this work, we propose a ***native*** *skeleton-guided diffusion model for controllable HIG called* **HumanSD**. *Instead of performing image editing with dual-branch diffusion, we fine-tune the original SD model using a novel* **heatmap-guided denoising loss**. *This strategy effectively and efficiently strengthens the given skeleton condition during model training while mitigating the catastrophic forgetting effects.* HumanSD *is fine-tuned on the assembly of three large-scale human-centric datasets with text-image-pose information, two of which are established in this work. Experimental results show that* HumanSD *outperforms ControlNet in terms of pose control and image quality, particularly when the given skeleton guidance is sophisticated. Code and data are available at: https://idea-research.github.io/HumanSD/.*

*Equal contribution. ‡ Work done during an internship at IDEA.
†Corresponding author.

1. Introduction

Controllable human image generation (HIG) aims to generate human-centric images under given conditions such as human pose [24, 33, 45], body parsing [46, 56], and text [19, 36, 42]. It has numerous applications (e.g., animation/game production [29] and virtual try-on [55]), attracting significant attention from academia and industry.

While earlier controllable HIG solutions based on generative adversarial networks (GANs) [10,23–26,33,43,50,53] and variational auto-encoders (VAEs) [7, 14, 34, 45] have been successfully applied in certain applications (e.g., virtual try-on), they have not gained mainstream acceptance due to their training difficulties and poor multi-modality fusion and alignment capabilities [51]. Recently, diffusion models [12, 35, 40] have demonstrated unprecedented text-to-image generation performance [32] and quickly become the dominant technique in this exciting field. However, it is difficult to provide precise position control with text information, especially for deformable objects such as humans.

To tackle the above problem, two concurrent controllable diffusion models were proposed in the literature: ControlNet [52] and T2I-Adapter [27]. Both models introduce an additional learnable diffusion branch on top of the frozen pre-trained stable diffusion (SD) model [35]. The additional branch enables the enforcement of various conditions such as skeleton and sketch during image generation, which greatly improves the original SD model in terms of controllability, thereby gaining huge traction from the community.

However, the learnable branch in such dual-branch diffusion models is essentially performing a challenging image feature editing task and suffers from several limitations. Consider the skeleton-guided controllable HIG problem that generates humans with specific poses. Given text prompts containing human activities, the SD branch may generate various images that are inconsistent with the skeleton guidance, e.g., humans could present at different places with various poses. Therefore, the extra condition branch needs to learn not only how to generate humans according to the given skeleton guidance but also how to suppress various inconsistencies, making training more challenging and inference less stable. Generally speaking, the larger the gap between skeleton guidance and original images produced by the frozen SD branch, the higher discrepancy between the given guidance and generated human images. Moreover, the inference cost of these dual-branch solutions largely increases compared to the original SD model.

In contrast to employing an additional trainable branch for controllable HIG, this work proposes a **native** skeleton-guided diffusion model, named *HumanSD*. By directly fine-tuning the SD model [35] with skeleton conditions concatenated to the noisy latent embeddings, as shown in Figure 2 (a), *HumanSD* can natively guide image generation with the desired pose, instead of conducting a challenging image editing task. To mitigate the catastrophic forgetting effects caused by model overfitting during fine-tuning, we propose a novel *heatmap-guided denoising loss* for diffusion models to disentangle between conditioned humans and unconditioned backgrounds in the training stage. Such a disentanglement forces the fine-tuning process to concentrate on the generation of foreground humans while minimizing unexpected overrides of the pre-trained SD model parameters that hurt the model's generation and generalization abilities.

Besides the algorithm, training data is another important factor determining model performance [38]. To improve the HIG quality of *HumanSD*, we fine-tune our model on three large-scale human-centric datasets containing high-quality images and the corresponding 2D skeletal information and text descriptions: *GHI*, *LAION-Human*, and *Human-Art*. Specifically, *GHI* and *LAION-Human* are established in this work. *GHI* has 1M multi-scenario images generated from SD with crafted prompts, and only the top 30% with the highest image quality are selected. For *LAION-Human*, it selects 1M human-centric images from the LAION-Aesthetics [37] via filtering.

The main contributions of this work include:

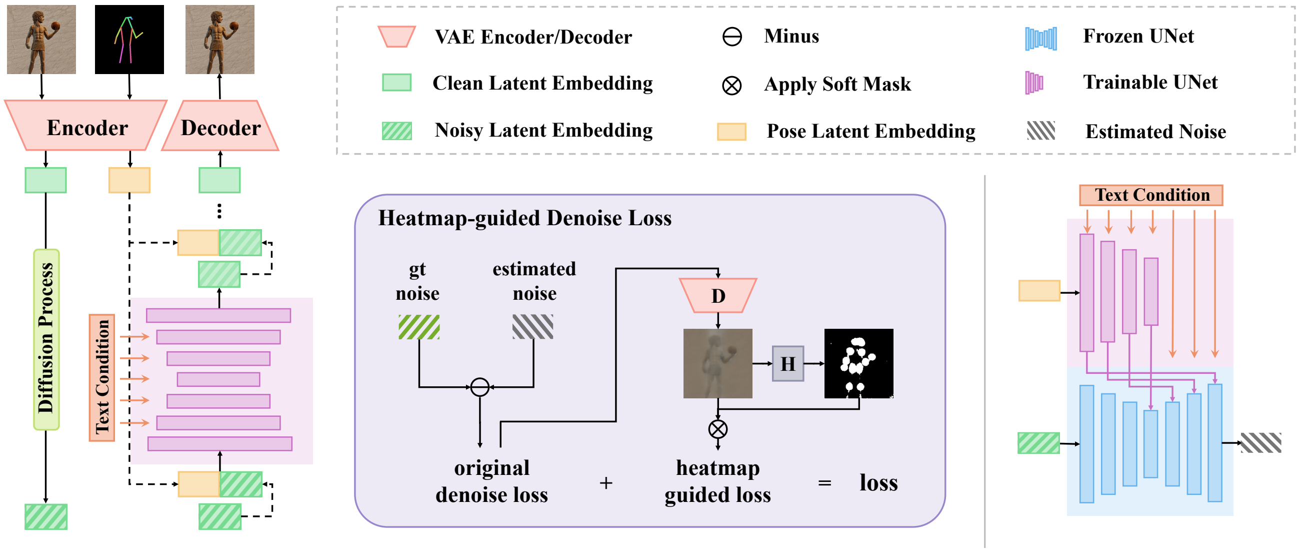
- We propose a new HIG framework *HumanSD* with a novel *heatmap-guided denoising loss*, to natively generate human images with highly precise pose control yet no extra computational costs during inference.
- We introduce two large-scale human-centric datasets with a standard development process, which facilitates multi-scenario HIG tasks with large quantities, rich data distribution, and high annotation quality.
- To demonstrate the effectiveness and efficiency of *HumanSD*, we apply a series of evaluation metrics covering image quality, pose accuracy, text-image consistency, and inference speed to compare our model with previous works in a fair experimental setting.

With the above, *HumanSD* outperforms state-of-the-art solutions such as ControlNet regarding pose control and human image generation quality, particularly when the given skeleton guidance is sophisticated.

2. Related Work

2.1. Pose-Guided Human Image Generation

During the past two decades, pose-guided controllable HIG [7,23–26,33,34,43,45,50,53] has gained lots of attention in academia and industry due to the pose's validity in motion description [17, 18, 44, 47–49]. With source images and pose conditions (e.g., skeletal images or body parsing), pose-guided HIG models output photorealistic images with source images' appearance and desired poses. These algorithms are mainly based on GANs [10,23–26,33,43,50,53]



(a) **(b)**

Figure 2: **Overview**. **(a)** shows **the proposed framework *HumanSD*** with a novel *heatmap-guided denoising loss*. Given the pose condition, our model inputs the corresponding skeletal image into the VAE encoder of a pre-trained SD to get the pose latent embedding. The embedding is then concatenated with the noisy latent embedding generated by diffusion and inputted into the UNet. In the training stage, the *heatmap-guided denoising loss* helps UNet concentrate on the specific areas humans are located in, especially when the human generation has poor performance, by increasing its weight in the loss function. **(b)** shows the recent SOTA method, **ControlNet** [52], which doubles an SD UNet encoder for condition extraction and freezes the original SD branch to maintain image generation ability.

or VAEs [7, 14, 34, 45]. Exclusively focusing on natural-scene manipulation, they fail in diverse cross-modality feature alignment due to limitations in model design, inappropriate condition injection strategies, and lack of diversity in training data, which lead to unrealistic and poor results with artificial scenario source images or arbitrary pose inputs. In addition, these models strongly depend on the source-target paired images that are hard to acquire and lack diversity.

Different from images, text has become a flexible, user-friendly, and informative condition with the rise of large vision-language models [31]. Some works involve text conditions to guide HIG but are limited to small-scale vocabulary pools and fail with open vocabulary [19,36,42]. Among the very recent works, ControlNet [52], T2I-Adapter [27], and GLIGEN [15] introduce methods of adding arbitrary conditions. ControlNet and T2I-Adapter add additional trainable modules to pre-trained text-to-image diffusion models [35]. The target of designing general frameworks makes them not well-targeted to humans that appear with diverse poses, fine-grained body parts, styles, viewpoints, sizes, and quantities. Moreover, their models suffer from trainable-frozen branch conflicts, thus showing inadequate pose control ability. Superior to previous work, *HumanSD* is efficient as well as high-precision in human pose control, and specially designed for open-world multi-scenario HIG.

2.2. Human Image Generation Datasets

Current HIG datasets such as iDesigner [9], DeepFashion [22], Market1501 [54], and MSCOCO [18] mainly focus on the real-scene human generation and provide noisy paired source-target images. These mainstream datasets have limited scenarios (e.g., dress-up, street photography), and are not generalizable to other scenarios such as cartoons, oil paintings, and sculptures.

Recently, *Human-Art* [13] provides 50K human-centric images in five natural and fifteen artificial scenes with precise pose and text annotations. Specially designed for multi-scenario human-centric tasks, *Human-Art* is suitable for validating the quality and diversity of existing generation methods. However, the limited data scale of *Human-Art* makes it inadequate for large model training. Laion-5B [37] is a publicly available dataset with sufficient text-image paired data but contains many human irrelevant images. ControlNet [52] adopts the human pose estimator OpenPose [5] on internet-scratched images to collect 200K pose-image-text pairs, most of which are real-scene images. Using these data pairs in training will lead to a significant distribution bias towards real yet low-diversity scenes.

This work provides a standard development process for large-scale multi-scenario text-image-pose datasets targeted at skeleton-guided HIG, which addresses the absence of suitable training and testing datasets.

3. Preliminaries and Motivation

This section introduces the details of the conflicts in recent SOTA SD-based HIG methods before outlining the motivation for designing *HumanSD* in section 4. These methods, notably ControlNet and T2I-Adapter, use the Latent Diffusion Model (LDM [35]) as the foundation for its high trainability and high-generation quality, which will be introduced in Section 3.1. Then, Section 3.2 states the conflicts in ControlNet and T2I-Adapter. Since these two models have a similar design, we take ControlNet as an example.

3.1. Preliminaries - The Latent Diffusion Model

LDM, more known as Stable Diffusion (SD), is a diffusion model [39] conducted on latent embeddings instead of images. Images are projected into latent embeddings by a VAE and then guided by text conditions in the latent space. LDM has a latent-space loss function with a similar form to vanilla diffusion models [12, 40]:

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

where z_0 is the latent embedding of a training sample x_0; ϵ_θ and ϵ are respectively the noise estimated by the UNet θ and the ground truth noise injected at the corresponding diffusion timestep t; c is the embedding of all conditions involved in the generation; $\bar{\alpha}_t$ is the same coefficient as that in vanilla diffusion models.

3.2. Conflicts in Dual-Branch Solution

In this section, we provide more detailed theoretical analyses on ControlNet [52] condition addition strategy. We argue that the conflict between the behavior of the frozen image-generation branch and the trainable condition-injection branch results in the degradation of pose control.

As shown in Figure 3, ControlNet is a plug-and-play approach for conditional image generation. It clones an SD branch to extract hierarchical features from the added condition and freezes the original SD branch to preserve generation ability. The trainable and frozen neural network blocks are connected with a convolution layer. The convolution layer takes trainable features as input and its output is added to the frozen features.

We denote the feature in the UNet of the original and the additional SD branches as $f_\theta^O(z, c_T, t)$ and $f_\theta^A(z, c, t)$, where c_T is the text condition, $c = c_T + c_P$ is the ensemble of c_T and pose condition c_P. Note that noise $\epsilon_\theta^O(z, c_T, t)$ and $\epsilon_\theta^A(z, c, t)$ can be viewed as the feature output by the last UNet layer. As shown in Figure 3, $f_\theta^O(z, c_T, t)$ can be divided into a positive part $f_\theta^{O+}(z, c_T, t)$ and a negative part $f_\theta^{O-}(z, c_T, t)$, based on their consistency with c_P.

$$f_\theta^O(z, c_T, t) = f_\theta^{O+}(z, c_T, t) + f_\theta^{O-}(z, c_T, t) \tag{2}$$

For dual-branch models with text-and-pose-guided generation, an ideal estimated feature $\bar{f}_\theta$ should satisfy:

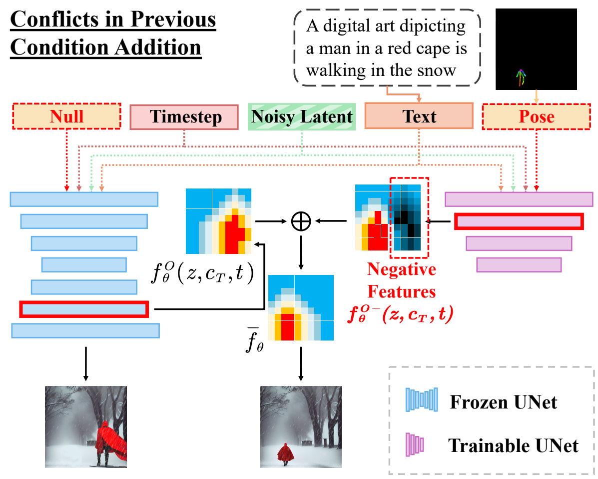


Figure 3: An example showing the **conflicts** in the behavior of the two branches in ControlNet [52].

$$\bar{f}_\theta = f_\theta^{O+}(z, c_T, t) + \tilde{f}_\theta^{+}(z, c, t), \tag{3}$$

where $\tilde{f}_\theta^{+}(z, c, t)$ ensures fine-grained pose control that cannot be guaranteed by $f_\theta^{O+}(z, c_T, t)$. We also have,

$$\bar{f}_\theta = f_\theta^O(z, c_T, t) + f_\theta^A(z, c, t) \tag{4}$$

Thus, we can obtain the feature in the additional SD branches as follows:

$$f_\theta^A(z, c, t) = \tilde{f}_\theta^{+}(z, c, t) - f_\theta^{O-}(z, c_T, t) \tag{5}$$

This leads to indirect noise generation during inference, where the additional (trainable) branch has to learn how to (1) identify the positive and negative parts of the estimated noise given the pose condition, (2) suppress the negative part, and (3) generate the extra positive part. The frozen SD branch results in a permanent existence of conflicts between the negative part and the extra positive part. In contrast, for fine-tuning-based methods with all parameters trainable, the models go through a smooth and stable training process, and naturally learn to process the pose conditions and the cross-condition balance, thus avoiding the conflict.

4. Method

To resolve the conflict in previous SD-based methods, we introduce *HumanSD*, a native skeleton-guided diffusion model for precise and efficient multi-scenario human image generation. Vanilla fine-tuning faces the problem of catastrophic amnesia and over-fitting. To address this issue, we propose a condition addition strategy with a novel loss, which is illustrated in Section 4.1 and Section 4.2. Lastly, we provide a dataset construction process for multi-scenario Human-centric Image Generation in Section 4.3.

4.1. Skeleton Condition Addition

As shown in figure 2 (a), our proposed *HumanSD* adds pose condition using a skeleton image with the same size as the input image, which provides explicit position information. In order to align the pose conditions with the latent embeddings of the input images, the skeleton image is then processed with the VAE encoder. Different from text conditions, we do not add pose latent embedding with attention in each UNet block, but directly concatenate it to the noisy latent embeddings. This ensures that information on the same density level is processed at the same stage, which results in improved structure information integration.

4.2. The Heatmap-guided Denoise Loss

Fine-tuning deep neural networks with no protection can easily lead to catastrophic forgetting, where the performance of previous tasks drastically degrades when learning to perform a new task. Directly fine-tuning diffusion models with new data and new conditions leads to the same problem (e.g., Anything Model [1], which is fine-tuned from SD to generate anime images, is unable to produce images in other styles). Such performance degradation partially results from the non-discriminatory learning on all pixels of the image. This, to some extent, is reasonable for the conditions with global information (e.g., general text descriptions). However, for conditions with local structure information (e.g., pose condition with specific position information), fine-tuning the whole image results in a quality decline of condition-invariant regions (e.g., background).

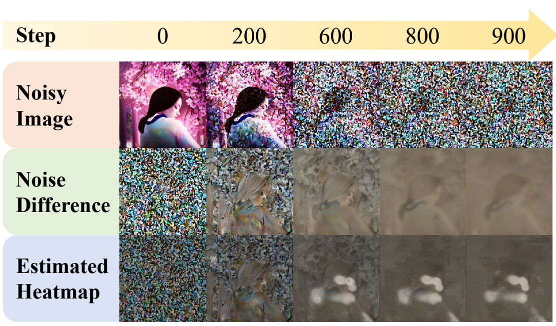


Figure 4: An illustration explaining **the calculation of** W_a in different diffusion steps.

To address this problem, we propose a *heatmap-guided denoising loss* to fine-tune the diffusion model in a safe mode when adding a new structure-aware condition, which pays special attention to the training of the newly added condition and leaves the condition-invariant parts of the image to the pre-trained backbone, thus reaching high performance in both generation quality and condition-image consistency. The *heatmap-guided denoising loss* takes effect by explicitly adding an aggregated heatmap weight W_a to the original loss of the diffusion model. The loss function is then modified from Equation 1 to Equation 6.

$$L_{\mathrm{h}} = \mathop{\mathbb{E}}_{t,z,\epsilon}\left[\left\|W_a \cdot \left(\epsilon - \epsilon_\theta\left(\sqrt{\bar{\alpha}_t}z_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, c, t\right)\right)\right\|^2\right] \quad (6)$$

One of the most straightforward designs for W_a is to assign bigger priority factors for feature pixels that are more related to the condition. However, diffusion is a step-by-step noise addition process, and not all steps are essential to condition injection. Therefore, assigning a constant weight map in all steps may disrupt the training process.

As a result, we need to (1) find out what the model learns at different steps and stages, and (2) determine a weight function $W_a(t)$ based on the step-wise model behavior. The first row of Figure 4 shows the decoded noisy latents in different steps; the second row shows the corresponding differences between the estimated noise and its ground truth (determined in the diffusion process), and the third row shows the corresponding heatmaps generated by a pre-trained human pose heatmap estimator [6] with the noise difference as inputs. Using the heatmap as the description of W_a, the diffusion model can learn better with greater concentration on the condition (human pose). More detailed implementation of *heatmap-guided denoising loss* can be found in Figure 2 (a) and supplementary materials.

4.3. The Dataset Construction Process

Diffusion models require enormous amounts of data for training and fine-tuning. To ensure diverse data distribution in image scenes, human actions, and appearances, we introduce a standard dataset development process and construct 2 large-scale datasets *GHI* and *LAION-Human*. Figure 5 illustrates examples and characteristics of each dataset.

GHI: It is an abbreviation for **G**enerated **H**uman **I**mages. Directly sampling data from SD's own learned distribution is a good way to maintain SD's generation capability with no new data distribution introduced. In order to maximize the exploitation of potential image possibilities in SD, we take advantage of prompt engineering [21, 28] to design prompts that are constructed with 18 sub-prompt parts including image scene style, human number, human characteristics, action, and background descriptions (e.g., a realistic pixel art of two beautiful young girls running in the street of Paris at midnight, in winter, 64K, a masterpiece.). We use pose estimator [6] trained on *Human-Art* to detect character poses in diverse scenes. Then filter out images with wrong human numbers, multi-arms and legs [7], and low body integrity based on detection results. The selection strategy ensures *GHI* contains relatively clean annotations for text and pose. We also do NSFW check to mitigate data biases. This leads to a total number of 1M pose-image-text pairs that include 14 scenes (taken from *Human-Art*) and 6826 human actions (taken from BABEL [30], NTU RGB+D 120 [20], HuMMan [4], HAA500 [8], and HAKE-HICO [16]) with 1 to 3 humans (with proportion of 7:2:1) in each image.

LAION-Human: Similar with ControlNet [52] and T2I-Adapter [27], we construct a dataset *LAION-Human* containing large-scale internet images. Specifically, we collect about 1M image-text pairs from LAION-5B [37] filtered by the rules of high image quality and high human estimation confidence scores. Superior to ControlNet and T2I-Adapter, we adopt a versatile pose estimator trained on *Human-Art*, which allows for selecting more diverse images such as oil paintings and cartoons. Importantly, *LAION-Human* contains more diverse human actions and more photorealistic images than data used in ControlNet and T2I-Adapter.

Human-Art: *Human-Art* [13] contains 50k images in 20 natural and artificial scenarios with clean annotation of pose and text, which can provide precise poses and multi-scenario for training and quantitative evaluation. We follow *Human-Art*'s setting to divide the training and testing sets.

Unless otherwise stated, we train *HumanSD* on the ensemble of *GHI*, *LAION-Human*, and the training set of *Human-Art* (denote as *Union* in following sections), and test on the validation set of *Human-Art*.

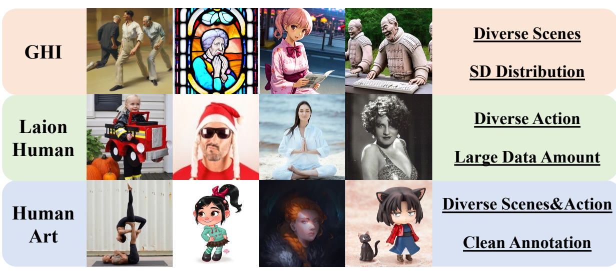


Figure 5: **Examples and characteristics** of the used datasets *GHI* & *LAION-Human*, and *Human-Art* [13].

5. Experiments

In this section, we validate that *HumanSD* outperforms currently SOTA SD-based (Section 5.2) and GAN-based (Section 5.3) methods on skeleton-guided HIG with 8 evaluation metrics explained in Section 5.1. Section 5.4 provides ablation studies on the *heatmap-guided denoising loss*, training datasets, and training iterations. Please refer to the supplementary material for implementation details.

5.1. Evaluation Metrics

To illustrate the effectiveness and efficiency of our proposed *HumanSD*, we use eight metrics covering four aspects: image quality, pose accuracy, text-image consistency, and inference time.

Image Quality: We report Fréchet Inception Distance (FID [11]) and Kernel Inception Distance (KID [3]), which are widely used to measure the quality of the syntheses. Specifically, we evaluate FID and KID on each *Human-Art* scenario and report the mean value, which reflects both quality and diversity of the generation.

Pose Accuracy: We adopt distance-based Average Precision (AP) [18], Pose Cosine Similarity-based [2] AP (CAP) and People Count Error (PCE) [7]. These metrics measure the difference between the given pose condition and the pose result extracted from the generated image. Distance-based AP evaluates the keypoint-wise distances between the ground truth and the generated pose. We also provide AP(m) for medium-sized humans (with resolutions ranging from 32^2 to 96^2 following MSCOCO [18]). We calculate CAP by simply replacing the distance error with the normalized cosine similarity error [2] in AP to evaluate the position-aligned similarity between the given pose and the generated pose. CAP eliminates the effect of absolute position and concentrates on pure action similarity. PCE measures the difference between the number of given skeletons and the generated humans. It effectively evaluates multi-person image generation, and partially reflects inconsistency [7] in single-person image generation, such as false numbers of heads, arms, and legs.

Text-image Consistency: The CLIP [31] Similarity (CLIPSIM [41]) evaluates text-image consistency between the generated images and corresponding text prompts. CLIPSIM projects text and images to the same shared space and evaluates the similarity of their embeddings.

Inference Time: We test inference time per image on one NVIDIA A100 80G to evaluate efficiency. Results are averaged over 20 random runs with batch size 1.

5.2. Comparison with SD-based Methods

We compare *HumanSD* with the very recent SOTA model ControlNet [52] and T2I-Adapter [27]. To ensure fairness, we report results trained on a subset of our proposed *LAION-Human*, including 0.2 million (0.2M) images with a data distribution similar to ControlNet and T2I-Adapter's training datasets in Table 1.

The superiority of *HumanSD* in pose controllability is validated by its remarkable performance in pose-related metrics. Compared with the best results among ControlNet and T2I-Adaptor, *HumanSD* (0.2M) shows a 34.8% to 109.1% performance boost on pose accuracy (e.g., AP, AP(m), and PCE). A combination of better condition injection and the *heatmap-guided denoising loss* leads to such performance enhancements. The results interpret the inevitable conflicts in ControlNet stated in Section 3.2. As indicated previously, with such conflicts, ControlNet may frequently be disrupted by the negative features that exist in the frozen branch, and fail to faithfully render the given pose. Instead, the native generation process and the *heatmap-guided denoising loss* of *HumanSD* simplify the pose guidance and ensure generation quality. Figure 1 I and Figure 1 III further demonstrate *HumanSD*'s expertise in handling challenging poses.

Specifically, since text prompts barely indicate the position (corresponding to the metric AP) and the size (corre-

Models \ Metrics	Image Quality FID ↓	KID $_{(\times 10)}$ ↓	Pose Accuracy AP ↑	AP(m) ↑	CAP ↑	PCE ↓	Text-image Consistency CLIPSIM ↑	Inference Time Second per Image ↓
Stable Diffusion	41.55	2.99	0.09	0.00	49.05	1.76	**33.86**	**3.88**
T2I-Adapter [27]	29.07	2.67	18.20	11.93	55.98	2.73	33.29	5.28
ControlNet [52]	27.24	2.60	18.45	11.71	57.18	2.47	33.16	6.37
HumanSD (0.2M)	**26.28**	**2.56**	**31.85** $_{72.6\%\uparrow}$	**24.95** $_{109.1\%\uparrow}$	**59.11** $_{3.3\%\uparrow}$	**1.61** $_{34.8\%\downarrow}$	32.98	3.89

Table 1: **Quantitative comparisons between *HumanSD* and other SD-based models** (fair comparisons). *HumanSD* is trained for around 300 GPU hours (95K iterations) on 0.2M text-image-pose pairs randomly selected from *LAION-Human*, similar to T2I-Adapter and ControlNet. Results demonstrate *HumanSD*'s effectiveness and efficiency.

sponding to the metric AP(m)) of each person-to-generate, there undoubtedly exists a large behavior gap between the text-guided frozen branch and the pose-aware trainable branch in ControlNet and T2I-Adapter. This is validated by the almost zero scores of SD on AP and AP(m). Accordingly, ControlNet and T2I-Adapter undoubtedly face severe conflicts related to such properties. Unfortunately, the conflicts aggravate with the increasing complexity of the conditions or the scenarios (e.g., see Figure 1 IV). This is validated by their even worse performance on multi-person generation than SD reflected by PCE.

Moreover, *HumanSD* infers much faster compared to ControlNet and T2I-Adapter thanks to its single branch design. Although T2I-Adapter uses a more efficient trainable branch than ControlNet, introducing an additional condition learning branch is still time-consuming. The compression of its condition learning branch also leads to quality decline compared with ControlNet, as shown in Table 1.

The disruption in ControlNet may be trivial for CAP, as the pose information given by the text prompts and the skeletal images are likely to be similar (e.g., the pose indicated by the text 'standing' may be very similar to the actual skeleton of a person standing). Therefore, SD can reach a certain level of CAP even without pose conditions, and the improvement of *HumanSD* on CAP is relatively small.

For image quality, the three models show similar FID and KID, indicating that they all manage to preserve SD's basic abilities of image generation and text comprehension. *HumanSD* achieves such performance by concentrating on specific human regions in fine-tuning, while ControlNet and T2I-Adapter achieve this via their frozen SD branches. SD has relatively the worst style-wise FID and KID scores, indicating the disqualification of text information in guiding high-quality and diverse human-centric image generation.

However, both *HumanSD* and ControlNet show a performance decline in text-image consistency within a reasonable and acceptable range. Such negligible yet existent degradation lies in the potential inconsistency between the high-level text and the low-level pose conditions. It is worthwhile to explore how to better balance pose and text coherent control capabilities.

5.3. Comparison with GAN-based Methods

This section shows the incapability of previous GAN-based HIG methods on precise and diverse pose control. We compare Neural Texture Extraction and Distribution (NTED [33]) and Text-Induced Pose Synthesis (TIPS [36]), which are both textless pose-guided real-scene image generation methods. For NTED and TIPS, we use images randomly selected from DeepFashion [22] as source image inputs and the skeleton maps from the validation set of *Human-Art* as pose conditions. For *HumanSD*, we use text and images from the validation set of *Human-Art*.

As shown in Figure 6, NTED and TIPS easily fail given unconventional pose conditions. Specifically, NTED and TIPS have an AP score of 2.79 and 17.65 with untrained poses as input. Thus, we can conclude that previous GAN-based HIG methods are not qualified for open-scenario poses. This reflects the significance of *HumanSD* with precise pose control and multi-scenario generation ability.

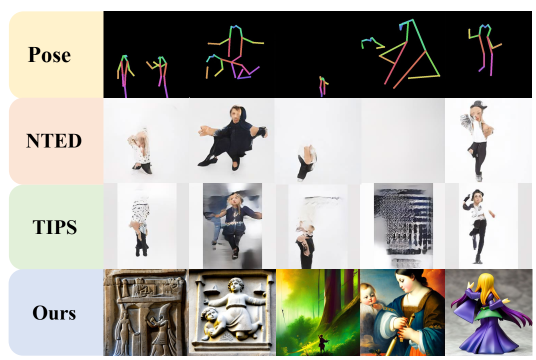


Figure 6: **Examples generated by previous methods** NTED and TIPS given unconventional pose conditions.

5.4. Ablation Study

In this section, we demonstrate that apart from the better condition injection that allows for a more native generation and avoids conflicts, the proposed *heatmap-guided denoising loss* also contributes to the better performance of *HumanSD*. Also, we explore how the proposed training datasets and the number of training iterations influence the final results. Unless otherwise stated, the model is trained on Union for around 300 GPU hours (95K iterations).

Impact of the heatmap-guided denoising loss. As shown in Table 2, adding the *heatmap-guided denoising loss* helps the back-propagation focus on optimizing weights more related to human generation. This further leads to more precise human pose guidance and thus boosts the AP score from 30.63 to 32.66. Meanwhile, focusing on the human generation improves the background's preservation. It thus safeguards the non-human-associated image information to be more related to text descriptions, which increases CLIPSIM from 32.55 to 32.98.

Model	**AP ↑**	**PCE ↓**	**CLIPSIM ↑**
w/ proposed loss	**32.66**	**1.56**	**32.99**
w/o proposed loss	30.63	1.57	32.55

Table 2: **Ablation on the loss function.**

Figure 7 shows qualitative visualizations of the impact of the *heatmap-guided denoising loss*. The *heatmap-guided denoising loss* contributes to more precise pose controllability (II(c), II(f)), better human detail fidelity (I(c), II(c)), improved text-image consistency (I(c), II(f)), and enhanced background quality (I(c), I(f), II(c), II(f)). Notably, *HumanSD* also generates remarkable results on humanoid figures like robots and animals (I(f)).

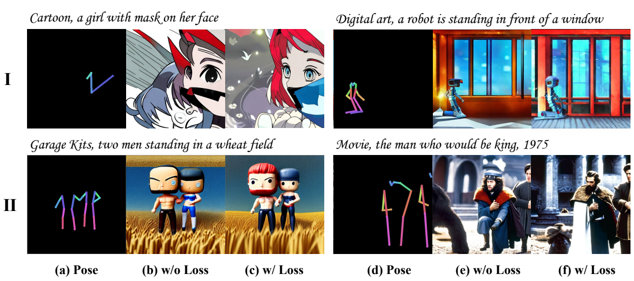


Figure 7: **Visualization of generated results w/ and w/o the *heatmap-guided denoising loss*.**

Impact of training datasets. To show the validity of the proposed datasets, we provide results on three training dataset settings. As shown in Table 3, using *GHI* alone can guarantee the generation with the most accurate human numbers and stronger text-image consistency. This is primarily owing to *GHI*'s better data distribution alignment with SD. However, due to the absence of real images, results generated by the model trained with *GHI* show low generation quality (e.g., blurred human limbs, unrealistic human structure), resulting in a low AP score. Training on *LAION-Human* can achieve relatively more satisfactory AP results. Compared with *LAION-Human*, combining all datasets to train a model can further improve AP performance, and obtain better trade-offs among PCE and CLIPSIM due to the increase in diversity.

Impact of training iterations. Fine-tuning iterations of *HumanSD* have a significant impact on generation results. In our experiments, the performance increases and then fluctuates at around 95K iterations (around 300 GPU hours). Table 4 shows the training results for 50K, 95K, and 150K iterations.

Dataset	**AP ↑**	**PCE ↓**	**CLIPSIM ↑**
Union	**32.66**	1.56	32.99
GHI	23.66	**1.50**	**33.61**
LAION-Human	31.93	1.60	32.98

Table 3: **Ablation on training datasets.** Details of the datasets are introduced in Section 4.3

Iterations	**AP ↑**	**PCE ↓**	**CLIPSIM ↑**
50K	25.47	1.58	32.39
95K	**32.66**	1.56	**32.99**
150K	32.17	**1.55**	32.51

Table 4: **Ablation on fine-tuning iterations.**

6. Conclusion

In this work, we have presented a new framework, named *HumanSD*, based on pre-trained SD for highly precise pose and text-conditioned human image generation. To concentrate on the generation of foreground humans and preserve pre-trained SD's generation ability, we proposed a novel *heatmap-guided denoising loss*. Moreover, we introduced large-scale human-centric datasets containing over 2M text-image-pose pairs for multi-scenario human-centric generative learning. Finally, we compared *HumanSD* with previous models on a series of evaluation metrics covering image quality, pose accuracy, and text-image consistency. Results demonstrate the effectiveness and efficiency of our proposed method and datasets.

Although the proposed *HumanSD* improves controllability for HIG, it still encounters certain limitations. (1) *HumanSD* still easily fails in image generation with extremely crowded scenes and complex/rare actions. (2) Despite being filtered, the large-scale text-image-pose pairs we trained on may still contain social biases and violent content. (3) The evaluation system is not yet comprehensive and complete.

Finally, we hope that this work can motivate future research with a focus on HIG for higher controllability, richer human scenarios, more conditions, and better image quality.

Acknowledgements

This work was supported in part by the Shenzhen-Hong Kong-Macau Science and Technology Program (Category C) of the Shenzhen Science Technology and Innovation Commission under Grant No. SGDX2020110309500101, in part by Research Matching Grant (CSE-7-2022) - RMG01, and in part by the Innovation and Technology Fund under Grant No. MRP/022/20X.

References

[1] Anything-v4.0. https://huggingface.co/andite/anything-v4.0. 5

[2] PoseNet similarity. https://github.com/freshsomebody/posenet-similarity. 6

[3] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. *arXiv preprint arXiv:1801.01401*, 2018. 6

[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. 5

[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. 3

[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. 5

[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, 3, 5, 6

[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. 5

[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. 3

[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. 2

[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. 6

[12] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:6840–6851, 2020. 2, 4

[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. 3, 6

[14] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 2, 3

[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. 3

[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. 5

[17] Jing Lin, Ailing Zeng, Haoqian Wang, Lei Zhang, and Yu Li. One-stage 3d whole-body mesh recovery with component aware transformer. 2023. 2

[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. 2, 3, 6

[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. 2, 3

[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. 5

[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. 5

[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. 3, 7

[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. 2

[24] Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool. Pose guided person image generation. volume 30, 2017. 2

[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. 2

[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. 2

[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. 2, 3, 5, 6, 7

[28] Jonas Oppenlaender. Prompt engineering for text-based generative art. *arXiv preprint arXiv:2204.13988*, 2022. 5

[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. 2

[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. 5

[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. 3, 6

[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. 2

[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. 2, 7

[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. 2, 3

[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. 1, 2, 3, 4

[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. 2, 3, 7

[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. 2, 3, 6

[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. 2

[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. 4

[40] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 2, 4

[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. 6

[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. 2, 3

[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. 2

[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. 2

[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. 2, 3

[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. 2

[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. 2

[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. 2

[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. 2

[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. 2

[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. 2

[52] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*, 2023. 1, 2, 3, 4, 5, 6, 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. 2

[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. 3

[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. 2

[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. 2