Supplementary Materials: 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

Content

A. Ethical and Social Impacts

Our work in text-based HIG using the HcP layer, which relies on reference images sourced from publicly available datasets and base models publicly released on the HuggingFace diffusers library [7], presents various ethical and social considerations. The primary concern is the potential impact on privacy and data protection. The generation of human images based on text inputs could unintentionally produce likenesses of real individuals, highlighting the need for guidelines to protect individual privacy and meanwhile prevent misuse of personal likenesses. Additionally, the inherited biases in the training datasets and base models can lead to stereotypical images. It's important to continuously monitor and adjust these biases to ensure a fair and inclusive representation of generated human figures.

Furthermore, while our method can enhance representation in digital media by generating diverse and accurate human figures, there is a risk of misuse in creating misleading or harmful content. Establishing ethical guidelines and usage policies is crucial to prevent the creation of deepfakes. Collaborative efforts with various stakeholders are necessary to develop responsible use cases and address potential misuse. In summary, our approach in tHIG, while offering the potential for creative and inclusive image generation, must be balanced with a commitment to ethical practices, privacy protection, and the promotion of diversity and inclusivity in text-based human figure synthesis.

B. Cross-Attention Layer Details

The cross-attention layer within the U-Net architecture of latent diffusion models plays a pivotal role in synthesizing detailed and contextually relevant images. This layer operates by computing a set of queries ($\mathbf{Q}$), keys ($\mathbf{K}$), and values ($\mathbf{V}$) based on the input latent representation $\mathbf{z}_{in}$ and the given text-conditioned embeddings $\mathcal{C}$.

First, the input latent representation $\mathbf{z}_{in}$is transformed into a query matrix $\mathbf{Q}$ using a weight matrix $\mathbf{W}_q$. This process converts the input into the query space:

$$\mathbf{Q} = \mathbf{W}_q \, \mathbf{z}_{in} \in \mathbb{R}^d \,, \tag{1}$$

Simultaneously, text-conditioned embeddings $\mathcal{C}$ from CLIP [4] text encoder, which embeds textual information, is used to generate the key $\mathbf{K}$ and value $\mathbf{V}$ matrices through their respective weight matrices $\mathbf{W}_k$and $\mathbf{W}_v$:

$$\begin{aligned} \mathbf{K} &= \mathbf{W}_k \, \mathcal{C} \in \mathbb{R}^{d \times N} \,, \\ \mathbf{V} &= \mathbf{W}_v \, \mathcal{C} \in \mathbb{R}^{d \times N} \,, \end{aligned} \tag{2}$$

The attention mechanism then computes the attention map $\mathcal{M}$ by applying a softmax function to the dot product of $\mathbf{Q}$ and $\mathbf{K}^T$, scaled by the square root of the dimensionality d. This step effectively captures the relevance of each

[*] Corresponding author

element in the context of the input latent representation:

$$\mathcal{M} = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}}\right) , \tag{3}$$

Finally, the output latent representation $\mathbf{z}_{out}$ is obtained by multiplying the value matrix $\mathbf{V}$ with a combined attention map $\hat{\mathcal{M}}$ in Eq. 2, which is an enhanced version of $\mathcal{M}$ incorporating the novel Human-centric Prior (HcP) layer introduced in our approach:

$$\mathbf{z}_{out} = \mathbf{V} \times \hat{\mathcal{M}} \in \mathbb{R}^d , \tag{4}$$

This augmented cross-attention mechanism, through $\hat{\mathcal{M}}$, effectively integrates human-centric prior information into the diffusion process, leading to more accurate and detailed human image synthesis in the generated images.

C. Detailed Experiment Settings

Following the work of Ju *et al.* [3], we train the HcP layer on the ensemble of LAION-Human, and the training set of Human-Art, and test on the validation set of Human-Art. Note that we only apply the text-based prompt in the validation set for inference evaluation.

Human-Art [3] contains 50k high-quality images with over 123k person instances from 5 natural and 15 artificial scenarios, which are annotated with bounding boxes, key points, self-contact points, and text information for humans represented in both 2D and 3D. It is, therefore, comprehensive and versatile for various downstream tasks.

In our study, we specifically focus on the *real human* category of the Human-Art dataset, as it directly aligns with our goal of addressing human image generation challenges. This category encompasses five sub-categories: *acrobatics*, *dance*, *drama*, *cosplay*, and *movie*. These sub-categories offer a diverse range of human actions and poses, providing an ideal context for training the HcP layer to handle complex human structures. Examples are shown in Figure 1.

Laion-Human [3]. Ju *et al.* also constructed a dataset LAION-Human containing large-scale internet images. Specifically, they collected about 1M image-text pairs from LAION-5B [5] filtered by the rules of high image quality and high human estimation confidence scores. Superior to ControlNet, a versatile pose estimator is trained on the Human-Art dataset, 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.

Implementation Details. Detailed implementation settings of training & inference stages are listed in Table 1. All experiments are performed on 8× Nvidia Tesla A100 GPUs.

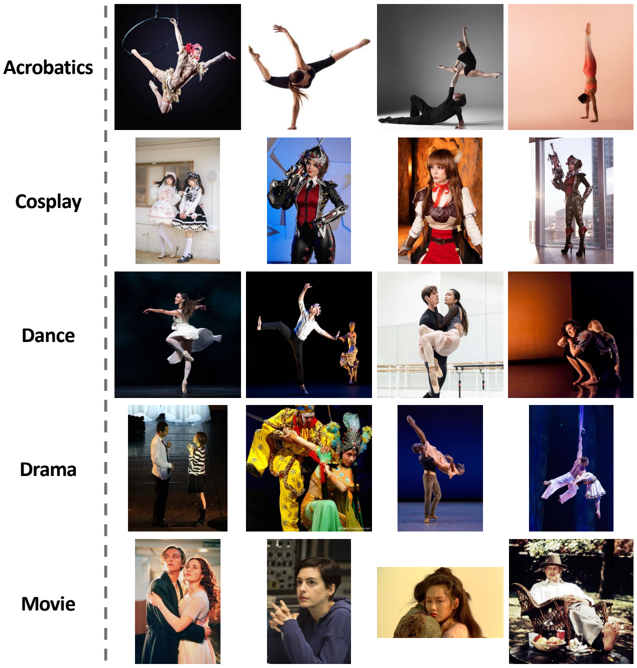

Figure 1. **Example images from the *realhuman* category in the Human-Art dataset.** Each sub-category (*acrobatics*, *dance*, *drama*, *cosplay*, and *movie*) is represented by four distinct images.

Table 1. List of implementation settings for both training and inference stages.

Implementation	Setting
Training Noise Schedular	DDPM [2]
Epoch	10
Batch size per GPU	8
Optimizer	Adam
Learning rate	0.0001
weight decay	0.01
Attention map ratio γ	0.1
loss ratio α	0.1
Training timestamp	1000
Training image size	512×512
Training pose image size	256×256
Sampling Noise Schedular	DDIM [6]
Inference step	50
guidance scale	7.5
Inference image size	512×512

D. Additional Ablations and Analyses

D.1. Attention Combination Weight γ

In this analysis, we focused on the ratio of attention map combination in Eq. 2 of the main text, as illustrated in Figure 2. This examination helps us understand the effects

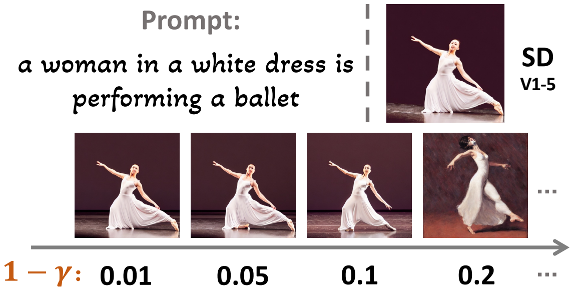

Figure 2. **Illustration of varying ratios (0.01, 0.05, 0.1, and 0.2) in the attention map combination on image generation.** Images generated for the given prompt at varying attention map combination ratios (0.01, 0.05, 0.1, 0.2).

of different ratios on the generated images. At lower ratios of 0.01 and 0.05, the images closely resemble those produced by the standard SD model, indicating that the Human-centric Prior (HcP) layer's adjustments are minimal and maintain the foundational characteristics of the SD outputs. The most effective correction occurs at a ratio of 0.1, where the HcP layer's influence is well balanced, significantly enhancing the accuracy of human figure generation while maintaining the original style and content of the SD model. However, a ratio of 0.2 leads to noticeable changes in both the content and style, diverging markedly from the SD model's outputs. Although this ratio corrects the human figures, it also significantly alters the overall image, affecting both the composition and thematic elements. In conclusion, these observations highlight the importance of an appropriate ratio to achieve a balance between correcting structural inaccuracies and preserving the original style and content of the images. The 0.1 ratio emerges as the most effective, offering an optimal blend of correction and preservation.

D.2. Learning Process with Pose Image

Figure 3 provides a visualization of the alignment between layer-specific ResNet features and corresponding scales of the human-centric attention map from the HcP layer. This visualization clearly demonstrates a notable similarity between the layer-specific ResNet features and the cross-attention maps at equivalent scales. This alignment plays a crucial role in our methodology. By ensuring that the ResNet features, which contain human-centric information, are closely aligned with the corresponding scales of the cross-attention layers, we enhance the model's ability to accurately incorporate human-centric details into the image generation process. This approach not only improves the structural integrity of the generated human figures but also ensures a more contextually accurate representation.

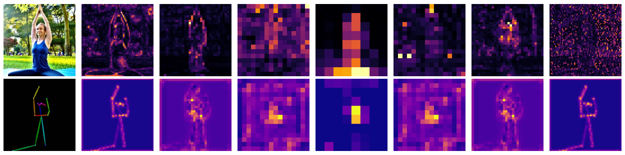

Figure 3. **Illustration of alignment between layer-specific ResNet features with corresponding scales and combined attention maps in each cross-attention layer.** The ResNet features are extracted from four different scale layers of a ImageNet pretrained ResNet50 model.

E. Additional Results

Due to space limitations, we only provide parts of the results in the main paper. In this section, we will report additional additional results, details, and analyses.

E.1. Human Evaluation Details

We provide the human evaluation setting details here. In the main text, we request participants to evaluate *anatomy quality* (AQ) and *text-image alignment* (TIA) for the prompt-generated image pairs. The former reflects viewers' experience in terms of the anatomical quality of the images. In contrast, the latter reflects viewers' subjective perceptions of text-image consistency between the generated images and corresponding text prompts. Before rating, we explain the whole procedure of our model and present some example pairs. When displaying these pairs, we explain the definition of AQ and TIA to each participant for a more precise understanding. Pairs that they require to rate are not included in these examples. Images are displayed in full-screen mode on calibrated 27-inch LED monitors (Dell P2717H). Viewing conditions are in accordance with the guidelines of international standard procedures for multimedia subjective testing [1]. The subjects are all university undergraduate or graduate students with at least two years of experience in image processing, and they claimed to browse images frequently. The percentage of female subjects is about 40%. All the subjects are aged from 20 to 27 years old. Before giving the final rating, we allow participants to watch each pair multiple times.

E.2. Controllable HIG Comparison

We provide additional comparisons with ControlNet [8] as shown in Figure 4. Note that in most cases, under the control of OpenPose images, ControlNet generates human images with the correct pose. The HcP layer does not interfere with generating human images with correct structures but acts to correct them in cases of errors. This makes the HcP layer an effective tool in preventing the generation of structurally incorrect human images.

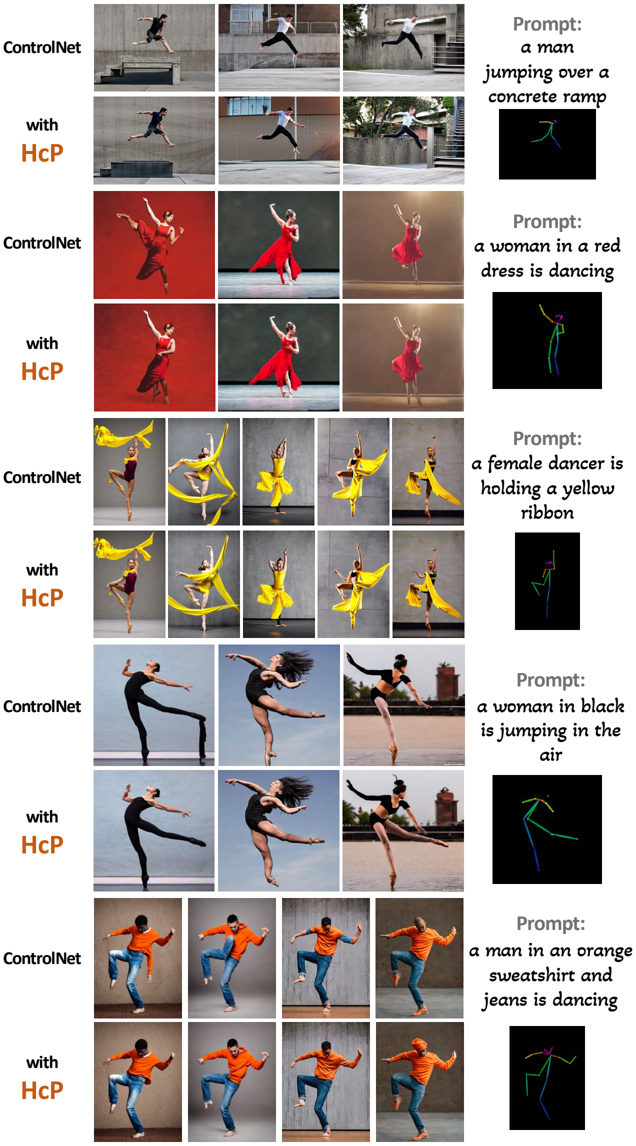
ControlNet
with HcP
Prompt: a man jumping over a concrete ramp
ControlNet
with HcP
Prompt: a woman in a red dress is dancing
ControlNet
with HcP
Prompt: a female dancer is holding a yellow ribbon
ControlNet
with HcP
Prompt: a woman in black is jumping in the air
ControlNet
with HcP
Prompt: a man in an orange sweatshirt and jeans is dancing

Figure 4. **Additional comparisons and compatibility with the controllable HIG application.** We selected ControlNet [8] as the basic model for controllable HIG and utilized OpenPose image as the conditional input.

E.3. Human-centric Prior Information Comparison

We provide additional results with different sources of human-centric prior information in Figure 5.

E.4. Qualitative Results

We provide additional qualitative results with baseline methods on three example prompts in Figure 7, and we also provide additional large diffusion model (SDXL) results on four example prompts in Figure 8.

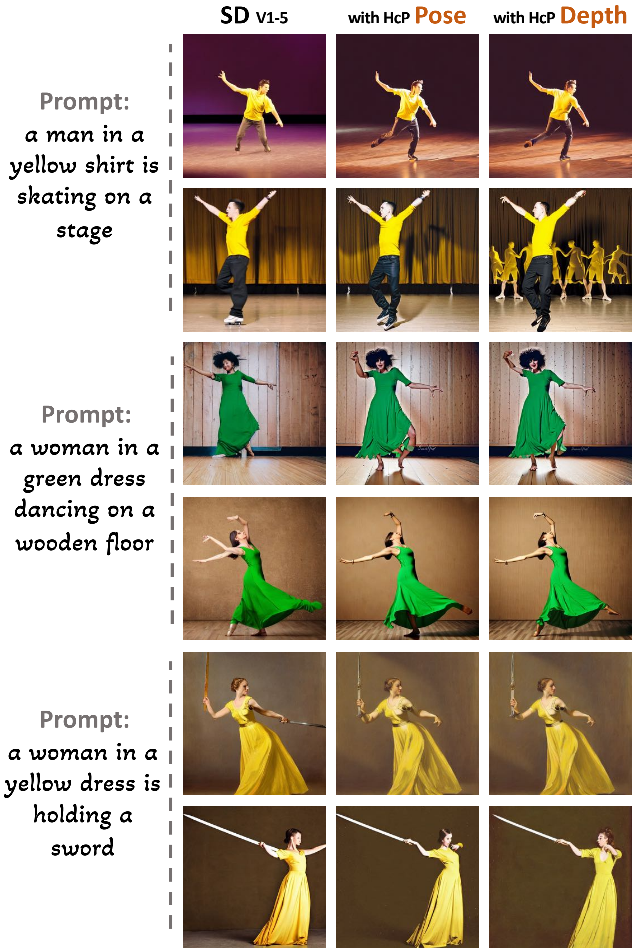
SD V1-5
with HcP Pose
with HcP Depth
Prompt: a man in a yellow shirt is skating on a stage
Prompt: a woman in a green dress dancing on a wooden floor
Prompt: a woman in a yellow dress is holding a sword

Figure 5. **Additional comparisons by using different sources of human-centric prior information**. The HcP layer is trained consistently for both Pose and Depth priors to ensure a fair and balanced comparison.

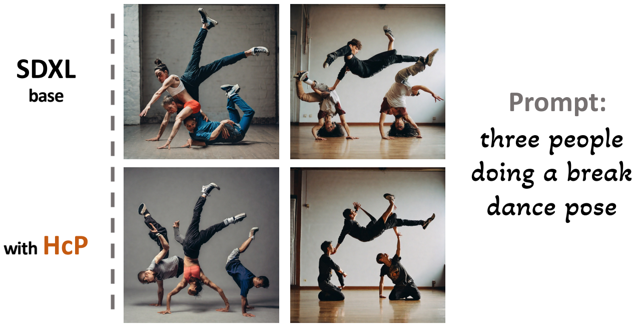

Figure 6. **Failure cases in complex action scenarios.** Two examples are generated using a pre-trained SDXL-base and an SDXL-base with the integrated HcP layer, in response to a more complex scene text prompt.

E.5. Failure Cases Analysis

We examine two instances where the generation of images based on the prompt "*three people doing a break dance pose*" fell short of expectations in Figure 6. The main reasons for these limitations are as follows. First, the generation of detailed facial features and limbs is less accurate. This inaccuracy may be due to the limitations of the SDXL-base model itself, particularly when depicting multiple individuals in a complex scene. Second, the intricacy of the '*break dance*' action, combined with the presence of multiple individuals, makes it more challenging to maintain accurate human structure in the generated images. Despite these challenges, it is noteworthy that the images generated with our HcP layer show improvements in human figure representation compared to those produced by the SDXL-base model. This highlights the HcP layer's effectiveness in enhancing image quality, even in complex scenarios involving detailed movements and multiple subjects.

F. Futuer work

Advancing from our present achievements, two crucial areas are highlighted for future development in text-based human image generation:

Diverse Data Learning: To improve the model's capability in handling complex scenarios, we plan to enrich our dataset with more varied and intricate human actions. This will enable continued learning and refinement, enabling better representation of dynamic human interactions.

Broader Priors Integration: We target to incorporate additional human-centric priors simultaneously, such as depth and edge, which will enhance the detail and realism of generated human figures, overcoming the limitations of relying solely on pose information.

References

[1] document Rec. ITU-R. Methodology for the subjective assessment of video quality in multimedia applications. *BT.1788*, pages 1–13, 2007. 3

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

[3] 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

[4] 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. 1

[5] 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. 2

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

[7] Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, and Thomas Wolf. Diffusers: State-of-the-art diffusion models, 2022. 1

[8] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, pages 3836–3847, 2023. 3, 4

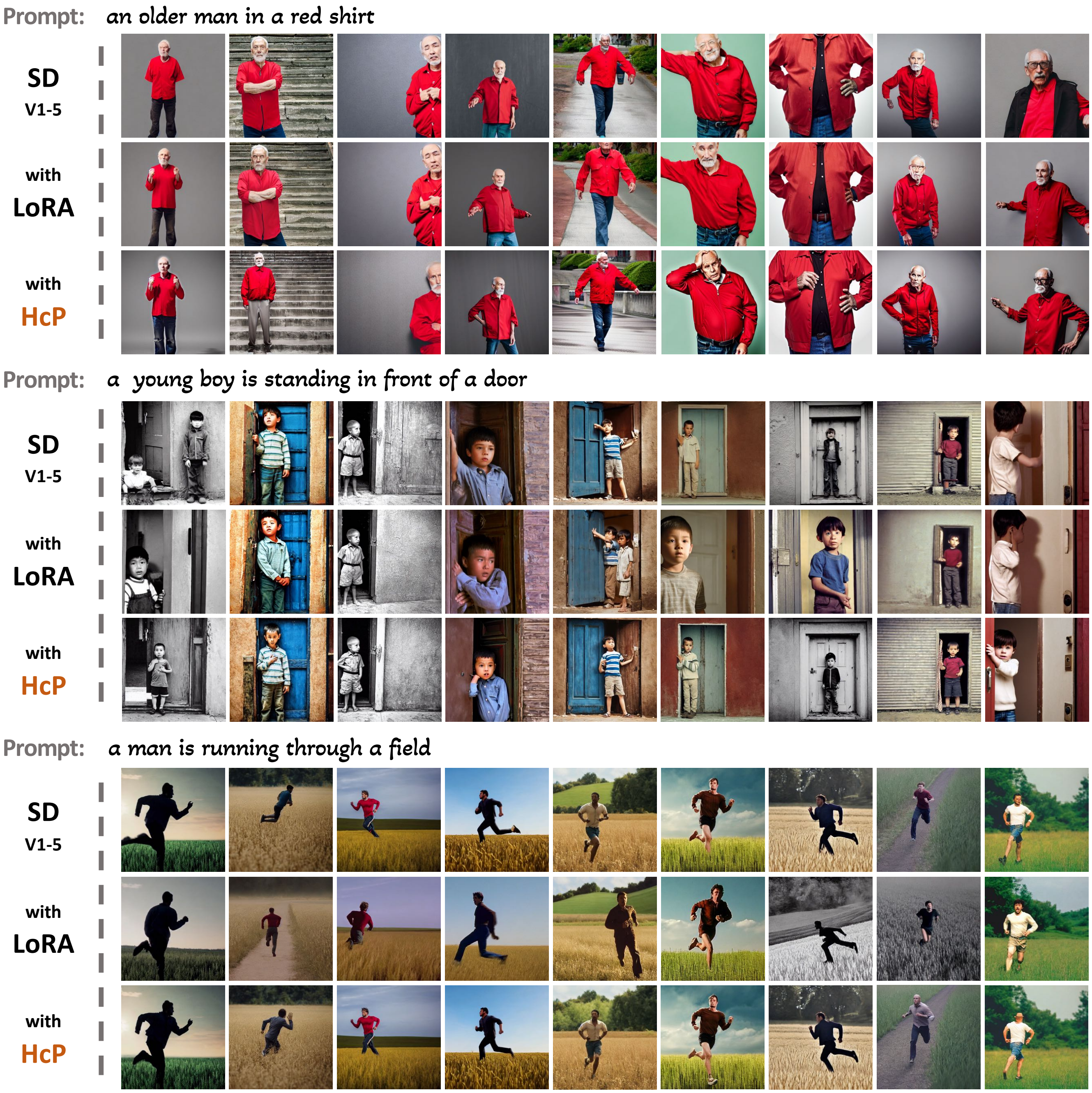
Prompt: an older man in a red shirt
SD V1-5
with LoRA
with HcP
Prompt: a young boy is standing in front of a door
SD V1-5
with LoRA
with HcP
Prompt: a man is running through a field
SD V1-5
with LoRA
with HcP

Figure 7. **Additional qualitative comparison with baseline methods on three example prompts**. We leverage the pre-trained SD v1-5 model for both "with LoRA" and "with HcP" models while keeping it frozen.

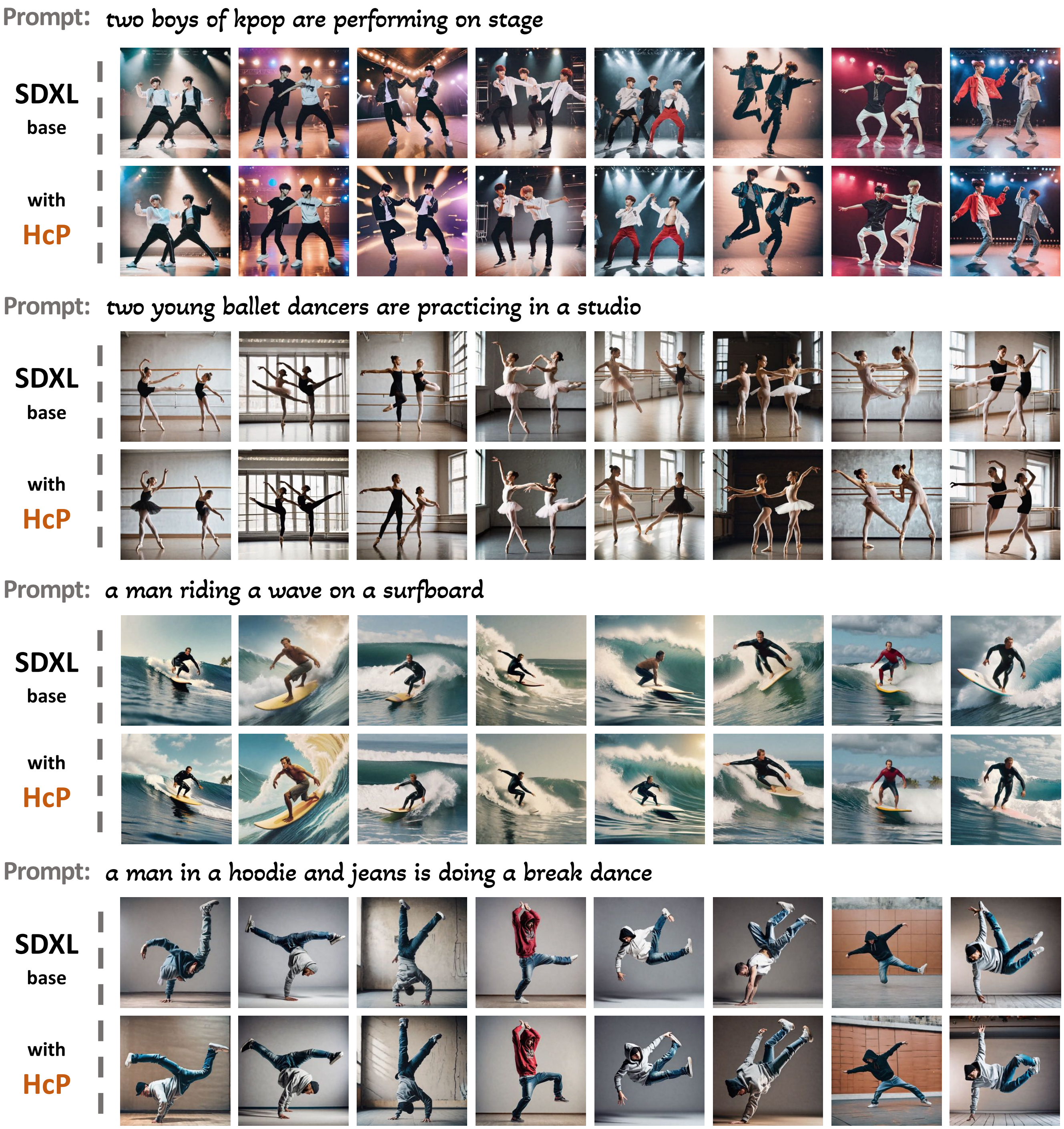
Prompt: two boys of kpop are performing on stage
SDXL base
with HcP
Prompt: two young ballet dancers are practicing in a studio
SDXL base
with HcP
Prompt: a man riding a wave on a surfboard
SDXL base
with HcP
Prompt: a man in a hoodie and jeans is doing a break dance
SDXL base
with HcP

Figure 8. **Additional results on larger diffusion model (SDXL-base) using HcP layer**. We leverage the pre-trained SDXL-base model for the "with HcP" model while keeping it frozen.