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

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

Abstract

We present a methodology for conditional control of human shape and pose in pretrained text-to-image diffusion models using a 3D human parametric model (SMPL). Fine-tuning these diffusion models to adhere to new conditions requires large datasets and high-quality annotations, which can be more cost-effectively acquired through synthetic data generation rather than real-world data. However, the domain gap and low scene diversity of synthetic data can compromise the pretrained model's visual fidelity. We propose a domain-adaptation technique that maintains image quality by isolating synthetically trained conditional information in the classifier-free guidance vector and composing it with another control network to adapt the generated images to the input domain. To achieve SMPL control, we fine-tune a ControlNet-based architecture on the synthetic SURREAL dataset of rendered humans and apply our domain adaptation at generation time. Experiments demonstrate that our model achieves greater shape and pose diversity than the 2d pose-based ControlNet, while maintaining the visual fidelity and improving stability, proving its usefulness for downstream tasks such as human animation. Our code is available at: https://ivpg.github.io/humanLDM

(a) Input SMPL control mesh and our SD [26] outputs

(b) Synthetic training data [37]

Figure 1. Our approach allows 3d parametric control over human pose and shape (a) in LDMs using SMPL [18] meshes. We train on synthetic data (b) and propose a domain adaptation technique to adapt model outputs into the original visual domain.

1. Introduction

The advent of advanced generative diffusion models, particularly Latent Diffusion Models (LDMs) [26], has revolutionized the field of image generation. Models for text-to-image synthesis such as Stable Diffusion [26], DALLE-3 [2] or Gemini [35] have made high-quality image synthesis from complex prompts accessible to a broad audience of users worldwide. Text alone, however, is not sufficient in many cases to describe exact scene layouts, subjects, or style. Recent techniques for added spatial control have enhanced the capabilities of these models by incorporating various forms of spatial guidance. ControlNet and T2I-Adapter [22, 39], for instance, augment LDMs with localized, task-specific conditions like edges and human poses, allowing for precise control over the generated images.

Our work focuses on precise control over humans in generated scenes (Fig. 1). While 2D pose conditions [22, 39] allow specifying constraints on human poses, body shapes are not controllable, and specified poses might suffer from 3D ambiguities, complicating precise human-centric illustration and animation. To address this, we aim to control body pose and shape using the commonly used 3D human parametric model, SMPL [18]. Our approach modifies the ControlNet [39] architecture to use SMPL parameter embeddings in the cross-attention blocks that typically attend to text condition embeddings. This modification enables the model to control global image content based on SMPL

guidance, seamlessly integrating human geometry information into the generation process (Fig. 1a).

However, training such control networks for large LDMs like Stable Diffusion (SD) [26] typically requires an annotated training dataset with a large and diverse set of high-quality images to retain the output fidelity of the original network. Most existing real-world datasets for 3D annotations, including SMPL parameters, do not meet these criteria, particularly regarding scale and diversity in scenes and body shapes. Synthetic datasets such as SURREAL [37] provide images with diverse body shapes and scenes at scale, and offer further advantages such as high-quality annotations and lower regulatory barriers. However, the visual fidelity of synthetic datasets is typically degraded. For instance, SURREAL [37] suffers from non-photorealistic shading (Fig. 1b), inadequate blending between subjects and background, as well as low image resolution, leading to a visual domain shift in SD networks fine-tuned on it. Our goal is to extract the human shape and pose information while discarding the visual aesthetics from such datasets. While traditional domain adaptation techniques [1, 13] aim to translate synthetic data to real-world appearance, adapting synthetic datasets to the full visual range of outputs of large text-to-image LDMs representing diverse and complex scenes like SD is challenging.

We introduce a technique for domain adaptation in the latent space of LDMs to extract the control condition of a synthentically-trained LDM control network and apply it in the visual domain of the original model using classifier-free guidance. For human body shape and pose control, we obtain a SMPL-conditioned vector from our synthetically trained ControlNet, which is then composed with outputs of another control network for visual domain guidance, effectively adapting the isolated SMPL-condition to the visual appearance and fidelity of the original SD model. Our results demonstrate significantly improved adherence to shape and pose compared to current control approaches [22, 39], while simultaneously retaining the visual fidelity (in terms of KID and Inception Score) of SD.

In summary, our contributions are as follows:

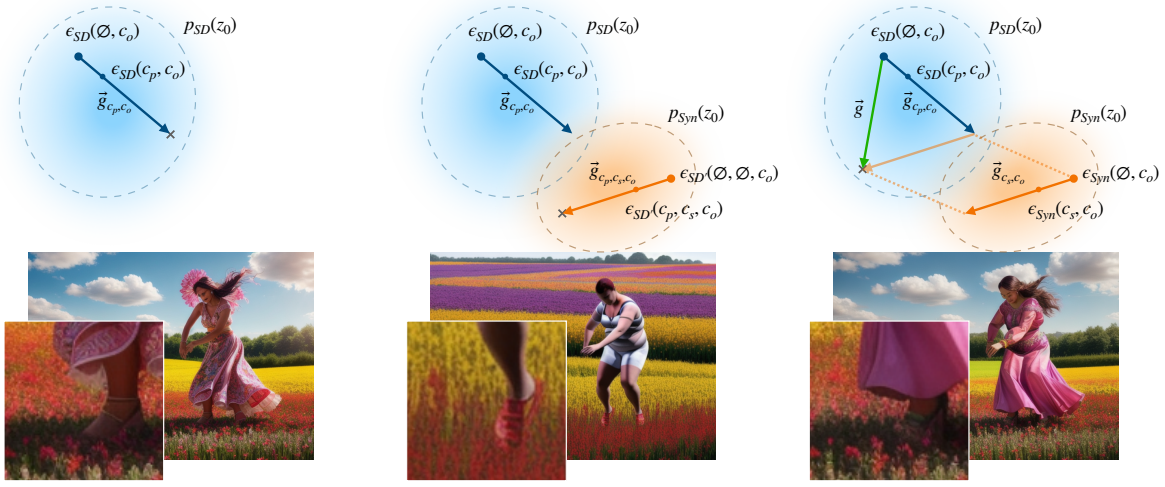
- We introduce a classifier-free guidance-based technique for domain adaptation in LDMs to adapt the visual domain of control models trained on synthetic data to the original high fidelity domain .
- We propose a SMPL-based control model for shape and pose control trained on the synthetic SURREAL [37] dataset.
- We demonstrate the efficacy of our approach in terms of visual fidelity, SMPL accuracy, and comparisons against state-of-the-art methods and ablated configurations.

2. Related Work

Image Generator Domain Adaptation. Domain adaptation [13] aims to shift the data distribution of a model to a new domain different from the one it was trained in. Various domain adaptation methods for image generation models have been explored, both to adapt model weights [6, 34], as well as outputs during inference [21, 24, 40]. The latent space of GANs such as StyleGAN [12] has been shown to be highly adaptable to new conditional domains such as CLIP [25]-embeddings [6, 24] or LDM-based classifier-free guidance [32]. Adding conditional control to LDMs without retraining on labeled data can be seen as a form of test-time domain adaptation (TTA) [15], adapting the model feature space [36] or the latent noise vector [21, 40] to the new conditional domain during generation. However, these methods do not address visual domain shifts. Our proposed generator domain adapation technique, similar to traditional TTA [15], transfers the visual domain from a synthetically trained source to a desired visual target domain at generation time. Instead of adapting training data [1] or model weights [34], we isolate the conditioning signal from its visual appearance using classifier-free guidance. Following Liu et al. [17], who proposed the composition of multiple conditional domains, we use guidance composition to combine the isolated SMPL guidance vector with a visual domain guidance vector from a control network trained in the original SD domain.

Controlling Text-to-Image Models. Several methods manipulate and control text-to-image diffusion models. ControlNet [39] and T2I-Adapter [22] enhance pretrained LDMs with spatially localized, task-specific conditions such as edges, depth, and human poses by integrating retrained model blocks or trainable adapters [22, 39]. DreamBooth [23] fine-tunes models for subject-specific generation and style control using diffusion guidance. InstructPix2Pix [4] enables text-based semantic editing and stylization of input images. Training-free methods, manipulate guidance vectors [20,33], null-text embeddings [21,40], or features [36]. DreamWalk [20] decomposes prompts into components and, similar to ours, uses compositional guidance [17] to emphasize style or spatial concepts during classifier-free guidance. In contrast to the preceding works that either assume availability of in-domain training data at scale or rely on less precise prompt- or feature-injection at runtime, our approach addresses adding fine-grained spatial control to text-to-image LDMs in the presence of only synthetic data.

SMPL-based Control. Only few methods have so far explored 3D body control in LDMs, mostly in the context of reference image control. Champ [30] uses SMPL conditioning, and MagicAnimate [38] uses dense-pose [7] conditioning of LDMs to enhance shape alignment and motion guidance in human image animation. By extracting pose and



(a) ControlNet (b) Naive Fine-Tuning (c) Guidance Composition

Figure 2. A pretrained ControlNet conditioned on 2d poses c_o with outputs ϵ_{SD} (a) generates text- and pose-guided images along a guidance vector $\vec{\boldsymbol{g}}_{\boldsymbol{c_p},\boldsymbol{c_o}}$ in the data domain p_{SD} (blue) of the SD model. To condition on SMPL parameters c_s for additional human shape control, the model is fine-tuned on a synthetic dataset (b), shifting the model outputs $\epsilon_{SD'}$ into the synthetic data domain p_{Syn} (orange). Our approach proposes guidance vector composition (c) to adapt the visual output domain to p_{SD} while retaining shape and pose control.

shape information from a reference video and conditioning an LDM on the rendered semantic maps, they can generate consistent and controllable human animations that replicate the actions observed in the reference video. Compared to these approaches, our method focuses on reference-free text-to-image generation with SMPL guidance while retaining the full representation capability of the original LDM.

3. Method

3.1. Problem Setting

Traditional fine-tuning of a generative model on conditional data typically shifts the output distribution to match the conditional training data. Mathematically, an unconditional generative model with parameters θ learns the probability distribution $P_\theta(\boldsymbol{x})$ of a training dataset $\mathcal{D} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, ...\}$. Fine-tuning with a second *annotated* dataset $\mathcal{D}' = \{(\boldsymbol{y}_1, \boldsymbol{c}_1), (\boldsymbol{y}_2, \boldsymbol{c}_2), ...\}$, optimizes θ to approximate the conditional distribution $P_{\theta'}(\boldsymbol{y} \mid \boldsymbol{c})$. This step shifts the marginal distribution of the model to the domain of the dataset $\mathcal{D}'$, effectively approximating $P_{\theta'}(\boldsymbol{y})$. This process retains visual output fidelity if the distributions $P(\boldsymbol{x})$ and $P(\boldsymbol{y})$ are similar. For instance, ControlNet [39] matches the visual characteristics of Stable Diffusion by training on similar data with added control annotations (see Fig. 2a). Here, the pretrained 2d-pose conditioned ControlNet remains within the original training domain P_{SD}. In contrast, fine-tuning on synthetic data, often with lower visual quality, can lead to a domain shift in $P_{\theta'}(\boldsymbol{y})$ and degrade output fidelity. As shown in Fig. 2b, a ControlNet fine-tuned on a 3d pose and shape dataset adheres to the target conditioning, but also shifts to the aesthetics of the synthetic data P_{Syn}. Faced with the challenge of fine-tuning LDMs with out-of-domain synthetic data to enable control over conditional information, we instead propose a domain adaptation technique that makes use of inherent properties of the LDM image generation process. The key idea is to isolate the conditional information from its domain and apply it within the original model's domain using classifier-free guidance to achieve $P_{\theta'} \sim P(\boldsymbol{x} \mid \boldsymbol{c})$. As depicted in Fig. 2c, our guidance composition approach applies shape and pose conditioning in the target domain, maintaining high visual fidelity.

3.2. Guidance Domain Adaptation

Next, we detail our general approach to guidance-based domain adaptation. We utilize two separate SD-control networks, with one responsible for visual domain guidance, i.e. targeting the desired appearance distribution P_{SD}, and the other for attribute guidance, i.e., targeting the conditional attribute distribution $P_{\text{Syn}}(\boldsymbol{y}|\boldsymbol{c})$. By composing these with classifier-free guidance, we isolate the attribute and shift its visual domain to the desired appearance.

Domain Guidance and Pose Guidance Networks. Fig. 3 illustrates the network components of our approach. We employ a ControlNet C_{SD} which serves as the domain appearance guidance net. In LDMs, latents z_{t-1} can be obtained from diffusion sampling [19, 31] using noise estimates ϵ_θ [10] dependent on the previous latent z_t at timestep t. See supplemental material Sec. A for

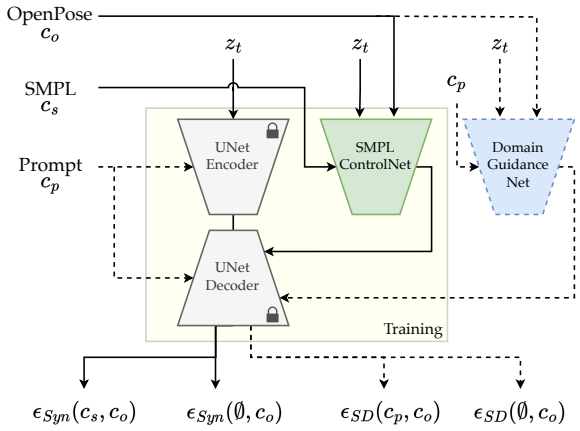


Figure 3. Overview of the networks involved in our SD-control approach during one denoising timestep t. During fine-tuning (solid lines), the overall model output $\epsilon_{\text{Syn}}(c_s, c_o) = \epsilon_\theta(z_t, t, \emptyset, C_{\text{SMPL}}(\boldsymbol{c}_\text{s}, \boldsymbol{c}_\text{o}))$ is adapted to a synthetic SMPL image dataset. During inference (solid and dotted lines), a pose-conditioned guidance network (C_{SD}) is executed alongside C_{SMPL} and a composite guidance vector is constructed from the outputs.

background on LDMs, ControlNet, and notation. Noise estimates steered with C_{SD} are defined as $\epsilon_{\text{SD}}(\boldsymbol{c}_p, \boldsymbol{c}_o) = \epsilon_\theta(z_t, t, \boldsymbol{c}_p, C_{\text{SD}}(\boldsymbol{c}_\text{p}, \boldsymbol{c}_\text{o}))$, where additional inputs are the prompt $\boldsymbol{c}_p$, and the spatial control condition $\boldsymbol{c}_o$. The appearance guidance net was trained on images within the original SD data domain, ensuring $\epsilon_{\text{SD}}(\boldsymbol{c}_p, \boldsymbol{c}_o) \sim P_{\text{SD}}$. The outputs in the *synthetic* data domain are defined as $\epsilon_{\text{Syn}}(\boldsymbol{c}_s, \boldsymbol{c}_o) = \epsilon_\theta(z_t, t, \emptyset, C_{\text{Syn}}(\boldsymbol{c}_\text{s}, \boldsymbol{c}_\text{o}))$, where C_{Syn} is an attribute guidance network finetuned in the synthetic data domain, thus $\epsilon_{\text{Syn}}(\boldsymbol{c}_s, \boldsymbol{c}_o) \sim P_{\text{Syn}}$, conditioned on an attribute $\boldsymbol{c}_s$ we want to isolate.

Classifier-Free Guidance. Classifier-free guidance (CfG) [11] steers the generative process of diffusion models by amplifying the conditional's gradient along a vector $\vec{\boldsymbol{g}}_{\boldsymbol{c}}$. During training, the model is simultaneously trained on both conditional ($\boldsymbol{c}$) and unconditional ($\emptyset$ or zeroed $\boldsymbol{c}$) objectives. These are combined during inference as:

$$\vec{\boldsymbol{g}}_{\boldsymbol{c}} = \boldsymbol{\epsilon}_\theta(\boldsymbol{z}_t, t, \boldsymbol{c}) - \boldsymbol{\epsilon}_\theta(\boldsymbol{z}_t, t, \emptyset),$$
$$\boldsymbol{z}_{t-1} \leftarrow \boldsymbol{\epsilon}_\theta(\boldsymbol{z}_t, t, \emptyset) + w\vec{\boldsymbol{g}}_{\boldsymbol{c}},$$

where w controls the guidance strength (CfG scale), and the latent z_{t-1} is the output of one step of diffusion sampling (e.g., DDIM [31]). This guidance vector helps in amplifying the desired characteristics specified by $\boldsymbol{c}$ while retaining the overall quality of the generated images. Applying CfG with outputs from the appearance guidance ControlNet C_{SD} thus results in a guidance vector $\vec{\boldsymbol{g}}_{\boldsymbol{c_p},\boldsymbol{c_o}}$ conditioned on a spatial condition ($\boldsymbol{c}_o$) and pointing towards the prompt ($\boldsymbol{c}_p$):

$$\vec{\boldsymbol{g}}_{\boldsymbol{c_p},\boldsymbol{c_o}} = \epsilon_{\text{SD}}(\boldsymbol{c}_\text{p}, \boldsymbol{c}_\text{o}) - \epsilon_{\text{SD}}(\emptyset, \boldsymbol{c}_\text{o}), \tag{1}$$

which we denote as the appearance guidance vector.

Guidance Composition. Now consider the latents in the synthetic domain $\epsilon_{\text{Syn}}(\boldsymbol{c}_s, \boldsymbol{c}_o)$. To shift their visual aesthetics into the original data domain (P_{SD}), we isolate the condition $\boldsymbol{c}_s$ using classifier-free guidance and compose it with the appearance guidance vector $\vec{\boldsymbol{g}}_{\boldsymbol{c_p},\boldsymbol{c_o}}$:

$$\vec{\boldsymbol{g}}_{\boldsymbol{c_s},\boldsymbol{c_o}} = \epsilon_{\text{Syn}}(\boldsymbol{c}_\text{s}, \boldsymbol{c}_\text{o}) - \epsilon_{\text{Syn}}(\emptyset, \boldsymbol{c}_\text{o}) \tag{2}$$
$$\vec{\boldsymbol{g}} = w_1\vec{\boldsymbol{g}}_{\boldsymbol{c_p},\boldsymbol{c_o}} + w_2\vec{\boldsymbol{g}}_{\boldsymbol{c_s},\boldsymbol{c_o}} \tag{3}$$
$$\boldsymbol{z}_{t-1} \leftarrow \epsilon_{\text{SD}}(\emptyset, \boldsymbol{c}_o) + \vec{\boldsymbol{g}} \tag{4}$$

where w_1 and w_2 are weighting factors for the guidance vectors. Effectively, the guidance vector $\vec{\boldsymbol{g}}_{\boldsymbol{c_p},\boldsymbol{c_o}}$ thus points in the direction of the prompt and visual domain, while the guidance vector $\vec{\boldsymbol{g}}_{\boldsymbol{c_s},\boldsymbol{c_o}}$ points towards adherence to condition $\boldsymbol{c}_s$. Since neither conditional nor unconditional ϵ_{Syn} receives prompt information, increasing the magnitude of $\vec{\boldsymbol{g}}_{\boldsymbol{c_s},\boldsymbol{c_o}}$ through w_2 does not impart more visual appearance information, thereby isolating the $\boldsymbol{c}_s$ from its visual appearance. Linearly combining both vectors shifts $\vec{\boldsymbol{g}}_{\boldsymbol{c_s},\boldsymbol{c_o}}$ into the original data distribution. We ablate the contributions of the guidance vector components in Sec. 4.6 and provide additional ablations in Sec. B.4 and a guidance scale analysis in Sec. B.5 of the supplemental material.

3.3. SMPL-Control Guidance

We apply our proposed guidance-based domain-adaptation approach to condition Stable Diffusion on SMPL models. As illustrated in Fig. 3, the shared spatial condition $\boldsymbol{c}_o$ to the domain guidance network C_{SD} and the attribute guidance network C_{Syn} is a 2D map of OpenPose [5] pose joints. The attribute guidance network $C_{\text{SMPL}} = C_{\text{Syn}}$ is a ControlNet additionally conditioned on SMPL embeddings ($\boldsymbol{c}_s$), see the supplemental material (Sec. A) for a background on the SMPL model. The visual domain guidance network C_{SD} is a pretrained ControlNet or a T2I-Adapter [22] and receives a prompt ($\boldsymbol{c}_p$) in addition to $\boldsymbol{c}_o$. We apply guidance composition on the output latents as detailed in Eq. (4). Overall, the composite guidance vector $\vec{\boldsymbol{g}}$ encodes prompt, SMPL-based shape and pose control while retaining visual appearance (see Fig. 2c).

Training. To obtain C_{SMPL}, we condition a ControlNet architecture on SMPL parameters by fine-tuning the cross-attention blocks, which originally attended to prompt embeddings ($\boldsymbol{c}_p$), using embeddings of SMPL parameters ($\boldsymbol{c}_s$) instead. The shape (β_{SMPL}) and pose (θ_{SMPL}) parameters of the SMPL model are concatenated and embedded by a single linear layer, i.e. $\boldsymbol{c}_s = \text{emb}(\theta_{\text{SMPL}}, \beta_{\text{SMPL}})$, to match the expected input dimensions of the cross-attention blocks. Further, also the SD-UNet only receives empty prompts during training, effectively removing prompt conditioning from C_{SMPL}. During training, the network is initialized with the weights of a 2d-pose ($\boldsymbol{c}_o$) conditioned ControlNet. After minimizing the training objective, outputs $\epsilon_{\text{Syn}}(\boldsymbol{c}_s, \boldsymbol{c}_o)$

are adapted to the synthetic training dataset domain. The output images closely resemble the degraded aesthetics of its synthetic training dataset (e.g., SURREAL [37]) manifesting in non-realistic skin, faces and clothing and free floating feet (see Fig. 2b). To enable classifier-free guidance, C_{SMPL} is trained on conditional and unconditional inputs, i.e., $\epsilon_{\text{Syn}}(\boldsymbol{c}_o, \emptyset)$, in parallel by randomly zeroing $\boldsymbol{c}_s$, and in both conditional and unconditional cases uses the empty prompt embedding in the SD U-Net during training.

4. Results

4.1. Implementation Details

We train our models for a single epoch on 200K samples from the SURREAL [37] dataset, containing SMPL-annotated images, additionally we annotate these samples with 2D body-pose maps (images containing 25 color-coded joint keypoints) using OpenPose [5] to obtain $\boldsymbol{c}_o$. Notably, no prompts are used to train C_{SMPL}. We did not see an improvement in the metrics when training on more samples from the dataset. The training was carried out on a single NVIDIA RTX 4090 with a learning rate of 10^{-5}. We adapt the pytorch-based huggingface diffusers library to implement our method, and initialize the Stable Diffusion U-Net with SD1.5 weights[1] [26] and C_{SMPL} with weights of the ControlNet-OpenPose[2] [39]. During inference, we also use these weights for the domain appearance guidance net (C_{SD}), or alternatively use the pose-conditioned T2I-Adapter-OpenPose[3].

4.2. Visual Fidelity

Generation Method	IS ↑	KID ↓
ControlNet-OpenPose[2] [39]	<u>15.673</u>	0.00614
T2I-Adapter-OpenPose[3] [22]	**15.755**	<u>0.00609</u>
C_{SMPL}-ft-attn + CN	15.521	0.00615
C_{SMPL}-ft-attn + T2I	15.474	**0.00597**
Ablation Variants:		
C_{SMPL}-extra-attn	15.157	0.01470
C_{SMPL}-ft-all + CN	15.222	0.00637
C_{SMPL}-ft-all + T2I	15.356	0.00607
Reference: COCO Eval Split	15.280	-

Table 1. Inception Score (IS) and Kernel Inception Distance (KID) to the Coco Eval set on images generated from our evaluation set

We assess visual fidelity by using poses and prompts from an evaluation set of annotated images to generate similar images, against which we benchmark fidelity metrics.

Study Setup. We curated a subset of the MSCOCO dataset [16] consisting of 5276 images, each containing one person with at least 90% of keypoints visible and covering at least 10% of the width and length of the image. Using the image-based SMPL predictor Hierarchical Probabilistic Human Estimation (HPH) [29], which performs comparatively well for shape extraction, we predicted SMPL models and used them together with the 2D poses and COCO captions as inputs to the benchmarked models. This setup generated human-centric photographic images with visual appearance and semantic content similar to our curated images. We compare our proposed methods, which involve fine-tuning only cross-attention blocks with ControlNet (ft-attn + CN) and T2I-Adapter (ft-attn + T2I) guidance, and as an ablated variant, fine-tuning the entire model (ft-all). To assess domain gap, we also benchmark a ControlNet (C_{SMPL}-extra-attn) in which we retain prompt inputs and instead fine-tune additional SMPL-cross-attention blocks which are inserted after prompt cross-attention. During inference using C_{SMPL}-extra-attn we do not use guidance composition.

Figure 4. Varying shape parameters for fixed pose and prompt

Metrics. We computed the Inception Score (IS) [27], evaluating the quality and diversity of the generated images, and the Kernel Inception Distance (KID) [3] which is similar to the FID [9], but is more reliable for unbiased evaluation of distribution distances on smaller datasets.

Results. As shown in Table 1, our attention fine-tuning approach with CN or T2I guidance achieves visual fidelity comparable to the baselines. T2I guidance degrades slightly more in terms of IS from its baseline but achieves slightly higher visual similarity (lower KID) to COCO than ControlNet guidance. Fine-tuning all weights (C_{SMPL}-ft-all) slightly increases the visual domain shift (higher KID). However, considering the indirect mapping between prompt generation and COCO, the KID variations are small and likely within the margin of error. Notably, the SMPL-conditioned ControlNet without domain adaption (C_{SMPL}-extra-attn) shows a significant domain gap, indicated by a high KID, demonstrating the effectiveness of our domain adaptation approach.

[1] https://huggingface.co/runwayml/stable-diffusion-v1-5
[2] https://huggingface.co/lllyasviel/sd-controlnet-openpose
[3] https://huggingface.co/TencentARC/t2iadapter_openpose_sd14v1

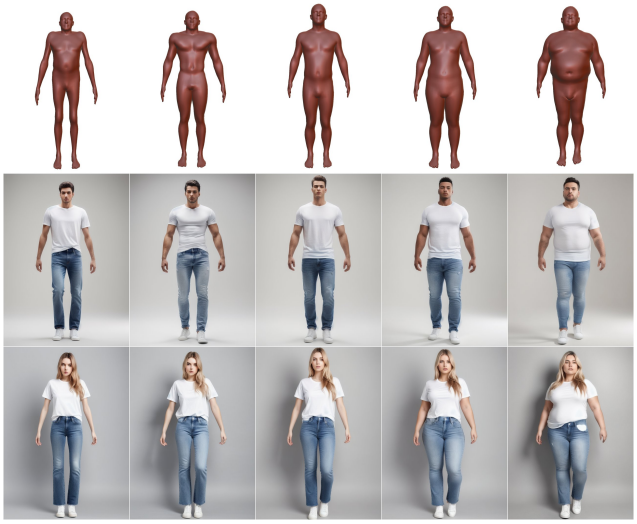


Figure 5. Sampling different shape parameters. The smpl shape (top row) is able to be accurately represented in the image of the clothed persons.

Method	PVET (SC) ↓		MPJPE (PA) ↓	
	AS	AS Ext.	AS	AS Ext.
ControlNet [39]	16.1	20.7	148.8	98.5
T2I-Adapter [22]	15.8	20.8	**118.7**	93.6
Champ [30]	-	17.6	-	**83.0**
C_{SMPL}-ft-attn + CN	14.9	16.0	135.3	98.1
C_{SMPL}-ft-attn + T2I	14.3	**15.2**	119.7	94.1
Ablation Variants:				
C_{SMPL}-ft-all + CN	14.8	16.5	135.5	98.6
C_{SMPL}-ft-all + T2I	**14.2**	15.7	120.1	94.5
GT Images [37]	-	12.1	-	75.3

Table 2. SMPL shape and pose generation accuracy in millimeters. We show the scale and translation corrected Per-Vertex-Error in T-Pose (PVET) and Mean Per Joint Position Error after Procrustes Analysis (MPJPE-PA) between input SMPL params and SMPL meshes extracted using HierProb [29] on the images generated from our evaluation set. AS-Ext uniformly samples body shapes w.r.t obesity and its SMPL models are rendered from a fixed frontal perspective enabling comparison with Champ [30].

4.3. Shape and Pose Accuracy

We assess pose and shape accuracy on a ground-truth set of SMPL parameters by measuring the distance to SMPL parameters estimated from generated images.

Study Setup. We first created an evaluation dataset of 5000 SMPL models, combining poses from the AIST [14] dance dataset with shapes from the SURREAL [37] dataset. This combination dataset (**AS**) addresses the limitations of each dataset: AIST provides diverse poses from different viewpoints but lacks shape diversity, while SURREAL offers varied shapes but limited pose diversity. Using AS, we generated images using our proposed approach, its configurations, and the ControlNet and T2I-Adapter baselines. Using HPH [29], we estimated the SMPL models from the generated images and compared these to the original ground-truth SMPL models.

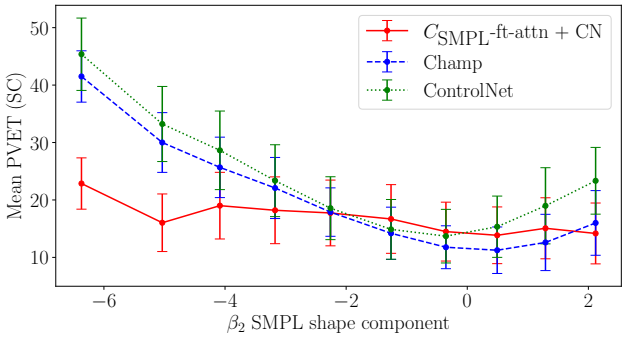
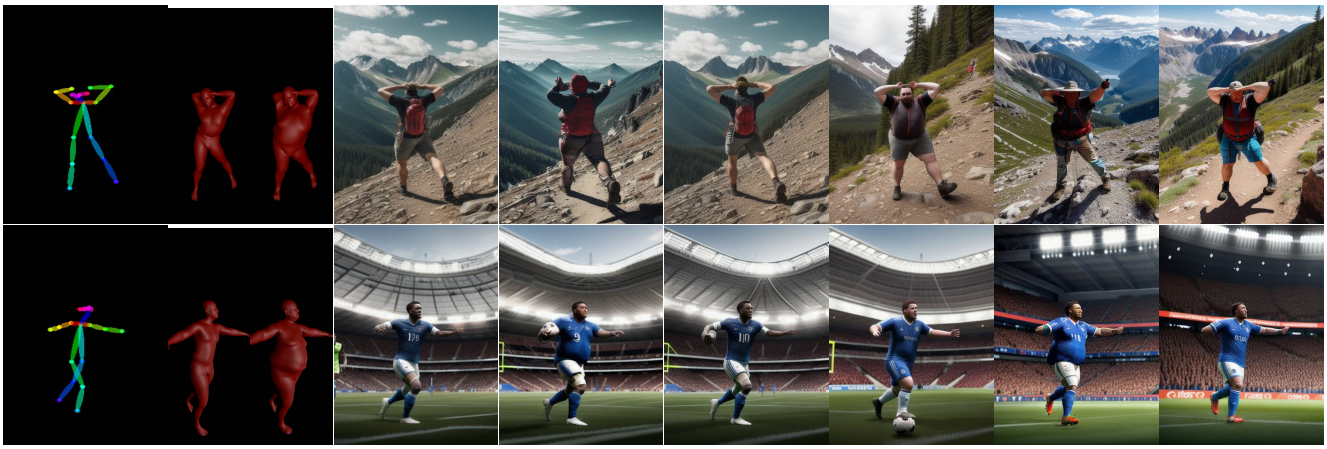


Figure 6. SMPL shape accuracy for extreme shapes. We show Per-Vertex-Error in T-Pose (PVET) for different shape extremes. The second shape component (β_2) strongly correlates with obesity. Error bars show the standard deviation.

SURREAL contains real body shape measurements, which are approximately normal distributed around a mean slim shape, with very limited samples for obese body types. To fairly evaluate the model performance on all body types, we created an extended dataset (AS-Ext) with added samples for obese shapes. Please refer to the supplemental material (Sec. B.1) for details on the creation of the evaluation dataset. Furthermore, to enable comparison with the state-of-the-art approach for SPML-based reference image control Champ [30], we rendered the SMPL models in AS-Ext in a full-frame frontal view to align inputs of our method and Champ in the absence of ground-truth camera matrices. The fixed-view rendered SMPL meshes additionally allow us to establish a baseline noise level for AS-Ext by direct application of HPH.

Metrics. For pose accuracy, we measure the Mean Per Joint Position Error after Procrustes Analysis. For shape accuracy, we measure the scale and translation corrected Per-Vertex-Error in T-Pose [28]. All values are given in mm.

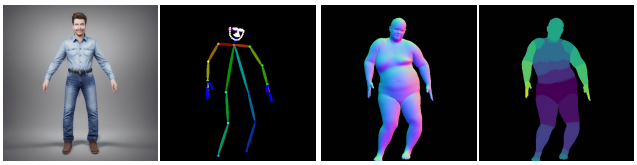
Results. Our results (Tab. 2) show significant improvements in shape accuracy over ControlNet and T2I-Adapter, and Champ on AS-Ext. A stronger deviation from the mean shape degrades the accuracy of ControlNet and Champ, as shown in Fig. 6, while our method achieves relatively constant performance. Interestingly, despite its SMPL conditioning, Champ [30] shows only marginal improvement over the shape-unaware ControlNet, while our model adheres to the conditioning across the range of shapes. Using T2I-Adapter demonstrates better pose adherence than ControlNet, reflected in lower MPJPE-PA values in both the baseline as well as when used in domain guidance (+T2I).



(a) Pose (b) SMPL s_1, s_2 (c) Ours$_{\text{CN}}(s_1)$ (d) Ours$_{\text{CN}}(s_2)$ (e) CN slim (f) CN not slim (g) Ours$_{\text{T2I}}(s_2)$ (h) T2I not slim

Figure 7. Comparison of pose and body shape control. We compare ours, ControlNet [39] (CN) and T2I [22] using slim and overweight body shapes (for ours) or prompts (for CN and T2I). Note that Ours$_{\text{CN}}$ refers to C_{SMPL} + CN, and Ours$_{\text{T2I}}$ refers to C_{SMPL} + T2I. Overall our method displays better stability and accuracy of shapes.

Compared to their baseline models, adding SMPL-guidance enhances pose adherence for ControlNet while having little effect for T2I-Adapter. Champ seems to improve the pose adherence, however other effects such as pose prediction being more accurate on the standardized rendered SMPL meshes may be possible.



(a) Ref. Image (b) Pose Map (c) Normal Map (d) Semantic Map

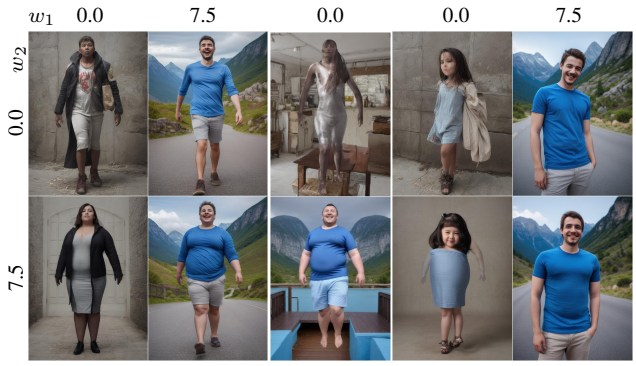
(e) SMPL-based control of reference image with Champ [30]

(f) SMPL-based text-to-image generation with ours

Figure 8. Comparison to SMPL-based reference image control using Champ [30]. Champ uses a SMPL-derived pose map (b), normal map (c), semantic map (d), depth map & mask (not shown) to manipulate a reference image (a). Our reference-free method (f) uses pose map (b) and SMPL parameters as inputs.

4.4. Visual Comparisons

In Fig. 4, we transition between two SMPL shape configurations, primarily affecting mid-body fat distribution, to illustrate a change from overweight to slim. In Fig. 5, we sample different shape parameters to showcase various body proportions for the same prompt: "a man/woman with a white t-shirt and jeans in front of a neutral background." In Fig. 7, we compare our method with ControlNet and T2I. Using a pose from AIST [14], we apply slim and overweight shapes to generate images. For ControlNet and T2I, we include "chubby" in the prompt to create the overweight image. The rows are organized with prompts "a man hiking" and "still of a football player in a FIFA game". Our method consistently adheres to shape information, whereas ControlNet's and T2I's interpretation of "chubby" varies. For example, the first row shows a significantly overweight person, while the shape of the football player is only minimally altered. Using other words that signify overweight creates similar results. Additionally, our method shows improved pose and background stability, as verified by an experiment on masked-out foregrounds and LPIPS measurements; see Sec. B.2 of the supplemental material. We also observe that, while metrically T2I achieves slightly better results for both original and combined with C_{SMPL} (Tabs. 1 and 2), it is more prone to body artifacts than its ControlNet counterpart. In Fig. 8 we show an example from Champ [30], and compare to ours in terms of shape adherence. We generate a reference image with SD of a person in neutral pose in frontal view. Champ then receives semantic maps derived from the SMPL mesh viewed from the same perspective as the reference image (Fig. 8) while our approach receives the SMPL model in parameter form and generates content according to a prompt instead of a reference image. Visually,



(a) Full Approach (b) No DA (c) No pose guidance

Figure 9. Comparing guidance configurations and the isolated effects of individual guidance vectors. (a) The full approach: $\epsilon_{\text{SD}}(\emptyset, \boldsymbol{c}_o) + w_1 \vec{\boldsymbol{g}}_{c_p,c_o} + w_2 \vec{\boldsymbol{g}}_{c_s,c_o}$ (b) Ablated model w/o domain adaptation (DA): $\epsilon_{\text{Syn}}(\emptyset, \emptyset, \boldsymbol{c}_o) + w_2 \vec{\boldsymbol{g}}_{c_p,c_s,c_o}$ (c) No pose guidance (i.e., using SD instead of C_{SD}): $\epsilon_{\text{SD}}(\emptyset) + w_1 \vec{\boldsymbol{g}}_{c_p} + w_2 \vec{\boldsymbol{g}}_{c_s,c_o}$

ours can achieve better and more natural body shape adherence than Champ.

4.5. Animation

Our method can also be applied to generate animated sequences, by using the control signal to drive frame transition-consistent animations using modified versions such as AnimateDiff [8]. We showcase examples in the supplemental material.

4.6. Ablating Guidance Vector Composition

Our classifier-free guidance-based approach consists of a composition of multiple models with different possible input condition configurations. In the following, we explore ablations of Eq. (4). In addition to the two guidance vectors $\vec{\boldsymbol{g}}_{\boldsymbol{c_p},\boldsymbol{c_o}}$ and $\vec{\boldsymbol{g}}_{\boldsymbol{c_s},\boldsymbol{c_o}}$ we ablate configurations using guidance vectors without pose-conditioning ($\vec{\boldsymbol{g}}_{\boldsymbol{c_p}}$) and without domain adaption, for which we use the ft-extra-attn which receives prompt in addition to pose and body ($\vec{\boldsymbol{g}}_{\boldsymbol{c_p},\boldsymbol{c_s},\boldsymbol{c_o}}$), defined as:

$$\vec{\boldsymbol{g}}_{\boldsymbol{c_p}} = \epsilon_{\text{SD}}(\boldsymbol{c}_p) - \epsilon_{\text{SD}}(\emptyset)$$
$$\vec{\boldsymbol{g}}_{\boldsymbol{c_p},\boldsymbol{c_s},\boldsymbol{c_o}} = \epsilon_{\text{Syn}}(\boldsymbol{c}_p, \boldsymbol{c}_s, \boldsymbol{c}_o) - \epsilon_{\text{Syn}}(\emptyset, \emptyset, \boldsymbol{c}_o)$$

In Fig. 9 we evaluate the effects of the composed guidance vectors by generating image grids using the prompt "A man wearing a blue shirt with a happy expression in front of a scenic and cinematic environment, best quality, photography". We compare configurations by selectively setting the guidance scales (w_1, w_2) to 0.0, effectively disabling the guidance components. Fig. 9a shows the configuration of our full approach. The first row and column illustrate the influence of isolated guidance vectors. The relatively stable background when altering w_2 indicates the independence between the vectors, allowing separate scaling of the adherence to the text prompt and SMPL conditioning. Fig. 9b demonstrates the impact of not shifting to the original data distribution, effectively running the original classifier-free guidance in the synthetic data domain. The domain gap from synthetic data results in compromised visual fidelity, with flat textures and floating effects, demonstrating that classifier-free guidance within the fine-tuned model's domain realizes both SMPL conditioning and the text prompt but at the cost of visual quality. The importance of the original 2D pose-conditioned ControlNet is visualized in Fig. 9c. When inferring only using the base SD U-Net, the shape and pose condition are not satisfactorily realized. In Sections B.4 and B.5 of the supplemental material we further demonstrate the negative impact of prompt conditions in ϵ_{Syn} and the visual influence of the guidance scales.

4.7. Limitations

Our approach, while effective in many scenarios, has notable limitations in generating body shapes that conflict with the prompt's content or implied context. For example, specifying athletic professions can in some cases neutralize the generation of obese body shapes. This issue is also present in the original SD model when specifying body weight via a prompt (see Sec. 4.4), suggesting a bias in SD training data or the model's learned representations. Our approach struggles to reconcile this conflicting information, leading to less accurate or unintended results. Please refer to the supplementary material for an example.

5. Conclusion

Our proposed method enables conditional control and generation of diverse human shapes and poses in pretrained text-to-image diffusion models by fine-tuning a ControlNet-based architecture on a 3D human parametric model (SMPL). To address the issue of limited real-world data, we train on synthetic data. To overcome the domain gap from training on less realistic and diverse synthetic scenes, we propose a domain adaptation technique using guidance-based isolation and composition. Our results show that composing the isolated SMPL condition with a domain guidance network can satisfactorily adapt the visual appearance from the synthetic to the original LDM domain. The interchangeability of ControlNet and T2I-Adapter for guidance domain adaptation suggests the potential for generalization in composing isolated conditions and domain guidance networks, possibly allowing multiple synthetically trained conditions to be composed together. Future research should explore the generalization of our guidance domain adaptation technique to other datasets, tasks, and domains.

References

[1] Amir Atapour Abarghouei and Toby Breckon. Real-time monocular depth estimation using synthetic data with domain adaptation via image style transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2800–2810, 06 2018. 2

[2] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8, 2023. 1

[3] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. In *International Conference on Learning Representations*, 2018. 5

[4] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18392–18402, 2023. 2

[5] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 43(1):172–186, 2019. 4, 5

[6] Rinon Gal, Or Patashnik, Haggai Maron, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. Stylegan-nada: Clip-guided domain adaptation of image generators. *ACM Transactions on Graphics (TOG)*, 41(4):1–13, 2022. 2

[7] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7297–7306, 2018. 2

[8] 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. 8

[9] 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. 5

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

[11] Ho, Jonathan, Jonathan Ho, Salimans, Tim, and Tim Salimans. Classifier-Free Diffusion Guidance. *arXiv.org*, July 2022. 4

[12] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8110–8119, 2020. 2

[13] Wouter M Kouw and Marco Loog. A review of domain adaptation without target labels. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 43(3):766–785, 2019. 2

[14] 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. 6, 7

[15] Jian Liang, Ran He, and Tieniu Tan. A comprehensive survey on test-time adaptation under distribution shifts. *International Journal of Computer Vision (IJCV)*, pages 1–34, 07 2024. 2

[16] 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. 5

[17] Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B Tenenbaum. Compositional visual generation with composable diffusion models. In *European Conference on Computer Vision (ECCV)*, pages 423–439. Springer, 2022. 2

[18] 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

[19] 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. 3

[20] Michelle Shu, Charles Herrmann, Richard Strong Bowen, Forrester Cole, and Ramin Zabih. DreamWalk: Style Space Exploration using Diffusion Guidance. *arXiv.org*, 2024. 2

[21] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6038–6047, 2023. 2

[22] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 4296–4304, 2024. 1, 2, 4, 5, 6, 7

[23] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023. 2

[24] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2085–2094, 2021. 2

[25] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learn-

ing transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, pages 8748–8763. PMLR, 2021. 2

[26] 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, 2, 5

[27] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in Neural Information Processing Systems (NeurIPS)*, 29, 2016. 5

[28] Akash Sengupta, Ignas Budvytis, and Roberto Cipolla. Synthetic training for accurate 3d human pose and shape estimation in the wild. *arXiv preprint arXiv:2009.10013*, 2020. 6

[29] 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. 5, 6

[30] Shenhao Zhu, Junming Leo Chen, Zuozhuo Dai, Yinghui Xu, Xun Cao, Yao Yao, Hao Zhu, and Siyu Zhu. Champ: Controllable and Consistent Human Image Animation with 3D Parametric Guidance. *arXiv.org*, 2024. 2, 6, 7

[31] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations (ICLR)*, 2020. 3, 4

[32] Kunpeng Song, Ligong Han, Bingchen Liu, Dimitris Metaxas, and Ahmed Elgammal. Stylegan-fusion: Diffusion guided domain adaptation of image generators. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 5453–5463, 2024. 2

[33] Adéla Šubrtová, Michal Lukáč, Jan Čech, David Futschik, Eli Shechtman, and Daniel Sýkora. Diffusion Image Analogies. In *ACM SIGGRAPH 2023 Conference Proceedings*, SIGGRAPH '23, pages 1–10, New York, NY, USA, July 2023. Association for Computing Machinery. 2

[34] Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *International Conference on Machine Learning (ICML)*, pages 9229–9248. PMLR, 2020. 2

[35] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 1

[36] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1921–1930, 2023. 2

[37] 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. 1, 2, 5, 6

[38] Zhongcong Xu, Jianfeng Zhang, Jun Hao Liew, Hanshu Yan, Jia-Wei Liu, Chenxu Zhang, Jiashi Feng, and Mike Zheng Shou. Magicanimate: Temporally consistent human image animation using diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1481–1490, 2024. 2

[39] 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, 2, 3, 5, 6, 7

[40] Jing Zhao, Heliang Zheng, Chaoyue Wang, Long Lan, Wanrong Huang, and Wenjing Yang. Null-text guidance in diffusion models is secretly a cartoon-style creator. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 5143–5152, 2023. 2