Controlling Human Shape and Pose in Text-to-Image Diffusion Models via Domain Adaptation Supplemental Material

Benito Buchheim Max Reimann Jürgen Döllner
Hasso Plattner Institute, University of Potsdam

A. Background

Latent Diffusion Models. Diffusion models add increments of random noise to input data during a variance-preserving Markov diffusion process [3, 12, 14], and then learn to reverse the diffusion process by denoising to construct desired data samples. Latent Diffusion Models (LDMs) [9] perform diffusion and denoising within a low-dimensional latent space $\boldsymbol{z}$, into which a Variational Autoencoder (VAE) [4] encodes an image $\boldsymbol{z}_0 = \mathcal{E}(I)$ using a pretrained encoder $\mathcal{E}(\cdot)$. Formally, the noisy latent representation of $\boldsymbol{z}_t$ is obtained by adding noise to z_o in every step t until $q(\boldsymbol{z}_T \mid \boldsymbol{z}_0)$ approximates a Gaussian distribution $\mathcal{N}(\mathbf{0}, \boldsymbol{I})$. In the denoising process, noise $\boldsymbol{\epsilon}_\theta(\boldsymbol{z}_t, t, \boldsymbol{c})$ is predicted for each timestep t from $\boldsymbol{z}_t$ to $\boldsymbol{z}_{t-1}$. Here, $\boldsymbol{\epsilon}_\theta$ represents a noise predicting neural network, typically a U-Net [10], while $\boldsymbol{c}$ denotes conditioning information. The training loss L minimizes the error between actual noise $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ and predicted noise $\boldsymbol{\epsilon}_\theta$:

$$L = \mathbb{E}_{\mathcal{E}(I), \boldsymbol{c}_\mathrm{p}, \boldsymbol{\epsilon}, t}\left[\omega(t) \|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\boldsymbol{z_t}, t, \boldsymbol{c}_\mathrm{p})\|_2^2\right], t = 1, ..., T \tag{1}$$

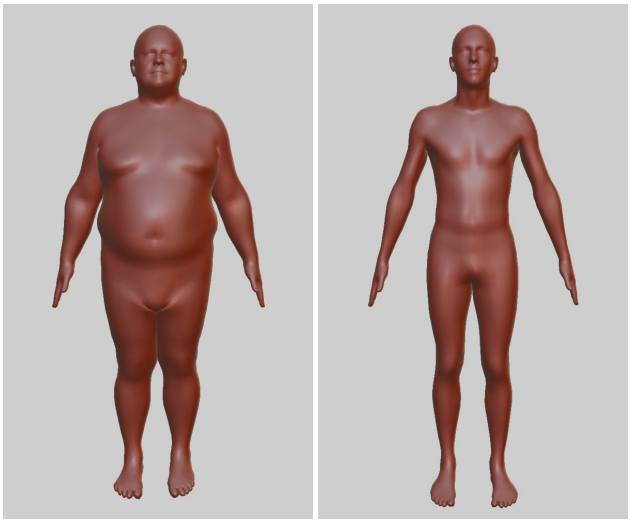
Here, $\omega(t)$ is a hyperparameter that adjusts the loss weighting at each timestep, and c_p are text prompt embeddings from CLIP in the case of text-to-image models such as Stable Diffusion [9]. After training, the model can progressively denoise from $\boldsymbol{z}_T \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ to $\boldsymbol{z}_0$ using a fast diffusion sampler [8, 13], with the final $\boldsymbol{z}_0$ decoded back into the image space I using a frozen decoder $\mathcal{D}(\cdot)$.

ControlNet. ControlNet [16] integrates spatially localized, task-specific image conditions into LDMs. It does this by cloning the blocks of a pretrained LDM, retraining these blocks, and then adding them to the original model output using zero convolution. These ControlNet blocks $\mathcal{C}$ modify the latent output features $\boldsymbol{f}$ of each U-Net decoder block:

$$\boldsymbol{f}_\mathrm{c} = \boldsymbol{f} + \mathcal{Z}\left(\mathcal{C}(\boldsymbol{x}, \boldsymbol{c}_\mathrm{f})\right) \tag{2}$$

Here, $\mathcal{Z}$ denotes zero-convolution and $\boldsymbol{c}_\mathrm{f}$ represents the task-specific spatial condition, which is incorporated in the fine-tuning of ControlNet blocks by minimizing the loss Eq. (1) with $\boldsymbol{\epsilon}_\theta(\boldsymbol{z}_t, t, \boldsymbol{c}_\mathrm{p}, \{\boldsymbol{f}_\mathrm{c}^i\})$, i.e., conditioned on all ControlNet block output features.

SMPL model. The Skinned Multi-Person Linear (SMPL) model [7], is widely used in computer graphics and vision for anatomically plausible and visually realistic human body deformations across various shapes and poses. SMPL combines a parametric shape space, capturing individual body shape variations, with a pose space encoding human joint positioning. The model uses low-dimensional parameters for pose, represented in joint angles $\theta_\mathrm{SMPL} \in \mathbb{R}^{24\times3\times3}$, and shape, represented in principal shape variation directions $\beta_\mathrm{SMPL} \in \mathbb{R}^{10}$, to produce a 3D mesh representation ($M \in \mathbb{R}^{3\times N}$) with $N = 6890$ vertices. A vertex weight evaluates the relationships between vertices and body joints.



(a) $\beta_2 = -4.0$ (b) $\beta_2 = 2.0$

Figure 1. Varying the second shape component β_2 of the SMPL [7] model. This component is highly correlated with body mass, which we use to evaluate the methods across different body types.

B. Evaluation Details and Additional Results

B.1. Shape and Pose Accuracy

The SMPL [7] model encodes body shape using shape components $\beta = \{\beta_1, \beta_2, ..., \beta_{10}\}$, derived through a principal component analysis of body meshes. For some of these components, identifiable and distinct effects on body shape can be determined. Specifically, the β_1 component strongly correlates with overall height, while β_2 strongly correlates with body mass (Fig. 1). The SURREAL [15] dataset provides body shape parameters collected from in-the-wild data. Although the data is of high quality and diversity, it contains limited samples for overweight and obese body types, with only about 4% of samples having a $\beta_2 \leq -2$. To fairly evaluate methods across body shapes, we augment the evaluation dataset (AS) to approximate a uniform shape distribution with respect to β_2. Specifically, for our extended analysis dataset (AS-Ext) we generate 500 pose-shape pairs for each increment of 1 of β_2 values between -6 and -2, using random poses from AIST [5] and shapes with corresponding β_2 sampled from SURREAL [15].

B.2. Background Stability

To measure background stability, we generated images using three prompts for 100 poses sampled from AIST, resulting in 300 prompt+pose samples. We use two prompts from Fig. 7 of the main paper and "a man dancing in the desert". For each sample, we generated images under both fit and not-fit shape conditions using our approach and ControlNet. We then used MaskRCNN [2] to mask out the person in each image. By computing the average LPIPS [17] between the masked versions of fit and not-fit pairs, we quantified the perceptual loss between the generated backgrounds when only the shape condition varied. ControlNet achieved an average LPIPS of 10.26, while our method improved on this with an average LPIPS of 9.13.

B.3. Animation

AnimateDiff [1] enables transition-consistent generation of frames, and can be controlled using a Control mechanism such as ControlNet. We use our approach to add 3d parametric SMPL control to the animations and generate several clips with both slim and obese variants. See Fig. 4 for frames from our animation. We provide several animated clips in our supplemental video. The animation appears smooth, with no notable degradation compared to using the original slim-body-type ControlNet. Faces appear less consistent, which is, however, a general limitation of the AnimateDiff method.

Figure 2. Ablation: including the prompt-condition into the SD U-Net of C_{SMPL}, i.e., $\epsilon_{\text{SD}}(\emptyset, \boldsymbol{c}_o) + w_1 \vec{\boldsymbol{g}}_{\boldsymbol{c_p}, \boldsymbol{c_o}} + w_2 \vec{\boldsymbol{g}}_{\boldsymbol{c_p}, \boldsymbol{c_s}, \boldsymbol{c_o}}$

B.4. Extended Guidance Ablation

Our proposed approach in Section 3.3 of the main paper always uses the empty prompt ($\emptyset$) for the attribute guidance SD U-Net's cross-attention layers during training and inference. That this is desirable can be seen in Fig. 2, in which we use an ablated configuration which also uses the prompt in the SMPL-guidance SD UNet, resulting in the guidance vector $\vec{\boldsymbol{g}}_{\boldsymbol{c_p}, \boldsymbol{c_s}, \boldsymbol{c_o}}$. As the two guidance vectors can no longer be scaled independently, increasing the guidance scales leads to an unstable background and exaggerated contrast. In Fig. 3 we also provide the prompt to the SD U-Net's cross-attention layers during inference, but do not use guidance composition, i.e., only using ϵ_{Syn}. We vary the combined prompt+SMPL guidance scale (w) for this configuration. While the body-shape is adhered to, the results exhibit a distinct synthetic appearance, which is exacerbated when increasing the shape adherence via w (in contrast to increasing w_2 of our proposed approach).

B.5. Guidance Scale Analysis

Our paper generally uses guidance scales (w_1, w_2) of 7.5 for all images, unless otherwise specified. These scales can

Figure 3. Ablated configuration without domain adaptation and no guidance composition. The prompt condition is supplied to the Stable Diffusion U-Net. We show results for different scales w in the ablated composition $\epsilon_{\text{Syn}}(\emptyset, \emptyset, \boldsymbol{c}_o) + w\vec{\boldsymbol{g}}_{\boldsymbol{c_p},\boldsymbol{c_s},\boldsymbol{c_o}}$.

be varied independently to adjust the strength of each component. In Fig. 5, we show an extended version of Fig. 9a from the main paper, by varying the guidance scale between 0 and 25. Moderate guidance scales below 25 yield high-fidelity images. Increasing the domain guidance scale (w_1) improves prompt adherence but introduces unstable backgrounds and artifacts in $w_1 = 25$. Similarly, increasing w_2 exaggerates the SMPL-shape, causing body artifacts at $w_2 = 25$. Higher w_2 values make images resemble the synthetic SURREAL dataset, while higher w_1 values decrease shape adherence. The diagonal ($w_1 = w_2$) provides the best combination of prompt adherence, shape adherence, and visual appearance.

Fig. 6 uses the same guidance scale setup with the prompt "a ballet dancer" and a dancing pose from AIST [5]. In this context, Stable Diffusion is biased towards the generation of slim body types, even with "obese" added to the prompt. As discussed in our limitations, our method similarly struggles to reconcile such conflicting information. Using an obese body shape for our SMPL-conditioned model only imparts the shape if w_2 is much larger than w_1, but this also results in a synthetic appearance. Intermediate scales also display artifacts in body composition. It should be noted that the severity of this limitation generally depends on factors such as prompt, seed, and pose. It can work well in some contexts, e.g., the football player in Fig. 7 of the main paper, likely due to specific training biases of Stable Diffusion.

B.6. Examples from Quantitative Evaluation

In Sections 4.2 and 4.3, we evaluated the fidelity and SMPL accuracy metrics of various model configurations and baseline models. We generated datasets using pose and shape inputs for each method, using a fixed seed. In Figs. 7 and 8, we provide visual examples from our experiments.

Fig. 7 presents outputs generated from the pose and caption inputs from the MSCOCO [6] dataset. As the models are only conditioned on high-level semantic information (text prompt, OpenPose pose map and SMPL parameters), the differences in visual content to the reference images (Fig. 7a) are expected and we are only interested in comparing visual fidelity. We calculated visual fidelity in terms of Kernel Inception Distance (KID) to measure the overall distance of the dataset to the reference real-world dataset [6]. As shown in the metrics, fine-tuning a ControlNet with additional cross-attention blocks and without domain adaptation (Fig. 7c) introduces a synthetic appearance. In contrast, our guidance composition avoids domain shift (Fig. 7d, e and g). Furthermore, it is evident that the domain guidance network (Fig. 7f and h) controls the overall layout of the image, while our method introduces only slight changes in the generated person. Fine-tuning only the attention blocks (ft-attn) keeps the background more stable compared to fine-tuning all blocks.

Fig. 8 shows outputs generated using the AIST [5]-SURREAL [15] SMPL models (see Section 4.3). It is visible that our models adhere better to the shape. In some cases, our ControlNet-guided (ft+CN) models also correct wrong body orientations introduced by ControlNet, while T2I-Adapter does not seem to suffer from this issue.

(a) Frame 1 (b) Frame 4 (c) Frame 8 (d) Frame 12 (e) Frame 16

Figure 4. Frames from a SMPL-controlled AnimateDiff [1] clip, for slim (top row) and overweight (bottom row) body types.

Figure 5. Varying guidance scales (w_1, w_2) in our proposed guidance composition: $\epsilon_{\text{SD}}(\emptyset, \boldsymbol{c}_o) + w_1 \vec{\boldsymbol{g}}_{\boldsymbol{c_p}, \boldsymbol{c_o}} + w_2 \vec{\boldsymbol{g}}_{\boldsymbol{c_s}, \boldsymbol{c_o}}$.

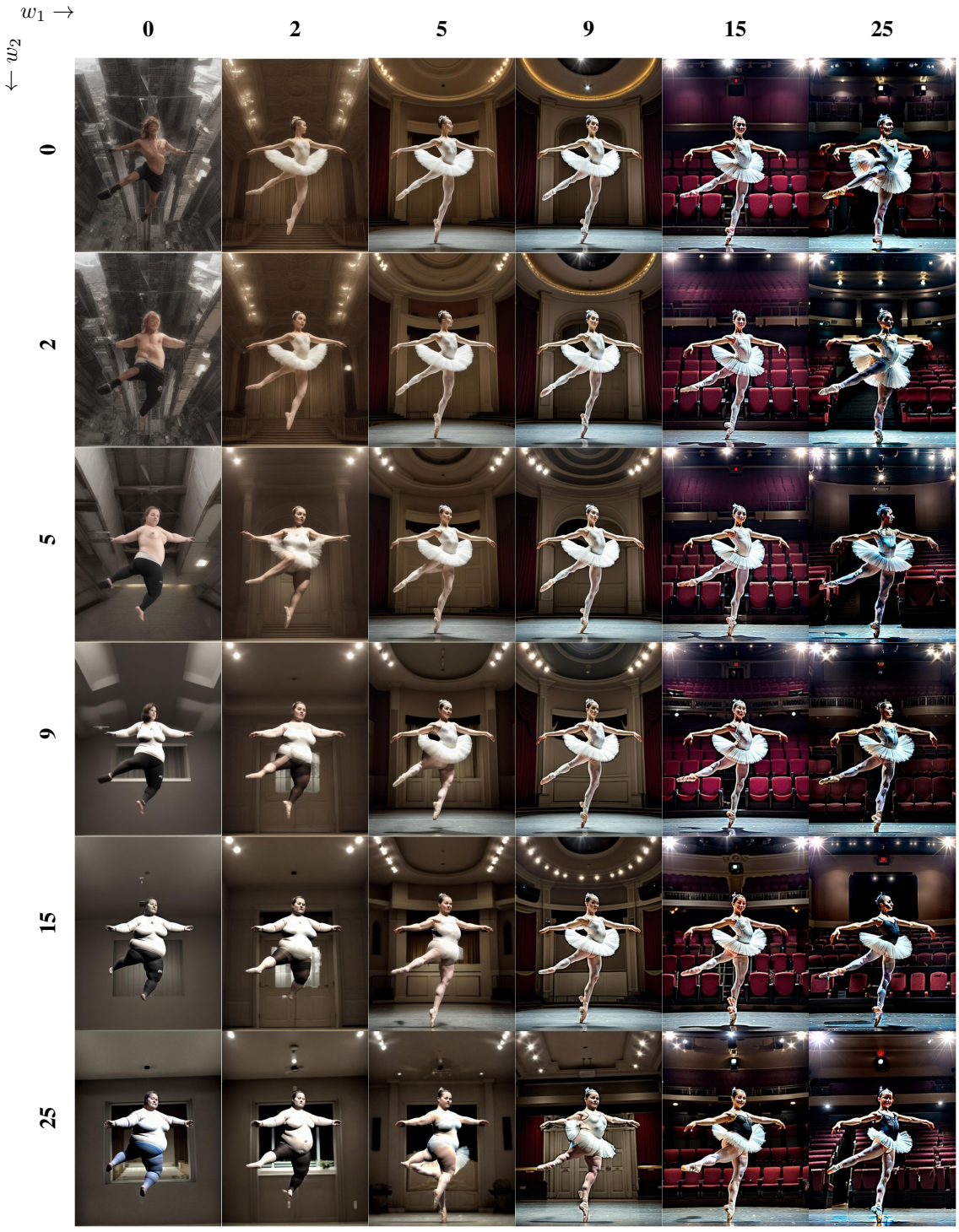
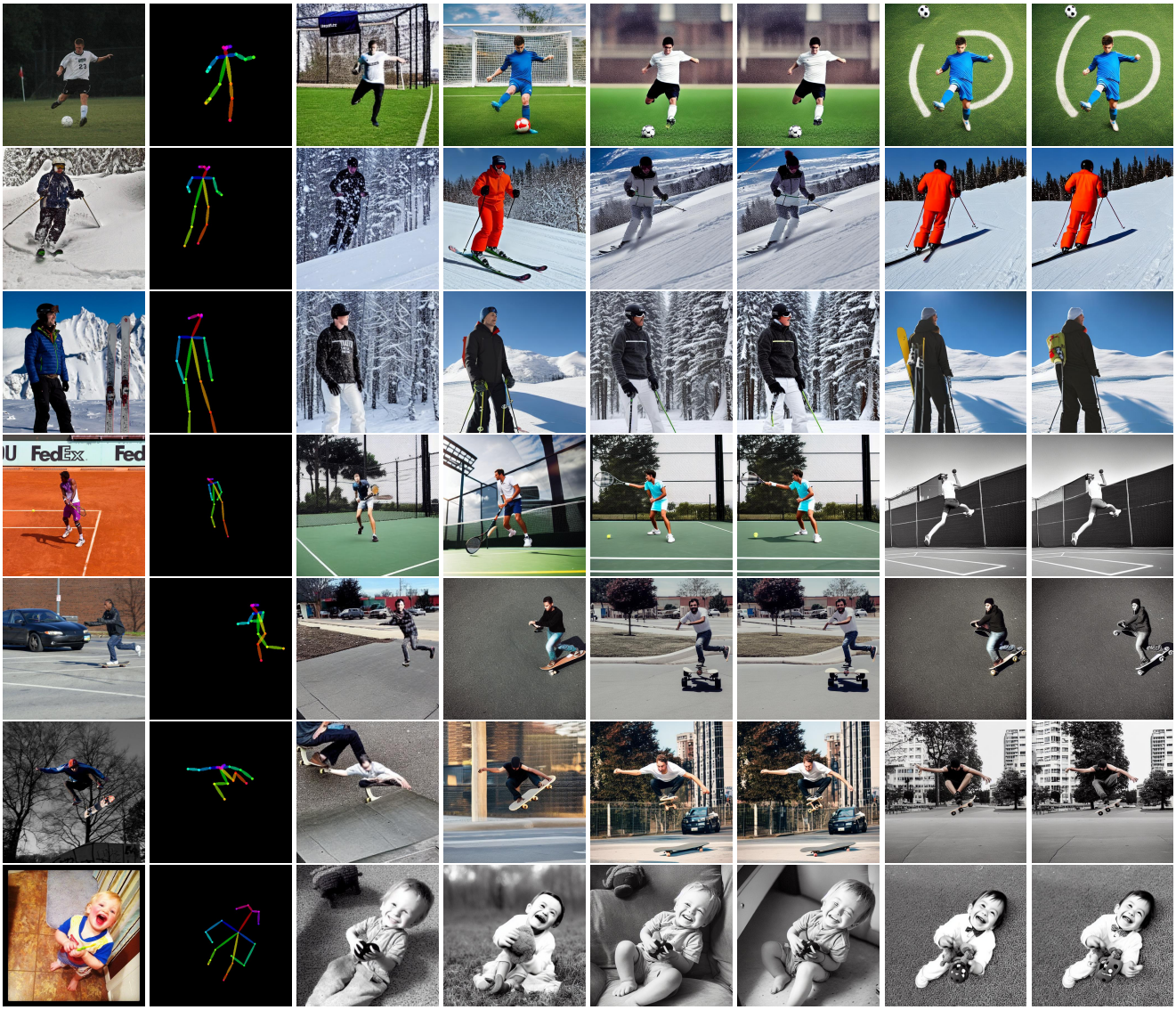
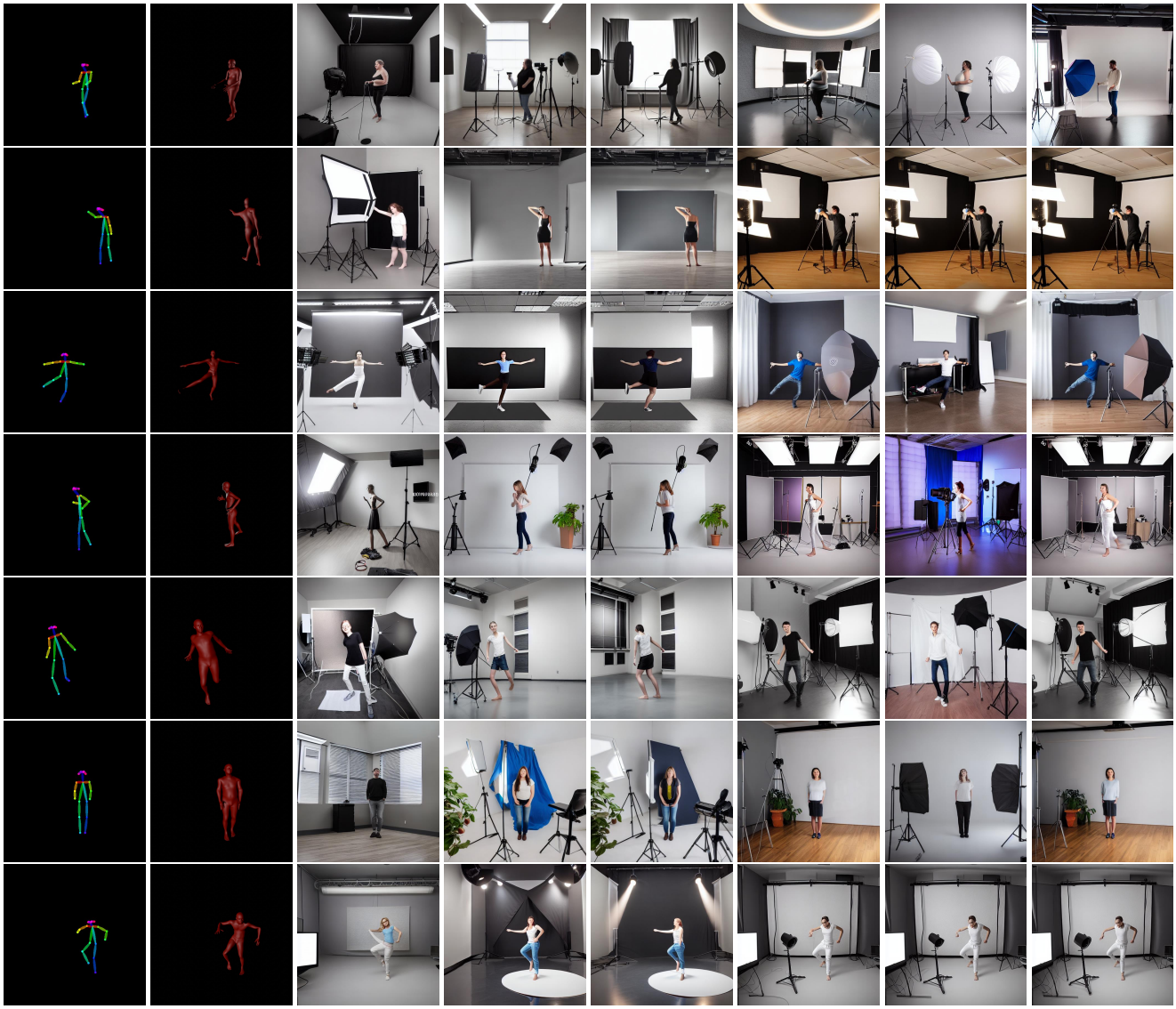


Figure 6. Limitation in Guidance Composition: The model struggles to reconcile conflicting information, such as obese SMPL-shape and slim ballet dancer produced by the pose-conditioned ControlNet.



(a) MSCOCO [6] (b) Pose (c) ft-extra-attn (d) ft-all+CN (e) ft-attn+CN (f) ControlNet (g) ft-attn+T2I (h) T2I-Adapter

Figure 7. Examples from the visual fidelity evaluation (Table 1, main paper). Our SMPL-finetuned ("ft-") models receive the text, pose and SMPL inputs, while the standard ControlNet and T2I-Adapter only receive the text and pose input. The text prompt and OpenPose Keypoints are provided by the MSCOCO [6] dataset, while SMPL annotations are predicted using HierProbHuman [11].



(a) Pose (b) SMPL (c) ft-all+CN (d) ft-attn+CN (e) ControlNet (f) ft-all+T2I (g) ft-attn+T2I (h) T2I-Adapter

Figure 8. Examples from the pose and shape accuracy evaluation (Table 2, main paper). Our SMPL-finetuned ("ft-") models receive the pose and SMPL inputs, while the standard ControlNet and T2I-Adapter only receive the pose input. The prompt for generating the entire dataset was "A person in a clean studio environment".

References

[1] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023. 2, 4

[2] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2961–2969, 2017. 2

[3] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 1

[4] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 1

[5] Ruilong Li, Shan Yang, David A Ross, and Angjoo Kanazawa. Ai choreographer: Music conditioned 3d dance generation with aist++. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (CVPR)*, pages 13401–13412, 2021. 2, 3

[6] 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 *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 3, 7

[7] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, Oct. 2015. 1, 2

[8] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 1

[9] 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)*, 2022. 1

[10] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention*. Springer, 2015. 1

[11] Akash Sengupta, Ignas Budvytis, and Roberto Cipolla. Hierarchical kinematic probability distributions for 3d human shape and pose estimation from images in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11219–11229, 2021. 7

[12] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning (ICML)*. PMLR, 2015. 1

[13] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations (ICLR)*, 2020. 1

[14] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations (ICLR)*, 2020. 1

[15] Gul Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 109–117, 2017. 2, 3

[16] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3836–3847, 2023. 1

[17] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 586–595, 2018. 2