This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Towards Effective Usage of Human-Centric Priors in Diffusion Models for Text-based Human Image Generation

Junyan Wang [1] Zhenhong Sun [2] Zhiyu Tan [3] Xuanbai Chen [4] Weihua Chen [5]
Hao Li [6*] Cheng Zhang [4] Yang Song [1]

[1] University of New South Wales [2] Australian National University [3] INF Technology
[4] Carnegie Mellon University [5] Alibaba DAMO Academy [6] Fudan University

Abstract

Vanilla text-to-image diffusion models struggle with generating accurate human images, commonly resulting in imperfect anatomies such as unnatural postures or disproportionate limbs. Existing methods address this issue mostly by fine-tuning the model with extra images or adding additional controls — human-centric priors such as pose or depth maps — during the image generation phase. This paper explores the integration of these human-centric priors directly into the model fine-tuning stage, essentially eliminating the need for extra conditions at the inference stage. We realize this idea by proposing a human-centric alignment loss to strengthen human-related information from the textual prompts within the cross-attention maps. To ensure semantic detail richness and human structural accuracy during fine-tuning, we introduce scale-aware and step-wise constraints within the diffusion process, according to an in-depth analysis of the cross-attention layer. Extensive experiments show that our method largely improves over state-of-the-art text-to-image models to synthesize high-quality human images based on user-written prompts. Project page: `https://hcplayercvpr2024.github.io`.

1. Introduction

Recent advancements in diffusion models have significantly improved **text-to-image** (T2I) generation, consistently enhancing the quality and precision of visual synthesis from textual descriptions [28, 31, 36, 39]. Within the paradigm of T2I, generating human images emerges as a specific focus, drawing substantial attention for its potential in applications such as virtual try-on [53] and entertainment [27]. Despite the remarkable advancements, human image generation still faces challenges, including the incomplete rendering of the human body, inaccuracies in the portrayal, and limb disproportions, such as the imperfect case shown in Figure 1. The challenges in generating human images arise from the diffusion model's inherent emphasis on broad generalization across diverse data, leading to a lack of detailed attention to human structure in the generated images. Resolving these issues is essential for advancing the field toward producing more realistic and accurate human images from textual descriptions.

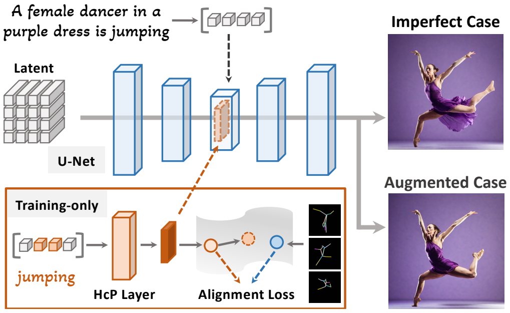

Figure 1. Existing text-to-image models often struggle to generate human images with accurate anatomy (upper branch). We incorporate human-centric priors into the model fine-tuning stage to rectify this imperfection (bottom branch). The learned model can synthesize high-quality human images from text without requiring additional conditions at the inference stage.

The straightforward method to tackle the challenges in human image generation involves using additional conditions during both the training and inference phases, *e.g.,* ControlNet [48]. While employing extra conditions like pose image guidance indeed improves the structural integrity of human, their reliance on additional conditions does not address the challenges inherent in human image generation. It restricts the direct creation of diverse images from text prompts, and requires extra conditions beyond text during inference, making the process tedious and less end-user friendly. Alternatively, another efficient approach employs fine-tuning methods, *e.g.,* LoRA [17], which adjust pre-trained models on specialized human-centric datasets

[*] Corresponding author

for more accurate human feature representation. While this approach can enhance human image generation, it may modify the model's original expressiveness and lead to catastrophic forgetting, resulting in outputs that are limited by the characteristics of the fine-tuning dataset.

Thus, our work concentrates on **text-based Human Image Generation** (tHIG), which relies exclusively on textual inputs without requiring additional conditions during inference. The primary objective of tHIG is to address the challenges in human image generation within diffusion models, enhancing their expressive power while leveraging the inherent diversity and simplicity of diffusion models to generate human images without additional conditions. To tackle the challenges in human image generation, we delved into several key factors for influencing the final output. Firstly, echoing the findings from [14], our analysis shows the role of *cross-attention maps* within diffusion models is a fundamental element, significantly impacting the structural content. This impact is particularly crucial in the generation of human body structures, where accurate representation depends critically on these maps' effectiveness. Furthermore, incorporating *human-centric priors*, such as pose image, has been shown to enhance human representation in synthesized visuals [19]. Aligning this with the inherent capabilities of existing T2I models provides a solid foundation for generating more realistic human figures.

Building on the outlined motivations, our work introduces a novel plug-and-play method for tHIG, which emerges from comprehensive insights into the diffusion process, with a particular focus on the crucial role of *cross-attention maps*. We present the innovative **H**uman-**c**entric **P**rior (HcP) layer, designed to enhance the alignment between the cross-attention maps and human-centric textual information in the prompt. Incorporating a specialized Human-centric Alignment loss, the HcP layer effectively integrates other auxiliary human-centric prior information, such as key poses, exclusively during the training phase. This inclusion improves the capability of the diffusion model to produce accurate human structure only with textual prompts, without requiring additional conditions during inference. Furthermore, our approach adopts a step and scale aware training strategy, guided by our in-depth analysis of the cross-attention layers. This strategy effectively balances the structural accuracy and detail richness in generated human images, while preserving the diversity and creativity inherent to T2I models.

We validate our HcP layer with Stable Diffusion [35]. The HcP layer can preserve the original generative capabilities of the diffusion model and produce high-quality human image generation without requiring additional conditions during the inference phase. Moreover, the HcP layer is compatible with the existing controllable T2I diffusion models (*e.g.,* ControlNet [48]) in a plug-and-play manner.

2. Related Work

Text-to-Image Generation. T2I as a rapidly evolving field, has witnessed the emergence of numerous model architectures and learning paradigms [3, 5–7, 21, 25, 31, 32, 39, 41, 43, 45, 46, 50–52]. Generative Adversarial Networks (GANs) based models [29, 40, 54] initially played a pivotal role in this field, establishing key benchmarks for quality and diversity in generated images. Recent advancements [28, 31, 35, 36] in diffusion models have significantly enhanced the capabilities of text-to-image generation. Diffusion models derive their effectiveness from a structured denoising process [16], which transforms random noise into coherent images guided by textual descriptions. For example, latent diffusion [35] utilizes a latent space-based approach where it first converts text into a latent representation, which is then progressively refined into detailed images through a denoising process. In this work, we build upon these diffusion model advancements by introducing HcP layer, specifically designed to enhance HIG.

Human Image Synthesis. Human image synthesis is an area of significant interest due to its broad applications in industries such as fashion [11, 12] and entertainment [27]. Most efforts [4, 19, 22, 26, 33, 34, 42, 47–49] to address the challenges of diffusion models in accurately representing human structure have relied on introducing additional conditions during both training and inference stages. For example, HumanSD [19] proposes a native skeleton-guided diffusion model for controllable human image generation by using a heatmap-guided denoising loss. However, this approach often complicates the image generation process and can limit the diversity of output images. Our work introduces the HcP layer and employs a targeted training strategy that enhances human image synthesis in diffusion models without additional conditions, which ensures the high-quality generation of human images.

Image Editing via Cross-Attention. Recent advancements in text-driven image editing have shown significant progress, especially within the paradigm of diffusion-based models [1, 9, 20, 23]. Kim *et al.* [20] show how to perform global changes, whereas Avrahami *et al.* [1] successfully perform local manipulations using user-provided masks for guidance. Progress in text-driven image editing primarily relies on refining the cross-attention layers within U-Net architectures [10, 14, 44]. For example, the work of Hertz *et al.* [14] presents several applications which monitor the image synthesis by editing the textual prompt only. This includes localized editing by replacing a word, global editing by adding a specification, and even delicately controlling the extent to which a word is reflected in the image. However, our approach enhances the influence of certain text embeddings during image generation, ensuring efficiency without additional conditions at inference.

3. The Proposed Approach

Our approach starts with an in-depth analysis of the observation from the behavior of the cross-attention layer during the diffusion process. Based on this analysis, we propose the Human-centric Prior layer with Human-centric Alignment loss to infuse human-centric prior knowledge. Subsequently, we detail the training strategy on both scale and step aware aspects. Figure 4 illustrates the procedure associated with the proposed HcP layer in the pre-trained latent diffusion model.

3.1. Analysis of Cross-Attention Layer

For the tHIG task, the aim is to generate a diverse set of images using a given text-to-image generation model driven by human-authored prompts. However, there exist certain issues in the context of generating human images, such as structural inaccuracies and inconsistent body proportions. As demonstrated in [14], the detailed structures in the generated images crucially depend on the interaction between the pixels and the text embedding at the cross-attention layers of U-Net. Consequently, we further examine the relationship between human body structures and each text token embedding in human image generation through the cross-attention process. For instance, in challenging cases like the prompt "*a young woman doing yoga on the beach*," we observe significant issues in rendering accurate human poses and proportions, as illustrated in Figure 2. Note that all observations and analysis are conducted on the publicly available Stable Diffusion v1-5 model [35].

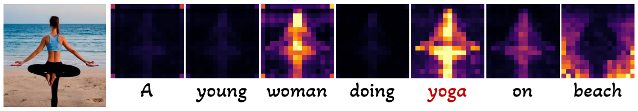

Figure 2. Average cross-attention maps across all timestamps of a text-conditioned diffusion process. These maps contain semantic relations with texts that affect the generated image, exemplified by the inaccurate duplication of legs in the generated human figure.

We can see that the cross-attention map corresponding to "*woman*" and "*yoga*" closely reflects the human pose, and the map for "*beach*" corresponds to background. This strong correlation between attention maps and texts indicates that cross-attention layers, guided by specific text embeddings, play a pivotal role in shaping the semantic content of the image. This also implies that insufficient capabilities of cross-attention layers can affect the results of generated images. Building on this observation, we conduct a comprehensive analysis as shown in Figure 3, to identify and address the underlying causes of prevalent issues in tHIG.

Step-wise Observation. The inference of the diffusion model is essentially a denoising process. Given each step in the diffusion process incrementally refines the output image, it's essential to analyze the impact of early vs. later steps, especially in the context of human-centric image generation. As illustrated in Figure 3, the anatomical structure of the human subject becomes distinguishable in the very early steps. Later steps, while refining and enhancing the image, primarily focus on the optimization of finer details rather than significant structural alterations. This indicates the role of the initial steps is determining the overall structure and posture of the generated human figure, while later steps work on refining details to improve the final output.

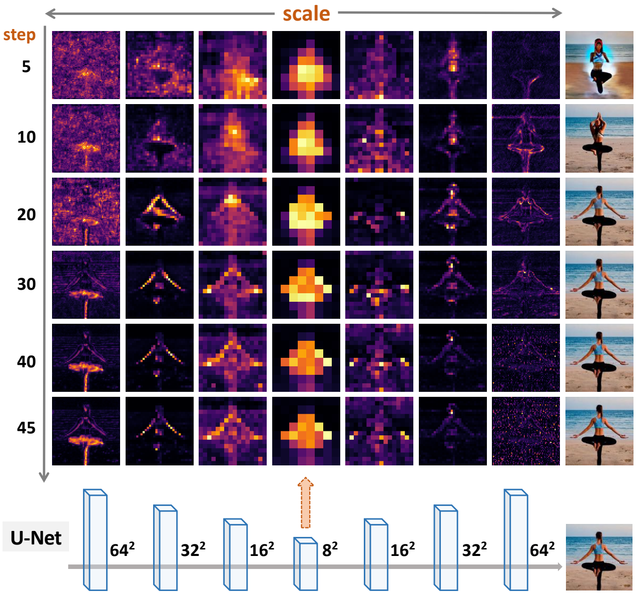

Figure 3. The cross-attention maps, as influenced by the fixed token 'yoga', are across various stages of the U-Net architecture at different inference timesteps. The vertical axis represents the inference timestep when using DDIM [38], while the horizontal axis corresponds to the different scale stages within the U-Net framework. The right side displays generated images at each step.

Scale-wise Observation. Based on our step-wise observations, we further investigate the role of resolution scale in synthesizing human images, particularly within the U-Net architecture of diffusion models. As illustrated in Figure 3, we observe that as the resolution decreases (towards the middle of the U-Net architecture), mid-stage timesteps predominantly determine the structural aspects of the human figure. At the smaller resolution scale, located at the midpoint of the U-Net, all timesteps collectively influence the human structure, with early timesteps playing a more significant role. Conversely, as the resolution increases again (moving towards the output layer), the early timesteps become key in defining the structure. These observations underscore the complexity inherent in the cross-attention layers and the pivotal role of different scales and steps in the human image generation process.

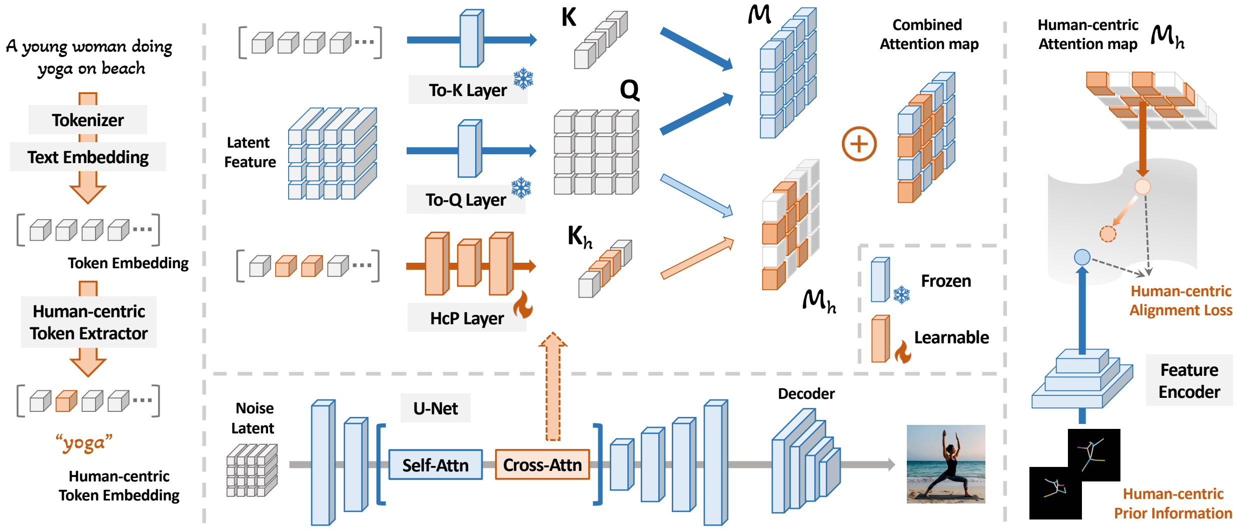

Figure 4. Overview of the proposed **learnable** Human-centric Prior layer training in the **frozen** pre-trained latent diffusion model. The left part shows the process of human-centric text tokens extraction, the middle part indicates the overall process of the HcP layer plugged into the U-Net framework, and the right part shows the HcP layer training with the proposed human-centric alignment loss.

3.2. Human-centric Prior Layer

As we discussed in Section 3.1, text embeddings related to humans and actions significantly influence the human structure in the generated image, which is particularly evident within the associated cross-attention maps. Therefore, we suggest that by enhancing the sensitivity of diffusion models to human-centric textual information during the denoising process, we can improve the structural accuracy and details in the generated images. To do this, we propose an additional learnable module, the Human-centric Prior (HcP) layer, to strengthen the interactions between the latent features and the human-centric textual within the cross-attention maps. This module is integrated without altering the pre-existing expressive capacity of the cross-attention layers, whose parameters remain frozen during training.

Within the latent diffusion framework, the structure allows the cross-attention layer to effectively incorporate textual information into the image synthesis process. Specifically, in this cross-attention mechanism, query $\mathbf{Q}$ represents the latent representation, capturing the spatial attributes at a specific resolution stage. On the other hand, both key $\mathbf{K}$ and value $\mathbf{V}$ are derived from the text-conditioned embeddings $\mathcal{C} = \{\mathcal{C}_1, \mathcal{C}_2, \dots, \mathcal{C}_n\}, \mathcal{C} \in \mathbb{R}^{N \times D}$, where N and D denote text token length and embedding dimension. Subsequently, we introduce an additional “Key” into the cross-attention mechanism, denoted as $\mathbf{K}_h$. This key is also derived from the text embeddings $\mathcal{C}$ via the HcP layer which is composed of multiple MLP networks. Then, $\mathbf{K}_h$ interacts with the Query, generating the human-centric attention map $\mathcal{M}_h$ as:

$$\mathcal{M}_h = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}_h^T}{\sqrt{d}}\right), \; \mathbf{K}_h = \phi(\mathcal{C}_h) \;, \tag{1}$$

where $\phi(\cdot)$ represents the transformation carried out by the HcP layer and d indicates the latent projection dimension of the keys and queries. Then the forward attention map of the cross-attention layer in the pre-trained denoising network is defined as the combination of the human-centric attention map $\mathcal{M}_h$ and the original attention map $\mathcal{M}$:

$$\hat{\mathcal{M}} = \gamma\mathcal{M} + (1-\gamma)\mathcal{M}_h \;, \tag{2}$$

where γ denotes the attention combination weight. Note that the HcP layer is a plug-and-play module that can be combined with any cross-attention layers. This integration not only preserves the expressive power of the existing pre-trained U-Net, but also addresses the issues of human structure generation within the image synthesis process. Subsequent subsections will describe the training process for the HcP layer to incorporate human-specific information.

3.3. Human-centric Alignment Loss

Acknowledging the diffusion model’s deficiency in focusing on the details of human structure, we focus on enhancing human-specific information within the HcP layer. Meanwhile, key pose images, effective in representing human body structures, are leveraged as essential sources of human-centric prior information. Consequently, we have designed a novel loss function that aligns this human-centric prior information with the HcP layer, thereby addressing the structural challenges in human image generation.

Concretely, a pre-trained entity-relation network is first deployed to extract human-centric words from textual prompts. For instance, *woman* and *yoga* from the phrase “*A young woman doing yoga on beach*”. Upon identifying human-centric terms, we only train corresponding indices within the attention map. This selective approach ensures the training focus of the human-centric attention map to the relevant regions. We then utilize a pre-trained encoder, such as ResNet50, to extract features $\mathbf{H}$ from the corresponding key pose images that provide a reference for human-centric characteristics. These features are aligned with the human-centric attention map $\mathcal{M}_h$, facilitated by a specially designed Human-centric Alignment Loss. This loss is computed using cosine distance, formulated as:

$$\mathcal{L}_{hca}(\mathbf{H}, \mathcal{M}_h) = \frac{1}{|\mathcal{I}_h|} \sum_{i \in \mathcal{I}_h} [1 - \mathcal{D}(\mathbf{H},\ \mathcal{M}_h[i])] \ , \tag{3}$$

where $\mathcal{D}(\cdot, \cdot)$ denotes the cosine similarity function and $|\mathcal{I}_h|$ is the count of human-centric word indices. By minimizing the cosine distance in this manner, the human-centric attention map becomes more focused on human-centric prior information, as illustrated in the right part of Figure 4. Notably, refinement is constrained to areas related to relevant tokens, with human-centric prior information directing the synthesis of human structures.

3.4. Scale & Step Aware Learning

Our detailed scale and step analysis in the inference phase (Section 3.1) reveal a critical insight that the formation of human structure is closely linked to the resolution scale at different U-Net stages. Based on this observation, we introduce a learning strategy that addresses the unique scale and step characteristics observed in the U-Net architecture. In this work, we first partition the U-Net of the Stable Diffusion v1-5 model into three distinct stages: *down*, *mid*, and *up*. This partition reflects the different resolution scales within the U-Net, as shown in Figure 5.

In order to dynamically adjust the loss weights λ at each stage of the U-Net, we utilize the **cosine function**, specifically adapted to the distinct characteristics of each scale. The formula for this dynamic adjustment is expressed as:

$$\lambda^l(t) = \begin{cases} \cos\left(\dfrac{t}{\mathbf{T}} \cdot \dfrac{\pi}{2}\right), & \text{if } l \in \text{down-scale} \\ \cos\left(\dfrac{t - \mathbf{T}}{\mathbf{T}} \cdot \dfrac{\pi}{2}\right), & \text{if } l \in \text{mid-scale} \\ \cos\left(\dfrac{2t - \mathbf{T}}{\mathbf{T}} \cdot \dfrac{\pi}{2}\right), & \text{if } l \in \text{up-scale} \end{cases} \tag{4}$$

where l denotes the cross-attention layer number in U-Net. For the down-scale stage, the loss weight follows a cosine function that varies in a straightforward manner with the progression of timestep t relative to the maximum timestep $\mathbf{T}$. This adjustment significantly impacts the human structural aspect at early timesteps. For the mid-scale stage, where the resolution is lower, the loss weight is adjusted through a cosine function centered around the midpoint of the timesteps. This adjustment allows a higher emphasis on the later ones. For the up-scale stage, as the resolution increases, the cosine function is designed to rapidly emphasize the middle timesteps, highlighting their importance in defining the human structure as the resolution scales up.

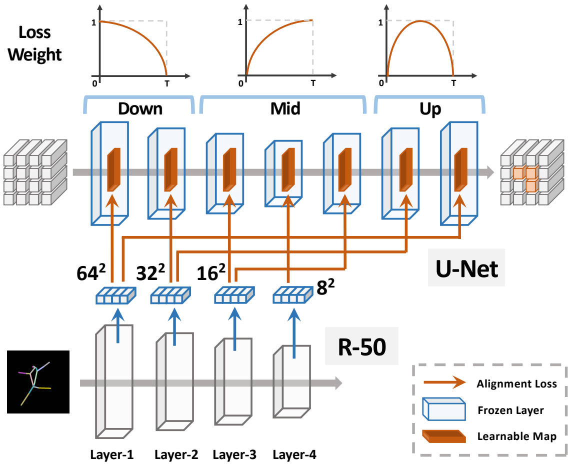

Figure 5. Alignment of layer-specific ResNet features with corresponding scale ($[64^2, 32^2, 16^2, 8^2]$) human-centric attention maps in each cross-attention layer of the U-Net architecture for human-centric alignment loss

This strategy is designed to optimize the learning process by adapting to the specific requirements at different scales and steps, as revealed in our prior cross-attention map analysis. It adjusts the learning focus, transitioning between structural definition and detailed texturing in accordance with the resolution scale.

Overall Optimization. Meanwhile, the denoising loss is also incorporated into the training process to further ensure the quality of the synthesized images. Therefore, the overall optimization objective can be expressed as follows:

$$\mathcal{L}_{ldm} = \mathbb{E}_{x, \epsilon \sim \mathcal{N}(0,1)}[||\epsilon - \epsilon_\theta(z_t, t)||_2^2] \ , \tag{5}$$

$$\mathcal{L}^t = \alpha \sum_{l \in L} (\lambda^l(t) \cdot \mathcal{L}^l_{hca}) + \mathcal{L}_{ldm} \ , \tag{6}$$

where L denotes the number of U-Net layers and α denotes the human-centric alignment loss weight. In contrast to other approaches, our method preserves the original generative capabilities of the model without altering its expressive power and focuses on refining the human structure within the generated images to ensure a more reasonable representation. Meanwhile, it operates without extra inputs, thereby maintaining diversity in the generative process.

4. Experiments

We validate the proposed HcP layer for HIG in various scenarios and introduce the experimental setup in Section 4.1, present the main results in Section 4.2, and detailed ablations and discussions in Section 4.3 and 4.4. Please see the Appendix for additional results and analyses.

4.1. Setup

Datasets. (1) *Human-Art* [18] contains 50k images in 20 natural and artificial scenarios with clean annotation of pose and text, which provide precise poses and multi-scenario for both training and quantitative evaluation. (2) *Laion-Human* [18] contains 1M image-text pairs collected from LAION-5B [37] filtered by the rules of high image quality and high human estimation confidence scores.

Evaluation Metrics. To comprehensively illustrate the effectiveness of our proposed method, we adopt three different types of metrics: (1) *Image Quality*: Frechet Inception Distance (FID) [15] and Kernel Inception Distance (KID) [2] to measure the quality of the syntheses. (2) *Text-image Consistency*: CLIP-Score [30] to evaluate text-image consistency between the generated images and corresponding text prompts. (3) *Human Evaluation*: This further evaluates the anatomy quality and examines the consistency between the text-image pairs using human's subjective perceptions.

Baselines. We compare HcP layer to the following methods. (1) *Stable Diffusion* (SD) v1-5 [35] without any modification. (2) *Low-rank Adaptation* (LoRA) [17] fine-tuned with SD model on both Human-Art training set and Laion-Human set. Additionally, we also compare with *ControlNet* [48] using the *OpenPose* condition, and *SDXL-base* [28].

Implementation Details. The total trainable parameters are from the proposed HcP layer which consists of three 1024-dimensional MLP blocks. We choose key pose image as human-centric prior information and use pre-trained ResNet-50 [13] as the human-centric prior information extractor. To align the scale of each layer's features of ResNet-50 with those from the cross-attention layer in U-Net, the input pose images are resized to 256 × 256. We select the top eight features with the highest variance across channels from the last four stages of ResNet50. These are leveraged as the multi-heads for the cross-attention layer of U-Net, with the head number set to 8. During training, we use the AdamW optimizer [24] with a fixed learning rate of 0.0001 and weight decay of 0.01, and we set $\gamma = 0.9$ and $\alpha = 0.1$ for loss control. In the inference stage, we adopt DDIM sampler [38] with 50 steps and set the guidance scale to 7.5. All experiments are performed on 8 × Nvidia Tesla A100 GPUs. More implementation details in Appendix C.

4.2. Main Results

We validate the superiority of the HcP layer by combining with pre-trained SD and making comparisons with vanilla SD, and SD enhanced with LoRA, from both qualitative and quantitative perspectives.

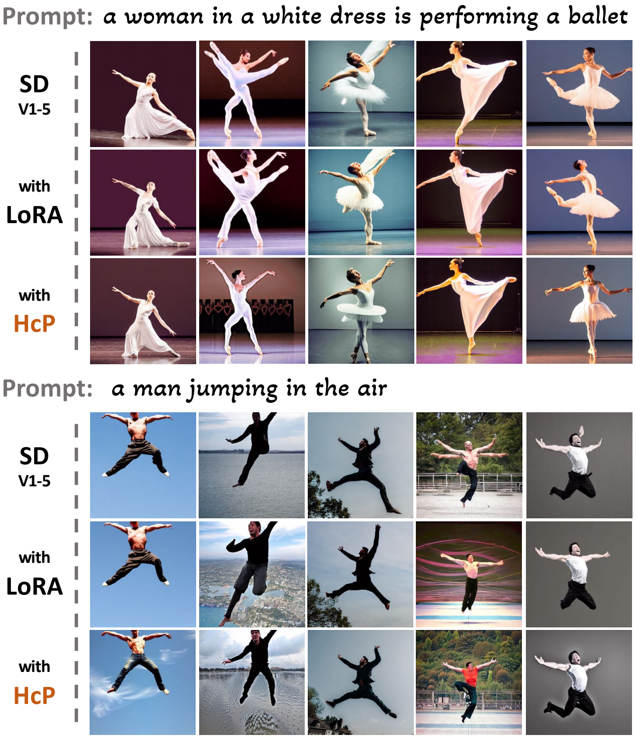
Prompt: a woman in a white dress is performing a ballet
SD V1-5
with LoRA
with HcP
Prompt: a man jumping in the air
SD V1-5
with LoRA
with HcP

Figure 6. **Qualitative comparison with baseline methods on two example prompts**. We leverage the pre-trained SD v1-5 model for both "with LoRA" and "with HcP" models while keeping it frozen. More examples across domains are included in the Appendix E.4.

Qualitative Evaluation. As shown in Figure 6, for simpler actions like "*jumping*", the pre-trained SD enhanced with LoRA shows improved quality in human image generation, but its effectiveness diminishes with more complex actions such as "*ballet*". Furthermore, LoRA somehow alters the depiction of the original diffusion model, especially for background content, indicating that it enhances the generation of human structures while simultaneously affecting the model's intrinsic capability to represent scenes. In contrast, our proposed method with the HcP layer shows consistently accurate human structure generation across a variety of actions, both simple and complex. Notably, our method retains the original expressive power of the pre-trained SD more effectively, maintaining both the background content and human structure more closely aligned with the original model, reflecting a more focused enhancement. This evaluation demonstrates the effectiveness of the HcP layer in addressing human image structure issues without significantly altering the model's overall image synthesis capabilities.

Quantitative Evaluation. According to the results in Table 1, the image quality metrics reveal that our HcP method does not compromise the original generation quality of the SD v1-5 model. Furthermore, our approach achieves a more significant increase in CLIP-Score compared with LoRA

Table 1. **FID, KID, and CLIP-Score results on Human-Art validation datasets.** ↓ indicates that lower FID and KID are better, reflecting higher image quality; ↑ denotes higher CLIP-Score indicating better alignment with textual descriptions.

Method	Quality		Consistency
	FID ↓	KID$_{\times 1k}$ ↓	CLIP-Score ↑
SD	33.31	9.38	31.85
+ LoRA	29.22	5.83	31.91
+ HcP	**28.71**	**5.62**	**32.72**

Table 2. **Human evaluation on the real-human category of Human-Art dataset.** Participants were asked to rate every pair by using a 5-point Likert scale (1 = poor, 5 = excellent), considering *anatomy quality* (AQ) and *text-image alignment* (TIA).

Method	Acrobatics		Cosplay		Dance		Drama		Movie	
	AQ	TIA	AQ	TIA	AQ	TIA	AQ	TIA	AQ	TIA
SD	1.6	2.2	3.5	4.1	2.0	2.5	2.0	1.8	3.0	3.4
+ LoRA	1.8	2.2	3.6	4.1	2.1	2.5	2.0	2.5	3.0	3.5
+ HcP	**2.7**	**3.5**	**3.8**	**4.3**	**3.5**	**4.0**	**3.2**	**2.6**	**3.1**	**3.6**

fine-tuning. This improvement underscores the efficacy of the HcP layer in refining human structure generation, ensuring a more accurate depiction of human poses and proportions in alignment with the textual descriptions.

Human Evaluation. To further verify the efficacy of HcP, we invited participants to evaluate our prompt-generated image pairs under the guidelines of multimedia subjective testing [8]. To be specific, we use different methods to generate 200 images for different domains in the Human-Art dataset. The results presented in Table 2 demonstrate a significant improvement in generating human figures for complex domains ('acrobatics', 'dance', and 'drama') using our method, compared to both SD and LoRA. Additionally, our approach yields comparable results to these methods in simpler domains ('cosplay' and 'movie'). These findings further validate the effectiveness of our proposed method in improving the capability of the diffusion model to produce a more accurate human structure and meanwhile retaining the original expressive power of the pre-trained diffusion model. More details can be seen in Appendix E.1.

4.3. Ablation Study

Different Timestamp Stages. During the training phase with the DDPM, which involves a maximum of 1000 timesteps, we selectively apply the human-centric alignment loss in different time segments: *early*, *middle*, and *late* phases, as shown in Figure 7. When the human-centric alignment loss is introduced during the early timesteps, the influence on human image generation is comparatively minimal. Essentially, applying the alignment loss too early fails to fully leverage the human-centric prior information. Conversely, when applied during the mid or late timesteps, the human-centric alignment loss affects the generation of human structures. It leads to the creation of more accurate human images through efficiently utilizing human-centric prior information. This finding aligns with our inference stage observation in Section 3.1, which the initial steps are crucial in establishing the overall structure and posture of the generated human image, while later steps work on refining details to improve the quality of the final output.

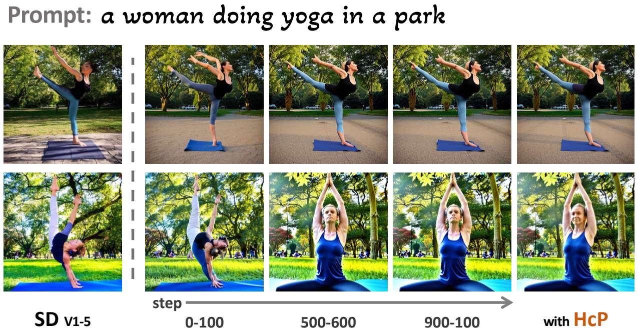

Figure 7. **Ablation on different timestamp stages**. The middle three images are the outcomes of training the model in three distinct phases (0-100, 500-600, and 900-1000 timesteps) without the cosine function for scale adjustments.

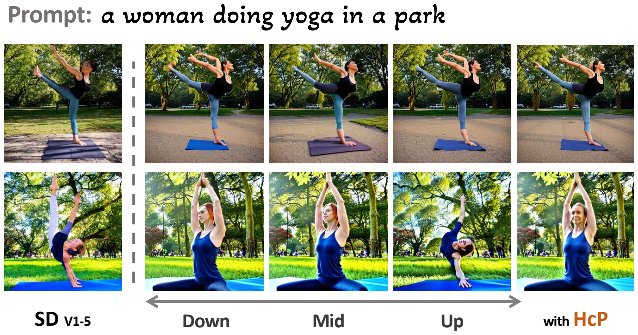

Figure 8. **Ablation on cosine function in different scale stages**. The middle three images are the outcomes of training the model at the down, mid, and up scales without cosine function adjustments.

Scale-Aware Training. In this validation, we separately excluded the cosine function adjustment at the down-scale, mid-scale, or up-scale stages of the U-Net, as results shown in Figure 8. As illustrated, the absence of the cosine function adjustment in the mid-scale leads to outcomes nearly unchanged from the final images, though with certain limitations. This corroborates our observation that at the smaller resolution scale, all timesteps collectively contribute to shaping the human structure. Significant deviations in results are observed when the cosine function adjustment is not applied in either the up or down scales, especially in the up-scale, which reinforces our observation regarding the distinct influence of different scale stages. Meanwhile, these further validate the appropriateness of applying cosine function adjustments at each scale in the U-Net architecture.

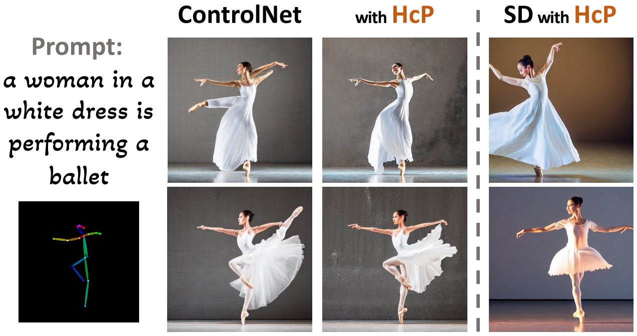

Figure 9. **Comparisons and compatibility with the controllable HIG application**. By plugging the HcP layer trained on pre-trained SD into ControlNet [48], our method can further boost both the quality and consistency compared with the original ControlNet-OpenPose model.

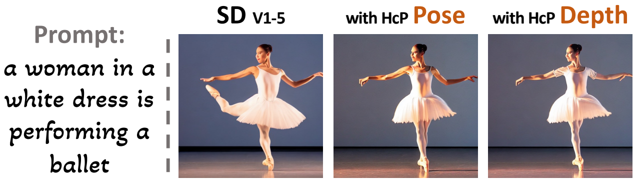

Figure 10. **Comparisons by using different sources of human-centric prior information**. The middle and right images utilize pose and depth images as the human-centric prior information respectively. More results can be seen in Appendix E.3.

4.4. Discussion

Controllable HIG. Considering the adaptable design of our proposed HcP layer as a plug-and-play approach, it can also be extended to Controllable HIG applications. According to Figure 9, despite having a defined pose, ControlNet still encounters challenges in accurately generating human structures. Interestingly, by simply plugging the proposed HcP layer, which is fine-tuned only on the SD instead of the ControlNet model, into the ControlNet, human images with more precise structure and posture are obtained. Moreover, even utilizing only the pre-trained SD model with the HcP layer without relying on any additional condition in the inference phase, our method can acquire comparable results and ensure diverse generations based on only textual inputs. More comparisons can be seen in Appendix E.2.

Human-centric Prior Information. In Figure 10, we utilize depth images as an alternative source of human-centric prior information. The results demonstrate that depth images are also effective in correcting inaccuracies in human image generation. While depth prior can enhance the detailing of the generated images, they tend to slightly alter the details of the original human image, such as the textures of clothing, compared to pose images. In future work, we plan to investigate how to use multiple types of human-centric prior information to optimize the balance between detail enhancement and structural accuracy for generated images.

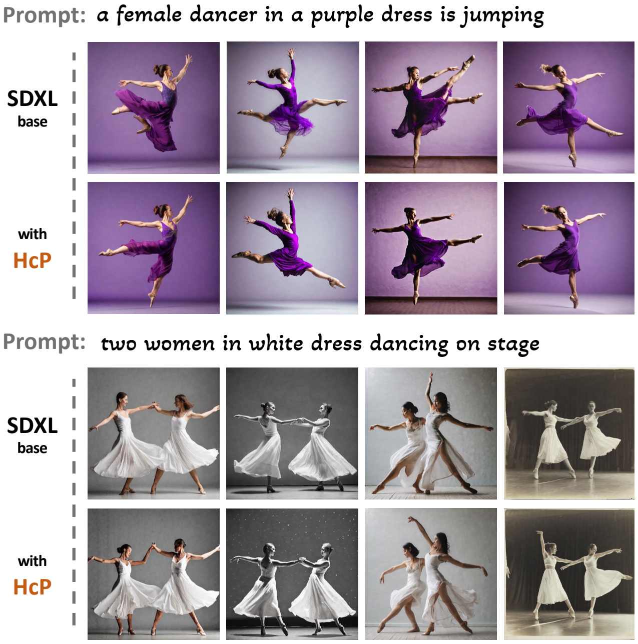
Prompt: a female dancer in a purple dress is jumping
SDXL base
with HcP
Prompt: two women in white dress dancing on stage
SDXL base
with HcP

Figure 11. **Results on larger diffusion model using HcP layer.** We leverage the pre-trained SDXL-base model for the "with HcP" model while keeping it frozen. More examples of different scenarios are included in the Appendix E.4.

Large Diffusion Model. To assess the effectiveness of our method on larger vision models, we evaluated it on SDXL-base, as shown in Figure 11. The results demonstrate that, while SDXL generally produces human images with better structure and detail compared to SD v1-5, it still exhibits some issues. For example, the proportions of the legs are not harmonious in the first image, and extra legs are in other figures. Notably, our method not only capably addresses these issues on the larger model but also enhances the overall fidelity and precision of the generated images.

5. Conclusion

In this work, we propose a simple yet effective method of using human-centric priors (HcP), *e.g.,* pose or depth images, to improve human image generation in existing text-to-image models. The proposed HcP layer effectively uses information about humans during the fine-tuning process without requiring additional input when generating images from text. Extensive experiments demonstrate that the HcP layer not only fixes structural inaccuracies in human structure generation but also preserves the original aesthetic qualities and details. Future work will explore the integration of multiple types of human-centric priors to further advance human image and video generation.

References

[1] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *CVPR*, pages 18208–18218, 2022. 2

[2] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. In *ICLR*, 2018. 6

[3] Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T Freeman, Michael Rubinstein, et al. Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704*, 2023. 2

[4] Weihua Chen, Xianzhe Xu, Jian Jia, Hao Luo, Yaohua Wang, Fan Wang, Rong Jin, and Xiuyu Sun. Beyond appearance: a semantic controllable self-supervised learning framework for human-centric visual tasks. In *CVPR*, pages 15050–15061, 2023. 2

[5] Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 2

[6] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *NeurIPS*, 34:8780–8794, 2021.

[7] Ming Ding, Wendi Zheng, Wenyi Hong, and Jie Tang. Cogview2: Faster and better text-to-image generation via hierarchical transformers. *NeurIPS*, 35:16890–16902, 2022. 2

[8] document Rec. ITU-R. Methodology for the subjective assessment of video quality in multimedia applications. *BT.1788*, pages 1–13, 2007. 7

[9] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *CVPR*, pages 12873–12883, 2021. 2

[10] Weixi Feng, Xuehai He, Tsu-Jui Fu, Varun Jampani, Arjun Reddy Akula, Pradyumna Narayana, Sugato Basu, Xin Eric Wang, and William Yang Wang. Training-free structured diffusion guidance for compositional text-to-image synthesis. In *ICLR*, 2022. 2

[11] Xiao Han, Licheng Yu, Xiatian Zhu, Li Zhang, Yi-Zhe Song, and Tao Xiang. Fashionvil: Fashion-focused vision-and-language representation learning. In *ECCV*, pages 634–651. Springer, 2022. 2

[12] Xiao Han, Xiatian Zhu, Licheng Yu, Li Zhang, Yi-Zhe Song, and Tao Xiang. Fame-vil: Multi-tasking vision-language model for heterogeneous fashion tasks. In *CVPR*, pages 2669–2680, 2023. 2

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 6

[14] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-or. Prompt-to-prompt image editing with cross-attention control. In *ICLR*, 2022. 2, 3

[15] 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. *NeurIPS*, 30, 2017. 6

[16] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 33:6840–6851, 2020. 2

[17] Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *ICLR*, 2021. 1, 6

[18] Xuan Ju, Ailing Zeng, Jianan Wang, Qiang Xu, and Lei Zhang. Human-art: A versatile human-centric dataset bridging natural and artificial scenes. In *CVPR*, pages 618–629, 2023. 6

[19] Xuan Ju, Ailing Zeng, Chenchen Zhao, Jianan Wang, Lei Zhang, and Qiang Xu. Humansd: A native skeleton-guided diffusion model for human image generation. *arXiv preprint arXiv:2304.04269*, 2023. 2

[20] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *CVPR*, pages 2426–2435, 2022. 2

[21] Diederik Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. *NeurIPS*, 34:21696–21707, 2021. 2

[22] Xian Liu, Jian Ren, Aliaksandr Siarohin, Ivan Skorokhodov, Yanyu Li, Dahua Lin, Xihui Liu, Ziwei Liu, and Sergey Tulyakov. Hyperhuman: Hyper-realistic human generation with latent structural diffusion. *arXiv preprint arXiv:2310.08579*, 2023. 2

[23] Zhiheng Liu, Ruili Feng, Kai Zhu, Yifei Zhang, Kecheng Zheng, Yu Liu, Deli Zhao, Jingren Zhou, and Yang Cao. Cones: Concept neurons in diffusion models for customized generation. *arXiv preprint arXiv:2303.05125*, 2023. 2

[24] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2018. 6

[25] Elman Mansimov, Emilio Parisotto, Jimmy Lei Ba, and Ruslan Salakhutdinov. Generating images from captions with attention. In *ICLR*, 2015. 2

[26] 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

[27] 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. 1, 2

[28] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 1, 2, 6

[29] Tingting Qiao, Jing Zhang, Duanqing Xu, and Dacheng Tao. Mirrorgan: Learning text-to-image generation by redescription. In *CVPR*, pages 1505–1514, 2019. 2

[30] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763. PMLR, 2021. 6

[31] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever.

Zero-shot text-to-image generation. In *ICML*, pages 8821–8831. PMLR, 2021. 1, 2

[32] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *ICML*, pages 1060–1069. PMLR, 2016. 2

[33] Yurui Ren, Xiaoming Yu, Junming Chen, Thomas H Li, and Ge Li. Deep image spatial transformation for person image generation. In *CVPR*, pages 7690–7699, 2020. 2

[34] Yurui Ren, Xiaoqing Fan, Ge Li, Shan Liu, and Thomas H Li. Neural texture extraction and distribution for controllable person image synthesis. In *CVPR*, pages 13535–13544, 2022. 2

[35] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 2, 3, 6

[36] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *NeurIPS*, 35:36479–36494, 2022. 1, 2

[37] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *NeurIPS*, 35:25278–25294, 2022. 6

[38] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 3, 6

[39] Ming Tao, Hao Tang, Fei Wu, Xiao-Yuan Jing, Bing-Kun Bao, and Changsheng Xu. Df-gan: A simple and effective baseline for text-to-image synthesis. In *CVPR*, pages 16515–16525, 2022. 1, 2

[40] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *CVPR*, pages 1316–1324, 2018. 2

[41] Xiaohan Xu, Ming Li, Chongyang Tao, Tao Shen, Reynold Cheng, Jinyang Li, Can Xu, Dacheng Tao, and Tianyi Zhou. A survey on knowledge distillation of large language models. *arXiv preprint arXiv:2402.13116*, 2024. 2

[42] 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

[43] Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. Diffusion models: A comprehensive survey of methods and applications. *ACM Computing Surveys*, 2022. 2

[44] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023. 2

[45] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *Transactions on Machine Learning Research*, 2022. 2

[46] Cheng Zhang, Xuanbai Chen, Siqi Chai, Henry Chen Wu, Dmitry Lagun, Thabo Beeler, and Fernando De la Torre. ITI-GEN: Inclusive text-to-image generation. In *ICCV*, 2023. 2

[47] Jinsong Zhang, Kun Li, Yu-Kun Lai, and Jingyu Yang. Pise: Person image synthesis and editing with decoupled gan. In *CVPR*, pages 7982–7990, 2021. 2

[48] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, pages 3836–3847, 2023. 1, 2, 6, 8

[49] Pengze Zhang, Lingxiao Yang, Jian-Huang Lai, and Xiaohua Xie. Exploring dual-task correlation for pose guided person image generation. In *CVPR*, pages 7713–7722, 2022. 2

[50] Shaofeng Zhang, Qiang Zhou, Zhibin Wang, Fan Wang, and Junchi Yan. Patch-level contrastive learning via positional query for visual pre-training. In *ICML*, 2023. 2

[51] Shaofeng Zhang, Feng Zhu, Rui Zhao, and Junchi Yan. Contextual image masking modeling via synergized contrasting without view augmentation for faster and better visual pre-training. In *ICLR*, 2023.

[52] Shaofeng Zhang, Feng Zhu, Rui Zhao, and Junchi Yan. Patch-level contrasting without patch correspondence for accurate and dense contrastive representation learning. *ICLR*, 2023. 2

[53] Xinyue Zhou, Mingyu Yin, Xinyuan Chen, Li Sun, Changxin Gao, and Qingli Li. Cross attention based style distribution for controllable person image synthesis. In *ECCV*, pages 161–178. Springer, 2022. 1

[54] Minfeng Zhu, Pingbo Pan, Wei Chen, and Yi Yang. Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis. In *CVPR*, pages 5802–5810, 2019. 2