PersonaCraft: Personalized and Controllable Full-Body Multi-Human Scene Generation Using Occlusion-Aware 3D-Conditioned Diffusion

Gwanghyun Kim[1*], Suh Yoon Jeon[1*], Seunggyu Lee[1], Se Young Chun[1,2†]
[1]Dept. of Electrical and Computer Engineering, [2]INMC & IPAI
Seoul National University, Republic of Korea
{gwang.kim, euniejeon, leeseunggyu, sychun}@snu.ac.kr

Figure 1. **PersonaCraft** generates realistic, personalized images of multiple individuals with complex occlusions, preserving facial identity and body shape using occlusion-aware 3D pose and shape conditioned diffusion. PersonaCraft outperforms baselines in body shape personalization (blue arrows indicate failures) and naturalness (yellow arrows highlight artifacts, with zoomed-in views in yellow boxes).

Abstract

We present PersonaCraft, a framework for controllable and occlusion-robust full-body personalized image synthesis of multiple individuals in complex scenes. Current methods struggle with occlusion-heavy scenarios and complete body personalization, as 2D pose conditioning lacks 3D geometry, often leading to ambiguous occlusions and anatomical distortions, and many approaches focus solely on facial identity. In contrast, our PersonaCraft integrates diffusion models with 3D human modeling, employing SMPLx-ControlNet, to utilize 3D geometry like depth and normal maps for robust 3D-aware pose conditioning and enhanced anatomical coherence. To handle fine-grained occlusions, we propose Occlusion Boundary Enhancer Network that exploits depth edge signals with occlusion-focused training, and Occlusion-Aware Classifier-Free Guidance strategy that selectively reinforces conditioning in occluded regions without affecting unoccluded areas. PersonaCraft can seamlessly be combined with Face Identity ControlNet, achieving full-body multi-human personalization and thus marking a significant advancement beyond prior approaches that concentrate only on facial identity. Our dual-pathway body shape representation with SMPLx-based shape parameters and textual refinement, enables precise full-body personalization and flexible user-defined body shape adjustments. Extensive quantitative experiments and user studies demonstrate that PersonaCraft significantly outperforms existing methods in generating high-quality, multi-person images with accurate personalization and robust occlusion handling.

[*]Authors contributed equally. [†]Corresponding author.

Figure 2. SMPLx-ControlNet (SCNet) with SMPLx depth and normal improves occlusion handling in pose generation compared to 2D OpenPose, though fine-grained occlusions remain challenging. Using our occlusion-focused methods, OccNet and OccCFG, we achieve superior pose consistency and anatomical coherence.

1. Introduction

Recent advances in personalized image generation [2, 4, 13–19, 24, 26, 43, 47, 50, 56, 57, 63, 63, 64, 67, 68, 70, 73, 76, 78, 80, 82–84] and controllable image generation [11, 22, 31, 44, 69, 79, 86], powered by diffusion models [9, 27, 33–36, 59, 65, 83], have paved the way for synthesizing multi-human scenes with explicitly defined poses. Despite these breakthroughs, generating images with multiple individuals in complex, occlusion-heavy environments remains a formidable challenge. Conventional approaches that rely on 2D skeleton-based conditioning lack the depth cues necessary to capture overlapping body parts and intricate inter-person interactions, often resulting in ambiguous occlusions and anatomical distortions. Although recent works [11, 48, 86] have begun exploring the use of 3D pose information, they are limited to single-person scenarios, thereby restricting their applicability to multi-person scenes.

Furthermore, achieving full-body personalization remains a significant challenge. Most existing methods prioritize facial identity, often overlooking the distinct characteristics of body shape. Although several approaches [78, 85] have enabled full-body personalization using general image encoders [55], they often struggle to disentangle intrinsic body shape from clothing attributes.

To address these challenges, we propose PersonaCraft, a framework for controllable and occlusion-robust full-body personalized image synthesis of multiple individuals in complex scenes. Firstly, our approach integrates diffusion models with 3D human modeling, employing SMPLx-ControlNet (*SCNet*), to utilize critical 3D geometric information to overcome the limitations of conventional 2D pose conditioning. Through careful exploration of 3D pose representations, we found that combining SMPLx depth and normal maps offers the most robust conditioning, ensuring high pose consistency and anatomical naturalness. However, despite the benefits of 3D-aware conditioning, handling complex, fine-grained occlusions still remains challenging, as shown in Fig. 2. Secondly, we propose the Occlusion Boundary Enhancer Network (*OccNet*), which explicitly enhances occluded regions at a finer level by leveraging depth edge signals that highlight occlusion boundaries. Complemented by our Occlusion-Aware Classifier-Free Guidance (*OccCFG*) strategy, our OccNet adaptively reinforces conditioning in occlusion-heavy areas without compromising contrast in unoccluded regions. Together, these components ensure robust anatomical coherence, even in complex multi-human interactions with significant occlusions, as presented in Fig. 2.

Another key advantage of our 3D-aware pose conditioning, beyond enhancing pose accuracy with improved occlusion handling and anatomical coherence, is its ability to encode body shape through SMPLx body shape coefficients. Our method can seamlessly be combined with Face Identity ControlNet [70] to achieve *full-body multi-human personalization*, representing a significant advancement beyond prior methods focused solely on facial identity [25, 37, 70, 74, 78, 81, 85]. Moreover, to overcome the limitations of SMPLx-based body representation, especially in challenging cases where body shape characteristics conflict, we introduce a dual-pathway approach where an optional textual pathway refines body shape. PersonaCraft enables fine-scale, user-defined body shape control through reference-based and interpolation/extrapolation-based adjustments.

Through extensive quantitative and qualitative evaluations, including user studies, our PersonaCraft shows significant improvements in occlusion handling, pose consistency, setting a new benchmark for anatomically precise controllable and personalized multi-human image synthesis.

2. Related Works

Single-Concept Personalization. Text-to-image (T2I) diffusion models and applications [9, 27, 33–36, 59, 65, 83] enable single-concept personalization, adapting pre-trained models for individual subjects. Early methods used optimization-based techniques [13, 16, 24, 26, 56, 57, 64] or textual embeddings [2, 17, 50, 67, 68, 80, 82]. LoRA methods [28, 66] reduced the need for many trainable parameters. Recent works [4, 14, 15, 18, 19, 43, 47, 63, 63, 70, 73, 76, 78, 83, 84] like IP-Adapter [78] and InstantID [70] adopt modular designs for fast personalization from a single reference image and human-centric modules. However, these methods face challenges in handling complex human poses, neglect body shape preservation, or fail to disentangle identity from clothing in personalized human image synthesis.

Multi-Concept Personalization. Recent multi-concept personalization methods [5, 20, 23, 45] use cross-attention to prevent concept entanglement. Custom Diffusion [39] integrates models via joint training or constrained optimization. Mix-of-Show [21] employs gradient fusion for identity preservation and regional sampling for attribute binding. Modular Customization [53] isolates concepts via orthog-

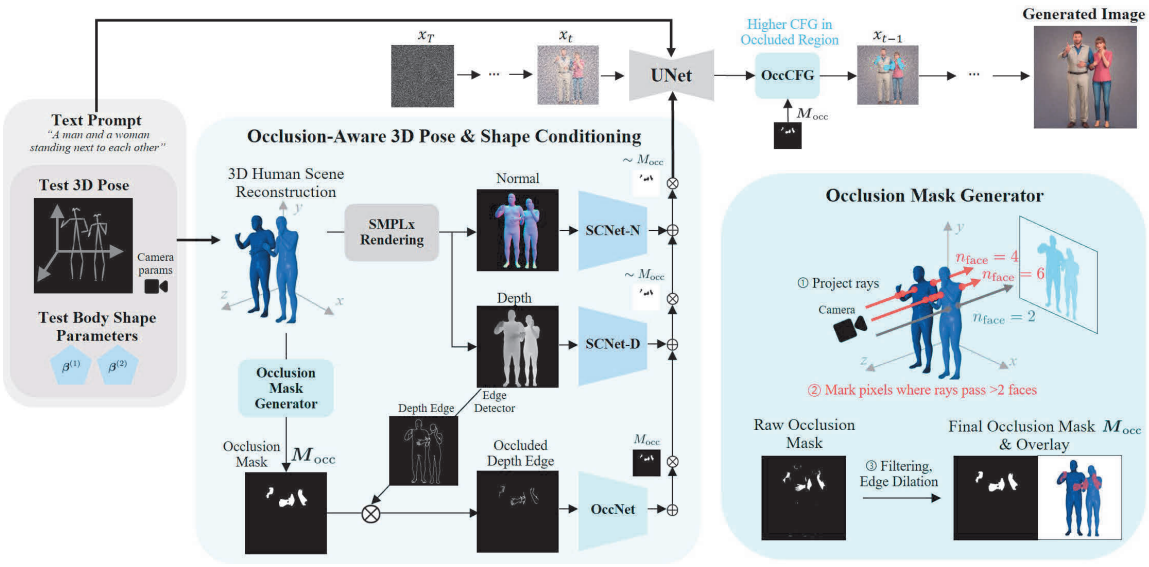

Figure 3. Overview of our occlusion-aware 3D pose and shape conditioning. We generate SMPLx renderings for SMPLx-ControlNet (*SCNet*) and derive occlusion masks for the Occlusion Boundary Enhancer Network (*OccNet*) by counting intersected faces and masking depth edges. Combining SCNet and OccNet residuals with occlusion masks enables the base U-Net to handle occlusion-aware 3D pose and shape conditioning. Occlusion-aware classifier-free guidance (*OccCFG*) further improves anatomical coherence in occluded regions.

onal directions. FastComposer [74] enhances training-free personalization. OMG [37] propose 2-stage multi-concept personalization without training. StoryMaker [85] ensures multi-character consistency via segmentation-constrained cross-attention. UniPortrait [25] proposess ID embeddings and routing modules for high fidelity and editability. ID-Patch [81] mitigates ID leakage and ensures precise identity-position association introducing ID patches. However, existing methods still exhibit limitations in maintaining occlusion robustness and precise control over body shapes.

Controllable Human Generation. Recent works, such as ControlNet [79] and HyperHuman [44], integrate T2I diffusion with 2D skeletons for controllable human synthesis. HumanSD [31] enhances skeleton-based generation with heatmap-guided denoising, while Stable-Pose [69] improves pose alignment using attention masking. DWpose [77] refines OpenPose [12] for better skeleton accuracy, and DensePose [22] maps RGB images to 3D surfaces. However, lacking depth cues, these methods struggle with occlusions in multi-human or complex pose scenarios. PODIA-3D [35] incorporates depth for pose guidance but not image synthesis. The SMPL [46] (Skinned Multi-Person Linear Model) and SMPLx [51] are parametric human body models that provide realistic 3D representations of human pose and shape. SMPL models the body using shape parameters and pose parameters with a differentiable blend skinning function, making it widely used in vision and graphics applications. SMPLx extends this by incorporating expressive facial and hand articulation, enabling more detailed human motion modeling. Recent works leverage SMPL for generative tasks: Champ [86] applies SMPL for single-human video synthesis, while HumanLDM [11] integrates SMPL with ControlNet for stable pose-conditioned human generation. However, these methods focus solely on single-human cases and do not address complex occlusions in multi-human scenarios.

3. Proposed Method: PersonaCraft

We propose PersonaCraft, a framework for controllable and occlusion-robust full-body personalized image synthesis of multiple individuals in complex scenes. PersonaCraft integrates diffusion models with 3D human modeling, leveraging SMPLx-ControlNet (*SCNet*) to incorporate 3D geometry, including depth and normal maps, for robust 3D-aware pose conditioning and enhanced anatomical coherence (Sec. 3.1). To handle fine-grained occlusions, we introduce the Occlusion Boundary Enhancer Network (*OccNet*), which exploits depth edge signals through occlusion-focused training, and the Occlusion-Aware Classifier-Free Guidance (*OccCFG*) strategy, which selectively reinforces conditioning in occluded regions without affecting unoccluded areas (Sec. 3.2). Furthermore, PersonaCraft seamlessly integrates with Face Identity ControlNet, enabling full-body multi-human personalization and surpassing prior approaches that focus solely on facial identity. Our dual-pathway body shape representation, combining SMPLx-based shape parameters with textual refinement, ensures precise full-body personalization and flexible user-defined body shape adjustments (Sec. 3.3).

Existing methods rely on 2D skeleton-based ControlNets [79], which lack depth cues and fail to capture identity-specific body shapes. We propose a 3D-aware pose conditioning, SMPLx-ControlNet (*SCNet*), that leverages SMPLx [51] to incorporate depth information, enabling occlusion-aware conditioning and body shape control.

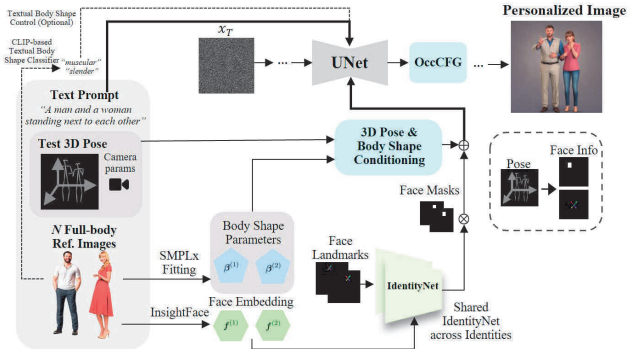

Figure 4. Full pipeline of PersonaCraft for multi-human full-body personalization. By integrating our method with face embeddings from InsightFace [1] and Face Identity ControlNet [70], we advance toward full-body multi-human personalization.

3.1. 3D-Aware Pose and Shape Conditioning for Multi-Human Generation

Existing methods rely on 2D skeleton-based ControlNets [79], which lack depth cues and fail to capture identity-specific body shapes in multi-human scenes. We propose a 3D-aware pose conditioning technique using SMPLx-ControlNet (*SCNet*). By leveraging SMPLx [51], we represent pose with depth information, enabling occlusion and controlling body shape. As shown in Fig. 3, given 3D poses $\{\boldsymbol{p}^{(i)}\}_{i=1}^{N_{\text{human}}}$, body shape parameters $\{\boldsymbol{\beta}^{(i)}\}_{i=1}^{N_{\text{human}}}$, and camera parameters $\boldsymbol{c}$, we construct a 3D human scene of N_{human} individuals with occlusions using SMPLx meshes. SMPLx renderings, like depth $\boldsymbol{d}_{\text{SMPLx}}$, serve as conditioning signals for the diffusion model, enabling high-fidelity image generation with improved identity preservation and occlusion handling.

We train SCNet $\mathcal{E}_\phi^{\text{SC}}$ by leveraging the pretrained ControlNet [79] framework, with the following objective:

$$\mathbb{E}_{\boldsymbol{x}_0,t,y,\boldsymbol{d}_{\text{SMPLx}},\boldsymbol{\epsilon}}\left[\left\|\boldsymbol{\epsilon}-\boldsymbol{\epsilon}_\theta\left(\boldsymbol{x}_t,t,y,\mathcal{E}_\phi^{\text{SC}}(\boldsymbol{d}_{\text{SMPLx}})\right)\right\|_2^2\right], \quad (1)$$

where $\boldsymbol{\epsilon}\sim\mathcal{N}(0,1)$, $\boldsymbol{\epsilon}_\theta$ is a base text-to-image diffusion model, t is the diffusion timestep, y is the text prompt, $\boldsymbol{x}_t$ is the noisy latent representation of the image at timestep t, and it gets progressively refined during the denoising process. We evaluate different SMPLx renderings—depth, normal, and RGB—for conditioning and find that combining depth SCNet (SCNet-D, $\mathcal{E}_{\phi_D}^{\text{SC}}$, for occlusion cues) and normal SCNet (SCNet-N, $\mathcal{E}_{\phi_N}^{\text{SC}}$, for surface orientation) by integrating the residuals from each SCNet provides the most robust pose consistency and anatomical coherence (see Tab. 5). Therefore, we adopt depth-normal as the default 3D representation for SCNet.

3.2. Occlusion-Focused Generation

Although 3D-aware pose conditioning improves handling occluded regions, generating small, detailed occlusions remains difficult due to inter-person contact and self-occlusion, which cause anatomical distortions. To address this, we introduce occlusion masks from SMPLx-rendered 3D scenes, marking pixels as occluded if covered by multiple surfaces, as shown in Fig. 3 (right). For each pixel (i,j), we count the number of faces $n_{\text{faces}}(i,j)$ that a projected ray passes through. A pixel is considered occluded if the ray intersects more than two surfaces. The occlusion mask M_{occ} is defined as:

$$M_{\text{occ}}(i,j)=\begin{cases}1, & \text{if } n_{\text{faces}}(i,j)>2\\ 0, & \text{otherwise}\end{cases} \quad (2)$$

To reduce subtle occlusions and prevent overfitting, we filter small occlusions and refine the boundaries by dilating the mask's edges. This improves the representation of occluded regions, especially in poses with inter-person contact and self-occlusion.

Occlusion Boundary Enhancer Network. Using this occlusion mask, we introduce an Occlusion Boundary Enhancer Network (*OccNet*) by training the model to generate images based solely on *SMPLx depth edges* within the occlusion mask. The SMPLx depth edges, defined as $\boldsymbol{e}_{\text{SMPLx}}=\left|\frac{\partial \boldsymbol{d}_{\text{SMPLx}}}{\partial x}\right|+\left|\frac{\partial \boldsymbol{d}_{\text{SMPLx}}}{\partial y}\right|>\tau$, where τ is the edge threshold, highlight occlusion boundaries where depth changes abruptly. Although this information is limited due to the loss of 3D information inside the occluded regions, it serves as a crucial signal by highlighting the occlusion boundaries.

As shown in Fig. 3, we only provide the depth edges within the occluded region and train the model to generate images based on this sparse signal, which enables the model to focus more on generating the occluded region while preserving its boundaries, compared to training with the full conditioning. We train OccNet $\mathcal{E}_\omega^{\text{Occ}}$ using the following objective:

$$\mathbb{E}_{\boldsymbol{x}_0,t,y,\boldsymbol{e}_{\text{SMPLx}},\boldsymbol{M}_{\text{occ}},\boldsymbol{\epsilon}}\left[\left\|\boldsymbol{\epsilon}-\boldsymbol{\epsilon}_\theta\left(\boldsymbol{x}_t,t,y,\mathcal{E}_\omega^{\text{Occ}}\left(\boldsymbol{e}_{\text{SMPLx}}\odot\boldsymbol{M}_{\text{occ}}\right)\right)\right\|_2^2\right]. \quad (3)$$

Occlusion-Aware Classifier-Free Guidance. We observe that increasing classifier-free guidance (CFG) in occluded regions enhances anatomical consistency by strengthening 3D pose information, resolving local ambiguities. However, a uniformly high CFG strength induce over-saturation in non-occluded areas, as noted in prior works [40, 58, 59]. This highlights the need for a spatially adaptive guidance strategy [62] tailored for occlusion in human scene scenarios. To this end, we introduce Occlusion-Aware Classifier-Free Guidance (*OccCFG*) as shown in Fig. 3:

$$\hat{\boldsymbol{\epsilon}}=\boldsymbol{\epsilon}_{\text{uncond}}+(k_{\text{occ}}\boldsymbol{M}_{\text{occ}}+k_{\text{base}}(1-\boldsymbol{M}_{\text{occ}}))(\boldsymbol{\epsilon}_{\text{cond}}-\boldsymbol{\epsilon}_{\text{uncond}}) \quad (4)$$

where k_{occ} and k_{base} are CFG scales for occluded and non-occluded regions, respectively, $\boldsymbol{M}_{\text{occ}}$ is the occlusion mask, $\boldsymbol{\epsilon}_{\text{uncond}}$ is the unconditional noise prediction, and $\boldsymbol{\epsilon}_{\text{cond}}$ is the conditional noise prediction.

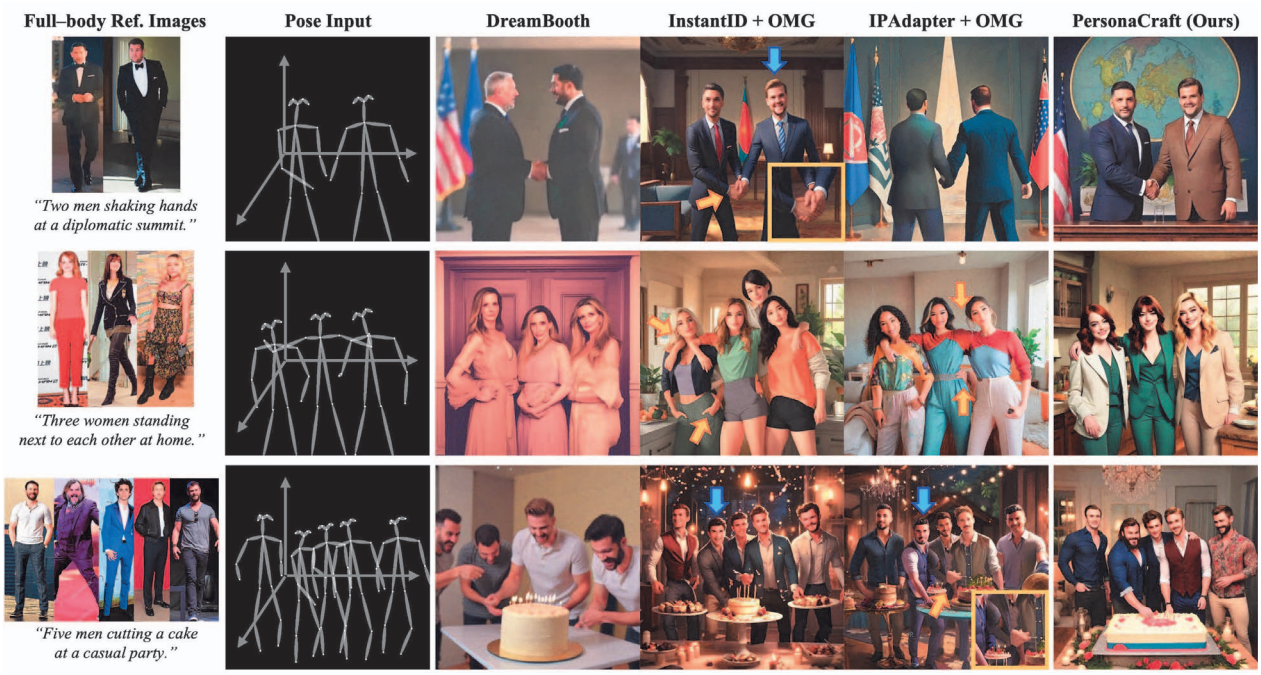

Figure 5. Qualitative comparison of personalized multi-human scene generation. Blue arrows indicate failures in body shape personalization, while yellow arrows highlight unnatural anatomical structures, with yellow boxes providing zoomed-in views.

Table 1. Quantitative evaluation of multi-human personalization across face identity, body shape preservation, pose accuracy, text alignment, and image quality. 'Single' denotes personalization for a single individual, while 'Multi' refers to cases with multiple identities (2–5), with 'Total' representing the averaged results.

Multi-Human Personalization	Face ID preservation↑			Body shape preservation↑			Pose		Text	Image quality	
	Single	Multi	Total	Single	Multi	Total	MPJPE (3D)↓	AP-0.5 (2D)↑	CLIP sim ↑	IS ↑	KID ↓
InstantID + OMG	0.369	0.199	0.227	0.520	0.345	0.374	112.6	0.258	0.267	3.923	0.101
IPAdapter+ OMG	0.228	0.129	0.142	0.565	0.376	0.401	116.3	0.244	0.266	4.133	0.102
IPA-Face + OMG	0.310	0.166	0.188	0.515	0.341	0.367	115.4	0.249	0.267	4.021	0.102
Ours	**0.421**	**0.298**	**0.317**	**0.630**	**0.548**	**0.560**	**60.65**	**0.506**	**0.273**	**4.237**	**0.093**

Table 2. Top-1 preference rates from the user study on naturalness, face identity, body shape, and text-image correspondence in personalized multi-human scene generation.

Top 1 (%)	Natural.↑	Face ID↑	Body shape↑	Text corr.↑
Textual Inversion	9.77	11.46	10.29	8.41
DreamBooth	15.4	11.13	13.33	13.01
InstantID + OMG	10.74	11.65	11.59	12.62
IPAdapter + OMG	12.75	11.78	12.62	14.3
IPA-Face + OMG	10.87	9.51	10.68	10.74
Ours	**40.45**	**44.47**	**41.49**	**40.91**

3.3. Full-Body Personalized Generation

Beyond robust pose, SMPLx's key advantage is encoding body shape with coefficients, enabling full-body multi-human personalization. Unlike prior methods that only focus on facial identity and struggle to disentangle garments from body shape, our approach explicitly isolates body shape as the true identity.

The overall process of full-body personalized image synthesis is illustrated in Fig. 4. Given full-body reference images $\{I_{\text{ref}}^{(i)}\}_{i=1}^{N_{\text{human}}}$ of N_{human} individuals, target poses $\{\boldsymbol{p}^{(i)}\}_{i=1}^{N_{\text{human}}}$, and a text prompt y, our goal is to generate a personalized image I_P that faithfully preserves both face and body identities while ensuring pose consistency. To achieve this, we use MultiHMR [8] to estimate SMPLx body shape parameters $\boldsymbol{\beta}^{(i)}$ and InsightFace [1] to extract face embeddings $\boldsymbol{f}^{(i)}$.

We then leverage our depth-normal SCNet $\mathcal{E}_{\phi_D}^{\text{SC}}$-$\mathcal{E}_{\phi_N}^{\text{SC}}$, OccNet $\mathcal{E}_{\omega}^{\text{Occ}}$, and face ControlNet IdentityNet [71] $\mathcal{E}_{\psi}^{\text{ID}}$. Let $F_{\theta}^{(k)}(\cdot)$ denote the k-th neural block and $\boldsymbol{s}^{(k)}$ the corresponding input feature map. When $\boldsymbol{R}_{\text{sc}} = \mathcal{E}_{\phi_D}^{\text{SC}}(\boldsymbol{d}_{\text{SMPLx}}) + \mathcal{E}_{\phi_N}^{\text{SC}}(\boldsymbol{n}_{\text{SMPLx}})$, $\boldsymbol{R}_{\text{occ}} = \mathcal{E}_{\omega}^{\text{Occ}}(\boldsymbol{d}_{\text{SMPLx}})$, and $\boldsymbol{R}_{\text{id}}^{(i)} = \mathcal{E}_{\psi}^{\text{ID}}(\boldsymbol{f}^{(i)}, \boldsymbol{p}_{\text{face}}^{(i)})$ are the residual features from each model, the next feature map $\boldsymbol{s}^{(k+1)}$ is obtained by adding these residual features to the neural block output $F_{\theta}^{(k)}(\boldsymbol{s}^{(k)})$, scaled by their conditioning weights α_{sc}, α_{occ}, and α_{id}, and modulated by the face masks $\boldsymbol{M}_{\text{face}}^{(i)}$ and occlusion mask $\boldsymbol{M}_{\text{occ}}$:

$$\begin{aligned}\boldsymbol{s}^{(k+1)} = F_{\theta}^{(k)}(\boldsymbol{s}^{(k)}) + \alpha_{\text{sc}}(1 - \boldsymbol{M}_{\text{occ}})\boldsymbol{R}_{\text{sc}} \\ + \alpha_{\text{occ}}\boldsymbol{M}_{\text{occ}}\boldsymbol{R}_{\text{occ}} + \alpha_{\text{id}} \sum_{i=1}^{N_{\text{human}}} \boldsymbol{M}_{\text{face}}^{(i)} * \boldsymbol{R}_{\text{id}}^{(i)}.\end{aligned} \tag{5}$$

Dual-Pathway Body Shape Personalization. We intro-

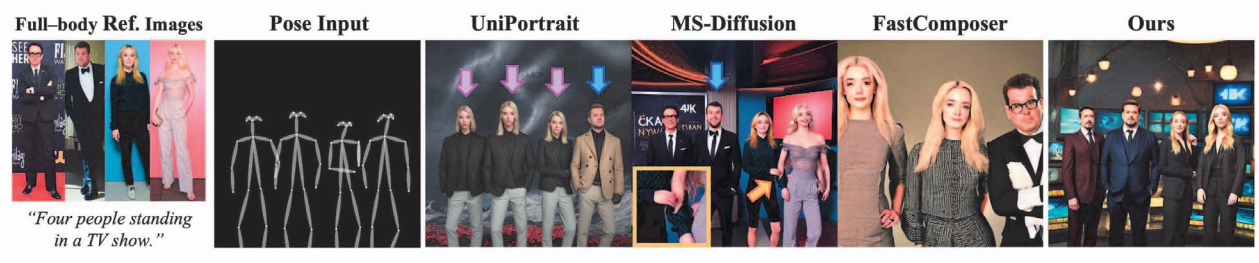

Figure 6. Comparison of personalized multi-human scene generation with the methods which are not direct baselines. Yellow arrows highlight anatomical inconsistencies in poses and occlusions, while Pink arrows indicate identity mixing or duplication in the baselines.

Figure 7. Qualitative comparison of pose-controlled multi-human generation. yellow arrows highlight unnatural anatomical structures, with yellow boxes providing zoomed-in views.

Table 3. Quantitative evaluation of pose-controlled human generation across pose accuracy, text alignment, and image quality.

Pose-Controlled Human Generation	Pose		Text	Image quality	
	MPJPE (3D) ↓	AP-0.5 (2D) ↑	CLIP sim ↑	IS ↑	KID ↓
T2I Adpater-SDXL	198.14	0.323	0.269	3.740	0.129
ControlNet-SDXL	114.93	0.291	**0.285**	3.062	0.138
ControlNet-Flux	102.64	0.393	0.218	3.651	0.107
HumanSD-SD2	103.62	0.357	0.261	3.610	0.082
Ours	**62.647**	**0.495**	0.274	**4.113**	**0.091**

Table 4. Top-1 preference rates from the user study on naturalness, pose consistency, and text-image correspondence in pose-controlled multi-human scene generation.

Top 1 (%)	Natural.↑	Pose consistency↑	Text corr.↑
T2I Adpater-SDXL	9.53	7.31	8.19
ControlNet-SDXL	17.72	18.19	19.36
Ours	**72.75**	**74.50**	**72.46**

duce a dual-pathway approach for body shape personalization, leveraging both SMPLx-based and text-based representations. Compared to textual descriptions of body shape, SMPLx-based representations provide more spatially fine-grained encoding of the identity. However, they are limited in accurately conveying body composition details, such as fat and muscle distribution [52]. In challenging cases where body shape characteristics conflict (e.g., a muscular physique with high body fat), our method optionally incorporates a textual pathway to enhance body shape representation. Specifically, a CLIP-based classifier [55] extracts body shape attributes in text, which are applied using regional prompting [6] and integrated with SCNet. Further details are provided in the supplementary material.

4. Experiments

Implementation Details. PersonaCraft is built on Stable Diffusion XL (SDXL) [54]. We extend the MPII dataset [3] with textual descriptions [32] and 3D human reconstructions from MultiHMR [8] using SMPLx [51]. SCNet and OccNet were fine-tuned on ControlNet [79] with a learning rate of 1e-5, Adam optimizer, batch size of 2, for 50,000 iterations on an NVIDIA A100 GPU. We use 30 denoising steps. We set the base CFG scale to 3 ($k_{\text{base}} = 3$) and increased the scale for occluded regions to 5 ($k_{\text{occ}} = 5$) to enhance occlusion handling. We set all conditioning scales (α_{sc}, α_{occ}, and α_{id}) to 0.8 across all cases.

Metrics. For personalized multi-human scene generation, face identity preservation is measured using FaceNet [61], following FastComposer [74]. Body shape preservation is

Figure 8. Effect of our occlusion-aware 3D pose & shape conditioning components. D and N denote the conditioning types of SCNet: depth, normal.

Figure 9. Effectiveness of dual-pathway body shape personalization in challenging cases where body shape characteristics conflict (e.g., a muscular physique with high body fat).

Table 5. Ablation study on SCNet, OccNet, and OccCFG for pose consistency. D, N, and R denote the conditioning types of SCNet: depth, normal, and RGB representations, respectively.

Ablation Study	Pose	
	MPJPE (3D)↓	AP-0.5 (2D)↑
SCNet-D	71.20	0.423
SCNet-N	79.30	0.398
SCNet-R	82.52	0.394
SCNet-D + SCNet-N	63.16	0.494
SCNet-D + SCNet-N + OccNet	62.92	**0.499**
SCNet-D + SCNet-N + OccNet + OccCFG	**62.64**	0.495

evaluated via cosine similarity of SMPLx body shape parameters β. Pose consistency is assessed using MPJPE (3D) [29], while 2D pose estimation is evaluated with AP-0.5 (2D) by comparing the target pose to the estimated pose from generated images. Text-image alignment is measured via CLIP-L/14 similarity. Image quality is quantified using Inception Score (IS) [60] and Kernel Inception Distance (KID) [10]. A user study evaluates perceptual quality.

For pose-controlled multi-human scene generation, the same key metrics are applied: MPJPE (3D) and AP-0.5 (2D) for pose accuracy, CLIP similarity for text-image alignment, and IS/KID for image quality.

Test Dataset. We test on the COCO-WholeBody dataset [30], containing 1∼5 people per image. A total of 1,000 images were selected (200 per group). Text prompts were extracted via BLIP [42]. MultiHMR [7] provided poses and 3D conditioning.

Additional details on experiments, results (including qualitative results), stylized artwork, PersonaCraft with LoRA and other face control modules, as well as ablation studies, are provided in the supplementary material.

4.1. Personalized Multi-Human Scene Generation

Baselines. We mainly compared PersonaCraft with baseline methods for single-shot, multi-identity, and pose-controllable synthesis, all implemented using SDXL [54]. Key baselines include OMG [37] with InstantID/IPAdapter/IPAdapter-Face (IPA-Face) [70, 78], and 2D pose ControlNet [79]. We also evaluated optimization-based methods like DreamBooth [56] and Texture Inversion [17]. Both qualitative and quantitative comparisons, as well as a user study, were conducted. Additionally, qualitative comparisons were made with Uni-Portrait [25], MS-Diffusion [72], and FastComposer [75].

Qualitative Results. As shown in Fig. 5, PersonaCraft surpasses baselines in multi-identity scene generation, particularly in identity preservation and occlusion handling. Our 3D-aware conditioning and occlusion-focused generation enhance the preservation of body shapes and identity features, and ensure pose fidelity, producing realistic generations with fewer artifacts. Baseline methods struggle with occlusion and body shape accuracy, often leading to distorted identities and unrealistic compositions, especially in multi-subject scenes. PersonaCraft excels in occlusion-aware full-body personalization, maintaining consistent identities across varying poses.

Quantitative Results. We evaluate PersonaCraft on face ID preservation, body shape, pose accuracy, text-image alignment, and image quality. As shown in Tab. 1, our method consistently outperforms baselines in identity and body shape preservation. It achieves the lowest MPJPE (3D) and highest AP (2D), indicating superior alignment with input poses. In terms of text-image alignment, ours also achieves the highest CLIP score. Furthermore, PersonaCraft outperforms baselines in IS and KID, producing perceptually aligned generations.

User Study. We conducted a user study evaluating naturalness, face identity, body shape, and text-image correspondence across baselines and our method. With 18,540 responses from 103 participants, Tab. 2 shows our method achieved the highest Top-1 preference across all metrics, demonstrating superior perceptual quality in personalized multi-human scene generation.

4.2. Pose-Controlled Multi-Human Scene Generation

Baselines. Key baselines include ControlNet-SDXL [54, 79], T2I Adapter-SDXL [49], ControlNet-Flux [41, 79], and HumanSD [41]. Both qualitative and quantitative comparisons were conducted for all baselines. Additionally, a user study was performed specifically for SDXL-based baselines, including ControlNet-SDXL and T2I Adapter-SDXL.

Qualitative Results. We evaluate pose-controlled multi-human scene generation by comparing PersonaCraft with existing methods in both single-human and multi-human settings (Fig. 7). Our method achieves superior pose alignment compared to baseline methods. Baselines relying on 2D pose conditioning (e.g., T2IAdapter-SDXL [49], ControlNet-

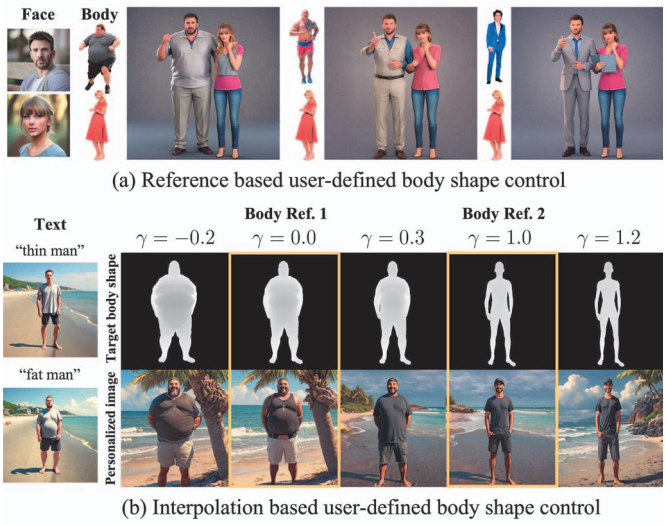

Figure 10. Result of PersonaCraft's user-defined body shape control. (a) Reference-based body shape control. (b) Interpolation and extrapolation-based control, where γ controls the interpolation or extrapolation between the reference body shapes.

SDXL [54, 79]) struggle with occlusion handling, leading to misaligned poses and distorted structures. ControlNet-Flux [41, 79] exhibits instability: at a lower conditioning scale (0.8), it fails to maintain anatomical coherence, while at a higher scale (1.0), it improves pose accuracy but reduces image fidelity. HumanSD [31] lacks precise pose control, producing unrealistic compositions. In contrast, PersonaCraft balances pose accuracy, anatomical coherence, and image fidelity, demonstrating the efficacy of our 3D-aware pose conditioning.

Quantitative Results. We evaluate pose-controlled multi-human scene generation in terms of pose accuracy, text-image correspondence, and image quality. As shown in Tab. 3, PersonaCraft achieves the lowest MPJPE (3D) and highest AP-0.5 (2D), demonstrating precise pose alignment and keypoint localization. Compared to ControlNet-based baselines, which suffer from structural inconsistencies and occlusion issues, our method maintains accurate pose fidelity. PersonaCraft also achieves the highest IS and lowest KID, indicating superior perceptual realism. While text-image correspondence scores are comparable across methods, prior studies [36, 38] highlight CLIP's limitations in evaluating complex multi-human scenes. A user study further analyzes perceptual quality.

User Study. We conducted a user study to assess naturalness, pose consistency, and text-image correspondence. Participants ranked their top three preferences among T2I Adapter-SDXL, ControlNet-SDXL, and PersonaCraft, collecting 15,390 responses from 114 participants. As shown in Tab. 4, PersonaCraft achieved the highest Top-1 preference across all metrics, excelling in naturalness (72.75%), pose consistency (74.50%), and text alignment (72.46%). These results confirm that our occlusion-aware 3D pose conditioning significantly improves perceptual quality, producing well-posed, visually coherent multi-human scenes.

4.3. Ablation Studies and Analysis

Comparison of 3D Pose Representations for SCNet. In Tab. 5, we evaluate different SMPLx renderings—depth, normal, and RGB—as conditioning inputs for SCNet. We find that depth yields the best pose consistency among these representations. Additionally, combining the depth and normal representations achieves the best performance. Therefore, we adopt this combination as the base conditioning for SCNet.

Effectiveness of OccNet and OccCFG. In Tab. 5, we also evaluate the effectiveness of OccNet and OccCFG. We found that OccNet and OccCFG improve pose consistency. Also, as analyzed in Fig. 8, we observe the effect of our Occlusion-aware 3D pose & shape conditioning components. Using only 2D pose leads to front-back ambiguities and structural inconsistencies, while relying solely on 3D pose (SCNet) struggles with fine-grained occluded regions. Using only OccNet improves occluded region synthesis but fails to maintain overall pose alignment. Our full model (SCNet + OccNet + OccCFG) effectively preserves pose structure while handling occlusions.

Dual-Pathway Body Personalization. As shown in Fig. 9, the SMPLx-based pathway captures fine-grained identity features but struggles with detailed body composition (e.g., muscle vs. fat distribution). The text-based pathway refines body shape using semantic attributes. Our approach combines both for more accurate and realistic personalization.

4.4. User-Defined Body Shape Control

A key feature of PersonaCraft is its user-defined body shape control, as shown in Fig. 10. Users can select a reference to adapt body characteristics for personalized images. The system also allows body proportions to be adjusted through interpolation or extrapolation between references, offering enhanced customization for precise body shape control.

5. Conclusion

We present PersonaCraft, a framework for occlusion-robust, full-body personalized image synthesis in complex multi-human scenes. PersonaCraft integrates diffusion models with 3D human modeling, using SMPLx-ControlNet for enhanced pose conditioning. To address occlusions at a fine level, we introduce Occlusion Boundary Enhancer and Occlusion-Aware Classifier-Free Guidance strategy, improving conditioning in occluded areas. Seamlessly combined with Face Identity ControlNet, our approach enables full-body personalization beyond facial identity. Extensive experiments show PersonaCraft outperforms prior arts in personalization accuracy and occlusion robustness in complex scenes.

Acknowledgement

This work was supported in part by Samsung Electronics MX Division and by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) [NO.RS-2021-II211343, Artificial Intelligence Graduate School Program (Seoul National University)]. Also, the authors acknowledged the financial support from the BK21 FOUR program of the Education and Research Program for Future ICT Pioneers, Seoul National University.

References

[1] Insightface: 2d and 3d face analysis project. Github, https://github.com/deepinsight/insightface. 4, 5

[2] Yuval Alaluf, Elad Richardson, Gal Metzer, and Daniel Cohen-Or. A neural space-time representation for text-to-image personalization. *ACM TOG*, 42(6):1–10, 2023. 2

[3] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *CVPR*, 2014. 6

[4] Moab Arar, Rinon Gal, Yuval Atzmon, Gal Chechik, Daniel Cohen-Or, Ariel Shamir, and Amit H. Bermano. Domain-agnostic tuning-encoder for fast personalization of text-to-image models. In *SIGGRAPH Asia*, 2023. 2

[5] Omri Avrahami, Kfir Aberman, Ohad Fried, Daniel Cohen-Or, and Dani Lischinski. Break-a-scene: Extracting multiple concepts from a single image. *arXiv preprint arXiv:2305.16311*, 2023. 2

[6] Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. Multidiffusion: Fusing diffusion paths for controlled image generation. *arXiv preprint arXiv:2302.08113*, 2023. 6

[7] Fabien Baradel*, Matthieu Armando, Salma Galaaoui, Romain Brégier, Philippe Weinzaepfel, Grégory Rogez, and Thomas Lucas*. Multi-hmr: Multi-person whole-body human mesh recovery in a single shot. In *ECCV*, 2024. 7

[8] Fabien Baradel, Matthieu Armando, Salma Galaaoui, Romain Brégier, Philippe Weinzaepfel, Grégory Rogez, and Thomas Lucas. Multi-hmr: Multi-person whole-body human mesh recovery in a single shot. In *ECCV*, pages 202–218. Springer, 2025. 5, 6

[9] 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), 2023. 2

[10] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. *arXiv*, 2018. 7

[11] Benito Buchheim, Max Reimann, and Jürgen Döllner. Controlling human shape and pose in text-to-image diffusion models via domain adaptation. *arXiv preprint arXiv:2411.04724*, 2024. 2, 3

[12] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017. 3

[13] Daewon Chae, Nokyung Park, Jinkyu Kim, and Kimin Lee. Instructbooth: Instruction-following personalized text-to-image generation. *arXiv preprint arXiv:2312.03011*, 2023. 2

[14] Wenhu Chen, Hexiang Hu, Yandong Li, Nataniel Rui, Xuhui Jia, Ming-Wei Chang, and William W Cohen. Subject-driven text-to-image generation via apprenticeship learning. *arXiv preprint arXiv:2304.00186*, 2023. 2

[15] Xi Chen, Lianghua Huang, Yu Liu, Yujun Shen, Deli Zhao, and Hengshuang Zhao. Anydoor: Zero-shot object-level image customization. *arXiv preprint arXiv:2307.09481*, 2023. 2

[16] Jooyoung Choi, Yunjey Choi, Yunji Kim, Junho Kim, and Sungroh Yoon. Custom-edit: Text-guided image editing with customized diffusion models. *arXiv preprint arXiv:2305.15779*, 2023. 2

[17] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *ICLR*, 2022. 2, 7

[18] Rinon Gal, Moab Arar, Yuval Atzmon, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. Designing an encoder for fast personalization of text-to-image models. *arXiv preprint arXiv:2302.12228*, 2023. 2

[19] Rinon Gal, Moab Arar, Yuval Atzmon, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. Encoder-based domain tuning for fast personalization of text-to-image models. *ACM TOG*, 42(4):1–13, 2023. 2

[20] Yuan Gong, Youxin Pang, Xiaodong Cun, Menghan Xia, Haoxin Chen, Longyue Wang, Yong Zhang, Xintao Wang, Ying Shan, and Yujiu Yang. Talecrafter: Interactive story visualization with multiple characters. *SIGGRAPH Asia*, 2023. 2

[21] Yuchao Gu, Xintao Wang, Jay Zhangjie Wu, Yujun Shi, Yunpeng Chen, Zihan Fan, Wuyou Xiao, Rui Zhao, Shuning Chang, Weijia Wu, et al. Mix-of-show: Decentralized low-rank adaptation for multi-concept customization of diffusion models. *NeurIPS*, 2023. 2

[22] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *CVPR*, 2018. 2, 3

[23] Ligong Han, Yinxiao Li, Han Zhang, Peyman Milanfar, Dimitris Metaxas, and Feng Yang. Svdiff: Compact parameter space for diffusion fine-tuning. *arXiv preprint arXiv:2303.11305*, 2023. 2

[24] Shaozhe Hao, Kai Han, Shihao Zhao, and Kwan-Yee K Wong. Vico: Detail-preserving visual condition for personalized text-to-image generation. *arXiv preprint arXiv:2306.00971*, 2023. 2

[25] Junjie He, Yifeng Geng, and Liefeng Bo. Uniportrait: A unified framework for identity-preserving single- and multi-human image personalization. *arXiv preprint arXiv:2408.05939*, 2024. 2, 3, 7

[26] Xingzhe He, Zhiwen Cao, Nicholas Kolkin, Lantao Yu, Helge Rhodin, and Ratheesh Kalarot. A data perspective on enhanced identity preservation for diffusion personalization. *arXiv preprint arXiv:2311.04315*, 2023. 2

[27] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 2020. 2

[28] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *ICLR*, 2021. 2

[29] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *TPAMI*, 36(7):1325–1339, 2013. 7

[30] Sheng Jin, Lumin Xu, Jin Xu, Can Wang, Wentao Liu, Chen Qian, Wanli Ouyang, and Ping Luo. Whole-body human pose estimation in the wild. In *ECCV*, 2020. 7

[31] Xuan Ju, Ailing Zeng, Chenchen Zhao, Jianan Wang, Lei Zhang, and Qiang Xu. Humansd: A native skeleton-guided diffusion model for human image generation. In *ICCV*, pages 15988–15998, 2023. 2, 3, 8

[32] Muhammad Saif Ullah Khan, Muhammad Ferjad Naeem, Federico Tombari, Luc Van Gool, Didier Stricker, and Muhammad Zeshan Afzal. Focusclip: Multimodal subject-level guidance for zero-shot transfer in human-centric tasks, 2024. 6

[33] Gwanghyun Kim and Se Young Chun. Datid-3d: Diversity-preserved domain adaptation using text-to-image diffusion for 3d generative model. In *CVPR*, 2023. 2

[34] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *CVPR*, 2022.

[35] Gwanghyun Kim, Ji Ha Jang, and Se Young Chun. Podia-3d: Domain adaptation of 3d generative model across large domain gap using pose-preserved text-to-image diffusion. In *CVPR*, pages 22603–22612, 2023. 3

[36] Gwanghyun Kim, Hayeon Kim, Hoigi Seo, Dong Un Kang, and Se Young Chun. Beyondscene: Higher-resolution human-centric scene generation with pretrained diffusion. In *ECCV*, 2024. 2, 8

[37] Zhe Kong, Yong Zhang, Tianyu Yang, Tao Wang, Kaihao Zhang, Bizhu Wu, Guanying Chen, Wei Liu, and Wenhan Luo. Omg: Occlusion-friendly personalized multi-concept generation in diffusion models. In *ECCV*, 2024. 2, 3, 7

[38] Max Ku, Dongfu Jiang, Cong Wei, Xiang Yue, and Wenhu Chen. Viescore: Towards explainable metrics for conditional image synthesis evaluation. *arXiv preprint arXiv:2312.14867*, 2023. 8

[39] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *CVPR*, pages 1931–1941, 2023. 2

[40] Tuomas Kynkäänniemi, Miika Aittala, Tero Karras, Samuli Laine, Timo Aila, and Jaakko Lehtinen. Applying guidance in a limited interval improves sample and distribution quality in diffusion models. *Advances in Neural Information Processing Systems*, 37:122458–122483, 2025. 4

[41] Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024. 7, 8

[42] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 7

[43] Zhen Li, Mingdeng Cao, Xintao Wang, Zhongang Qi, Ming-Ming Cheng, and Ying Shan. Photomaker: Customizing realistic human photos via stacked id embedding. *arXiv preprint arXiv:2312.04461*, 2023. 2

[44] 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, 3

[45] Zhiheng Liu, Yifei Zhang, Yujun Shen, Kecheng Zheng, Kai Zhu, Ruili Feng, Yu Liu, Deli Zhao, Jingren Zhou, and Yang Cao. Cones 2: Customizable image synthesis with multiple subjects. *arXiv preprint arXiv:2305.19327*, 2023. 2

[46] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Transactions Graphics*, 34(6): 248:1–248:16, 2015. 3

[47] Yiyang Ma, Huan Yang, Wenjing Wang, Jianlong Fu, and Jiaying Liu. Unified multi-modal latent diffusion for joint subject and text conditional image generation. *arXiv preprint arXiv:2303.09319*, 2023. 2

[48] Yifang Men, Yuan Yao, Miaomiao Cui, and Liefeng Bo. Mimo: Controllable character video synthesis with spatial decomposed modeling. *arXiv preprint arXiv:2409.16160*, 2024. 2

[49] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, 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. 7

[50] Lianyu Pang, Jian Yin, Haoran Xie, Qiping Wang, Qing Li, and Xudong Mao. Cross initialization for personalized text-to-image generation. *arXiv preprint arXiv:2312.15905*, 2023. 2

[51] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *CVPR*, 2019. 3, 4, 6

[52] MPI for Intelligent Systems Perceiving Systems Department. Smpl made simple faqs. 6

[53] Ryan Po, Guandao Yang, Kfir Aberman, and Gordon Wetzstein. Orthogonal adaptation for modular customization of diffusion models. *arXiv preprint arXiv:2312.02432*, 2023. 2

[54] 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. 6, 7, 8

[55] 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*, 2021. 2, 6

[56] 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. In *CVPR*, 2023. 2, 7

[57] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Wei Wei, Tingbo Hou, Yael Pritch, Neal Wadhwa, Michael Rubinstein, and

Kfir Aberman. Hyperdreambooth: Hypernetworks for fast personalization of text-to-image models. *arXiv preprint arXiv:2307.06949*, 2023. 2

[58] Seyedmorteza Sadat, Otmar Hilliges, and Romann M Weber. Eliminating oversaturation and artifacts of high guidance scales in diffusion models. In *ICLR*, 2024. 4

[59] 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*, 2022. 2, 4

[60] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *NeurIPS*, 29, 2016. 7

[61] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, pages 815–823, 2015. 6

[62] Dazhong Shen, Guanglu Song, Zeyue Xue, Fu-Yun Wang, and Yu Liu. Rethinking the spatial inconsistency in classifier-free diffusion guidance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9370–9379, 2024. 4

[63] Jing Shi, Wei Xiong, Zhe Lin, and Hyun Joon Jung. Instantbooth: Personalized text-to-image generation without test-time finetuning. *arXiv preprint arXiv:2304.03411*, 2023. 2

[64] James Seale Smith, Yen-Chang Hsu, Lingyu Zhang, Ting Hua, Zsolt Kira, Yilin Shen, and Hongxia Jin. Continual diffusion: Continual customization of text-to-image diffusion with c-lora. *arXiv preprint arXiv:2304.06027*, 2023. 2

[65] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2020. 2

[66] Yoad Tewel, Rinon Gal, Gal Chechik, and Yuval Atzmon. Key-locked rank one editing for text-to-image personalization. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–11, 2023. 2

[67] Yael Vinker, Andrey Voynov, Daniel Cohen-Or, and Ariel Shamir. Concept decomposition for visual exploration and inspiration. *ACM TOG*, 2023. 2

[68] Andrey Voynov, Qinghao Chu, Daniel Cohen-Or, and Kfir Aberman. $p+$: Extended textual conditioning in text-to-image generation. *arXiv preprint arXiv:2303.09522*, 2023. 2

[69] Jiajun Wang, Morteza Ghahremani Boozandani, Yitong Li, Björn Ommer, and Christian Wachinger. Stable-pose: Leveraging transformers for pose-guided text-to-image generation. *NeurIPS*, 37:65670–65698, 2025. 2, 3

[70] Qixun Wang, Xu Bai, Haofan Wang, Zekui Qin, and Anthony Chen. Instantid: Zero-shot identity-preserving generation in seconds. *arXiv preprint arXiv:2401.07519*, 2024. 2, 4, 7

[71] Qixun Wang, Xu Bai, Haofan Wang, Zekui Qin, Anthony Chen, Huaxia Li, Xu Tang, and Yao Hu. Instantid: Zero-shot identity-preserving generation in seconds, 2024. 5

[72] X Wang, Siming Fu, Qihan Huang, Wanggui He, and Hao Jiang. Ms-diffusion: Multi-subject zero-shot image personalization with layout guidance. *arXiv preprint arXiv:2406.07209*, 2024. 7

[73] Yuxiang Wei, Yabo Zhang, Zhilong Ji, Jinfeng Bai, Lei Zhang, and Wangmeng Zuo. Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. *arXiv preprint arXiv:2302.13848*, 2023. 2

[74] Guangxuan Xiao, Tianwei Yin, William T Freeman, Frédo Durand, and Song Han. Fastcomposer: Tuning-free multi-subject image generation with localized attention. *arXiv preprint arXiv:2305.10431*, 2023. 2, 3, 6

[75] Guangxuan Xiao, Tianwei Yin, William T Freeman, Frédo Durand, and Song Han. Fastcomposer: Tuning-free multi-subject image generation with localized attention. *International Journal of Computer Vision*, pages 1–20, 2024. 7

[76] Yuxuan Yan, Chi Zhang, Rui Wang, Yichao Zhou, Gege Zhang, Pei Cheng, Gang Yu, and Bin Fu. Facestudio: Put your face everywhere in seconds. *arXiv preprint arXiv:2312.02663*, 2023. 2

[77] Zhendong Yang, Ailing Zeng, Chun Yuan, and Yu Li. Effective whole-body pose estimation with two-stages distillation. In *CVPR*, 2023. 3

[78] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. 2023. 2, 7

[79] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *CVPR*, 2023. 2, 3, 4, 6, 7, 8

[80] Xu-Lu Zhang, Xiao-Yong Wei, Jin-Lin Wu, Tian-Yi Zhang, Zhao-Xiang Zhang, Zhen Lei, and Qing Li. Compositional inversion for stable diffusion models. *arXiv preprint arXiv:2312.08048*, 2023. 2

[81] Yimeng Zhang, Tiancheng Zhi, Jing Liu, Shen Sang, Liming Jiang, Qing Yan, Sijia Liu, and Linjie Luo. Id-patch: Robust id association for group photo personalization. *arXiv preprint arXiv:2411.13632*, 2024. 2, 3

[82] Ruoyu Zhao, Mingrui Zhu, Shiyin Dong, Nannan Wang, and Xinbo Gao. Catversion: Concatenating embeddings for diffusion-based text-to-image personalization. *arXiv preprint arXiv:2311.14631*, 2023. 2

[83] Yufan Zhou, Ruiyi Zhang, Jiuxiang Gu, and Tong Sun. Customization assistant for text-to-image generation. *arXiv preprint arXiv:2312.03045*, 2023. 2

[84] Yufan Zhou, Ruiyi Zhang, Tong Sun, and Jinhui Xu. Enhancing detail preservation for customized text-to-image generation: A regularization-free approach. *arXiv preprint arXiv:2305.13579*, 2023. 2

[85] Zhengguang Zhou, Jing Li, Huaxia Li, Nemo Chen, and Xu Tang. Storymaker: Towards holistic consistent characters in text-to-image generation. *arXiv preprint arXiv:2409.12576*, 2024. 2, 3

[86] Shenhao Zhu, Junming Leo Chen, Zuozhuo Dai, Qingkun Su, Yinghui Xu, Xun Cao, Yao Yao, Hao Zhu, and Siyu Zhu. Champ: Controllable and consistent human image animation with 3d parametric guidance. *arXiv preprint arXiv:2403.14781*, 2024. 2, 3