

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

Gwanghyun Kim^{1*}, Suh Yoon Jeon^{1*}, Seunggyu Lee¹, Se Young Chun^{1,2†}

¹Dept. of Electrical and Computer Engineering, ²INMC & IPAI

Seoul National University, Republic of Korea

{gwang.kim, euniejeon, leeseunggyu, sychun}@snu.ac.kr

A. Additional Results

A.1. Personalized Multi-Human Scene Generation

Additional Qualitative Comparison. As shown in Fig. S1 and S2, our proposed method demonstrates significant advantages over existing approaches. Notably, methods like InstantID+OMG [S28, S14] and IPAdapter+OMG [S33, S14], which rely on 2D skeleton-based pose conditioning, exhibit severe anatomical inaccuracies in challenging scenarios (highlighted by yellow arrows). These issues stem from the inherent limitations of 2D pose representations, which struggle to handle overlapping body parts and intricate interactions effectively. Blue arrows highlight cases where body shape preservation fails, where our approach maintains accurate body structure and pose fidelity while delivering superior performance in both face identity preservation and body shape consistency.

DreamBooth [S25], on the other hand, suffers even more pronounced issues due to its lack of pose guidance. This leads to severe anatomical distortions and, in some cases, the complete omission of individuals in multi-person scenes. Additionally, DreamBooth struggles with clothing-body shape displacement, where clothing styles are directly transferred without adapting to the individual’s body shape.

These findings further underscore the robustness and versatility of PersonaCraft, making it a state-of-the-art solution for personalized image generation in complex, real-world scenarios.

Comparison with Additional Baselines. We compared PersonaCraft against other baselines, including UniPortrait [S12], MS-Diffusion [S29], and FastComposer [S31]. While these methods share similar capabilities, they are not fully suited for our benchmark, making direct comparisons challenging. As shown in (Fig. S3), yellow arrows highlight anatomical inconsistencies in complex poses and occluded scenarios due to reliance on 2D pose representations or the

absence of pose control. PersonaCraft, in contrast, generates anatomically accurate and natural images under these conditions.

Additionally, MS-Diffusion copies clothing directly from full-body references without proper displacement. PersonaCraft integrates personalized body shapes and clothing displacement, maintaining consistency and realism.

These results highlight PersonaCraft’s superiority in generating accurate, identity-consistent images and handling occlusions and diverse poses with exceptional naturalness and customization.

A.2. Pose-Controlled Multi-Human Scene Generation

We assess pose-controlled multi-human scene generation by comparing PersonaCraft with existing methods in both single-human and multi-human contexts. As shown in Fig. S4, S5, and S6, our proposed method demonstrates significant advantages over existing approaches. Notably, methods that rely on 2D pose conditioning (e.g., T2IAdapter-SDXL[S18], ControlNet-SDXL [S34, S23]) struggle with occlusion handling, resulting in misaligned poses and distorted structures. ControlNet-Flux [S16, S34] shows instability: at a lower conditioning scale (0.8), it fails to preserve anatomical coherence, while at a higher scale (1.0), pose accuracy improves but image fidelity decreases. In contrast, PersonaCraft successfully balances pose accuracy, body realism, and image fidelity, highlighting the effectiveness of our 3D-aware pose conditioning.

A.3. PersonaCraft with Stylization

The proposed method is a plug-and-play approach, making it compatible with various style-specific LoRAs. To evaluate its effectiveness, we conducted experiments combining PersonaCraft with diverse style LoRAs, including Crayon [S21], Pastel [S5], 3D Render [S11], Pixel

*Authors contributed equally. †Corresponding author.

Art [S19], Illustration [S4], Frosting Lane [S2], Pokémon Trainer [S27], JoJo [S20], Graphic Novel [S9], and Cartoon [S3]. The results, shown in Fig. S7, highlight the method’s ability to adapt to different styles effectively. Notably, styles such as Pastel, Illustration, JoJo, and Pokémon Trainer introduce changes in facial and body characteristics, occasionally altering perceived identity, due to their bias. Nevertheless, the outcomes remain visually compelling and demonstrate the versatility of our approach.

A.4. Versatility of SCNet

To demonstrate the versatility of SCNet, we present results combining SCNet with various face identity personalization models, including InstantID [S28], PhotoMaker V2 [S17], and IPAdapter-Face [S33]. As shown in Fig. S8, SCNet enables robust body shape personalization and pose control when paired with these face models, achieving comprehensive full-body personalization and user-defined body shape adjustments. Notably, face personalization varies slightly depending on the chosen face module.

A.5. Effectiveness of Dual-Pathway Body Shape Personalization

Our dual-pathway design combining SMPLx and text improves body shape consistency in challenging cases, as shown in Fig. S9. A user study with 600 responses from 60 users confirmed its effectiveness: 76.17% preferred our method over SMPLx only (17%) and text-only (6.83%).

A.6. Robustness to 3D Conditioning Errors

Trained on SMPLx poses from MultiHMR [S8], our model preserves fidelity under pose perturbations (Fig. S10 left), supported by the robustness of the diffusion prior. Moreover, it reliably infers plausible occlusion boundaries even with misaligned or noisy depth inputs (Fig. S10 right). Shape inaccuracies are further alleviated by our dual-pathway body shape representation.

A.7. SMPLx vs. SMPL Comparison

SMPLx offers better representation of hand and face poses, as shown in the Fig. S11, providing better pose consistency and controllability compared to SMPL.

A.8. Ablation Study on Conditioning Scale

We analyze the effect of the conditioning scales of IdentityNet and SCNet on identity preservation when provided with face references and reference body shapes (SMPLx depth). As shown in Fig. S12, when the conditioning scale is set to 0 for both modules, the generated face and body shapes differ significantly from the reference. This indicates insufficient guidance from the reference inputs.

As the conditioning scales for IdentityNet and SCNet increase, the generated images progressively resemble the

reference face and body shape. This improvement demonstrates the critical role of conditioning strength in aligning the generated outputs with the given references. Optimal conditioning scales enable PersonaCraft to faithfully preserve both facial and body shape identities, ensuring high-quality personalization and consistency.

A.9. Ablation Study on Body Shape Parameters

For full-body personalized image generation, we extract the body shape parameters of the character to be personalized and use them for SMPLx rendering, which serves as the conditions for SCNet. In Tab. S1, we analyze the impact of incorporating the body shape parameters in this process. Using body shape parameters enhances body shape preservation during personalization. This indicates that leveraging the body shape parameters enables the generation of personalized images that more accurately reflect the character’s true physique.

Table S1. Evaluation of body shape preservation with and without the use of the body shape parameter.

	Single	Multi	Total
w/o body shape	0.615	0.520	0.539
w/ body shape	0.630	0.548	0.615

A.10. Ablation Study on Occlusion-aware 3D Pose & Shape Conditioning

Comparison of 3D Pose Representations for SCNet. In Fig. S13, we compare different combinations of SMPLx rendering-depth, normal, and RGB rendering-as conditioning inputs for SCNet. Using both depth and normal enables the model to leverage occlusion cues from depth and surface orientation information from normal. This leads to improved generation performance in occluded or complex body regions compared to using depth alone. However, incorporating RGB rendering in addition to depth and normal degrades image quality due to the use of multiple ControlNets for the same region. Therefore, we adopt the combination of depth and normal as base conditioning for SCNet. **Effectiveness of OccNet and OccCFG.** Also, the effect of our occlusion-aware 3D pose & shape conditioning components is analyzed in Fig. S13. Using only 2D pose leads to structural inconsistencies, while relying solely on 3D pose (SCNet) struggles with fine-grained occluded regions. Our full model (SCNet + OccNet + OccCFG) effectively preserves pose structure while handling occlusions.

In Fig. S14, we present a detailed analysis of OccCFG. We observe that increasing classifier-free guidance (CFG) in occluded regions improves anatomical consistency by leveraging stronger 3D pose information, effectively resolving local ambiguities. However, uniformly high CFG strength leads to over-saturation in non-occluded areas. We

Table S2. Additional comparison of baseline models fine-tuned on our training dataset (MPII). 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. (*: fine-tuned on our training dataset.)

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*	0.418	0.220	0.252	0.563	0.427	0.448	85.624	0.333	0.264	3.809	0.0996
IPAdapter+ OMG*	0.204	0.155	0.162	0.606	0.45	0.472	84.893	0.362	0.267	3.736	0.0984
IPA-Face + OMG*	0.350	0.181	0.207	0.568	0.429	0.451	86.373	0.355	0.265	3.930	0.0974
Ours	0.421	0.298	0.317	0.630	0.548	0.560	60.654	0.506	0.273	4.238	0.0931

found that applying CFG only in the human segmentation region also results in high CFG, whereas our OccCFG avoids issues in unoccluded regions while maintaining effective guidance in occluded areas.

A.11. Efficiency Analysis

As detailed in Table S3, we compared the inference times for multi-identity personalized image synthesis across various methods, specifically when generating images with three distinct identities. Optimization-based personalization methods like Textual Inversion [S10] and DreamBooth [S25] require a batch size of 4 and 500 optimization steps per identity. This lengthy process results in significantly extended inference times, rendering these approaches highly inefficient for real-time applications.

In contrast, PersonaCraft offers superior efficiency. While methods based on OMG [S14] involve a two-stage generation process, PersonaCraft completes inference in less than half the time (17.25s vs. 46.94s for OMG), despite comparable VRAM usage (22.51GB for Ours vs. 20.85GB for OMG). PersonaCraft’s modular design provides a significant advantage by eliminating the need to retrain priors or LoRAs, a requirement for methods like Textual Inversion (which takes 1636.15s and uses 11.00GB). This efficiency, coupled with its ability to generate high-quality personalized images, is particularly beneficial when synthesizing multiple identities in a single pass.

Table S3. Inference times for multi-identity personalized synthesis.

Method	Total Time (secs)
Text Inversion	1636.15
DreamBooth	770.713
InstantID + OMG	46.94
IPAdapter + OMG	44.62
IPA-Face + OMG	35.46
PersonaCraft (ours)	17.25

A.12. Additional Results with Baselines Fine-tuned on Our Training Dataset

We further compare PersonaCraft with key baselines fine-tuned on our training dataset: InstantID + OMG, IPAdapter + OMG, and IPA-Face + OMG for personalized multi-human scene generation, and ControlNet-SDXL for pose-controlled multi-human scene generation.

Personalized Multi-Human Scene Generation. Although the baselines are fine-tuned on our training dataset, Tab. S2 shows that 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. Additionally, PersonaCraft surpasses baselines in IS and KID, demonstrating enhanced perceptual quality and text-image coherence.

Pose-Controlled Multi-Human Scene Generation. Although the baselines are fine-tuned on our training dataset, Tab. S4 shows that PersonaCraft achieves the lowest MPJPE (3D) and highest AP-0.5 (2D), demonstrating superior pose alignment and keypoint localization.

Table S4. Additional comparison of the baseline fine-tuned on our training dataset (MPII). Quantitative evaluation of pose-controlled human generation in terms of pose accuracy, text alignment, and image quality. (*: fine-tuned on our training dataset.)

Pose-Controlled Human Generation	Pose		Text	Image quality	
	MPJPE (3D) ↓	AP-0.5 (2D) ↑	CLIP sim ↑	IS ↑	KID ↓
ControlNet-SDXL*	100.817	0.313	0.281	3.407	0.122
Ours	62.647	0.495	0.274	4.114	0.091

A.13. Failure Cases

While our method is versatile and can be applied to other ControlNet models, the performance of our face personalization depends significantly on the underlying face network. Additionally, the accuracy of 3D human model fitting is dependent on the performance of the fitting algorithm used. Variations in the quality of the fitting process may impact the final output, especially in cases where the reference data is incomplete or inaccurate as presented in Fig. S15. (Additional Experimental details are continued in Sec. B)

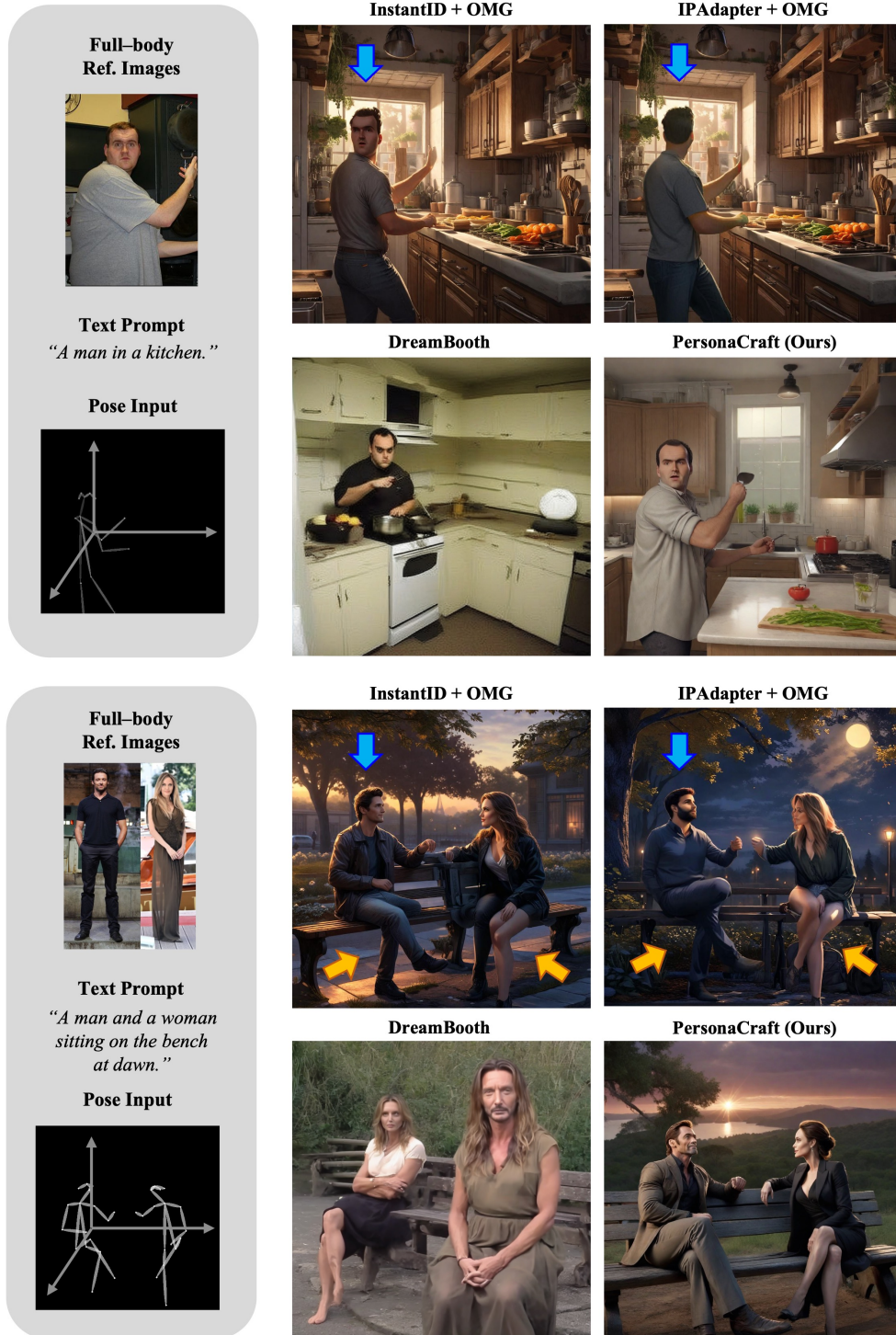


Figure S1. Additional comparison of generated images. **Yellow** arrows highlight anatomical issues due to 2D pose limitations. **Blue** arrows refer to the individuals evaluated for correct body shape preservation. PersonaCraft excels in identity, body shape consistency, and naturalness while being over twice as fast as OMG-based methods [S28, S33, S14].

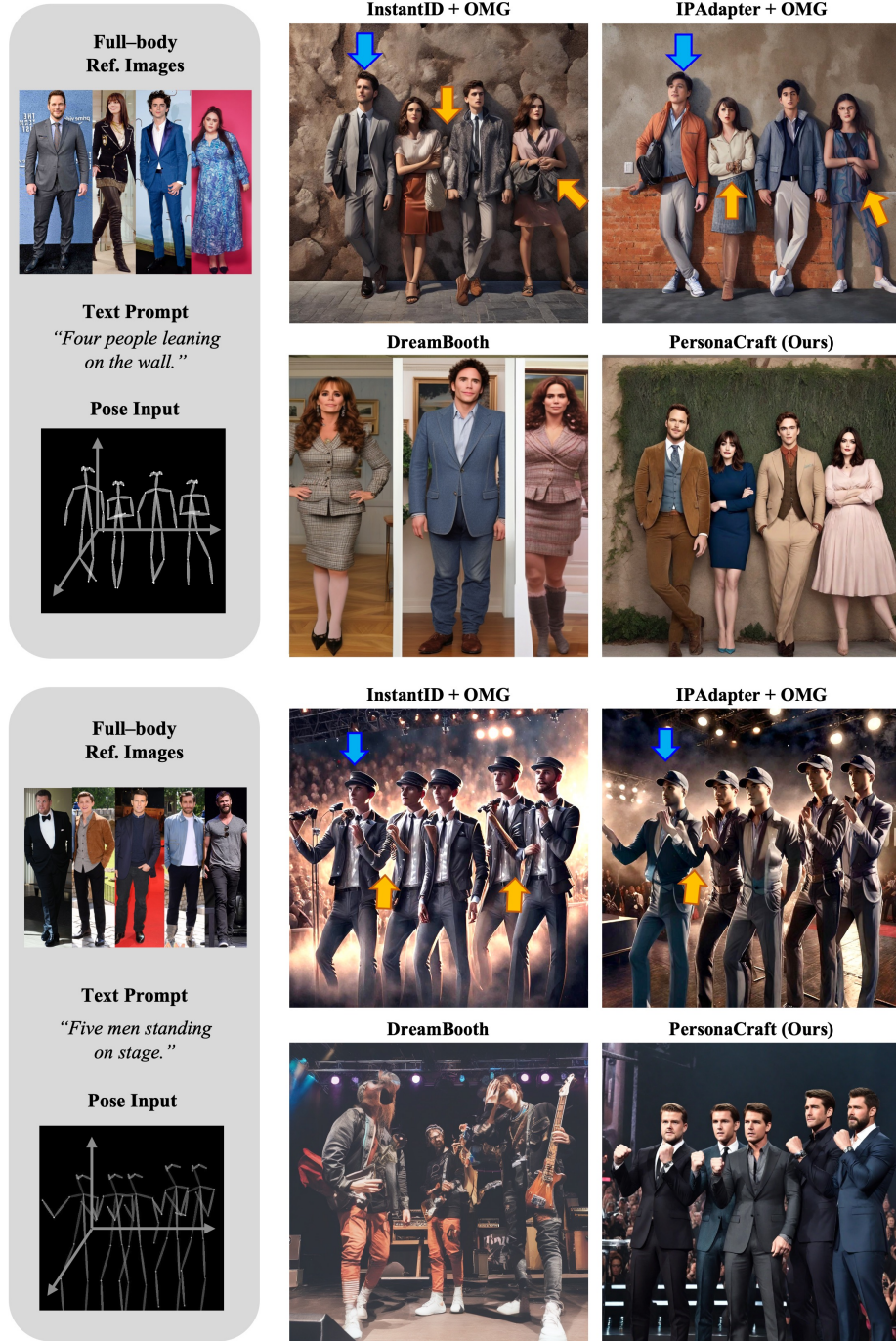


Figure S2. Additional comparison of generated images. **Yellow** arrows highlight anatomical issues in InstantID+OMG [S28, S14] and IPAdapter+OMG [S33, S14] due to 2D pose limitations. DreamBooth [S25] shows severe distortions and clothing mismatches. PersonaCraft excels in face identity, body shape consistency, and naturalness while being over twice as fast as OMG-based methods [S28, S33, S14].

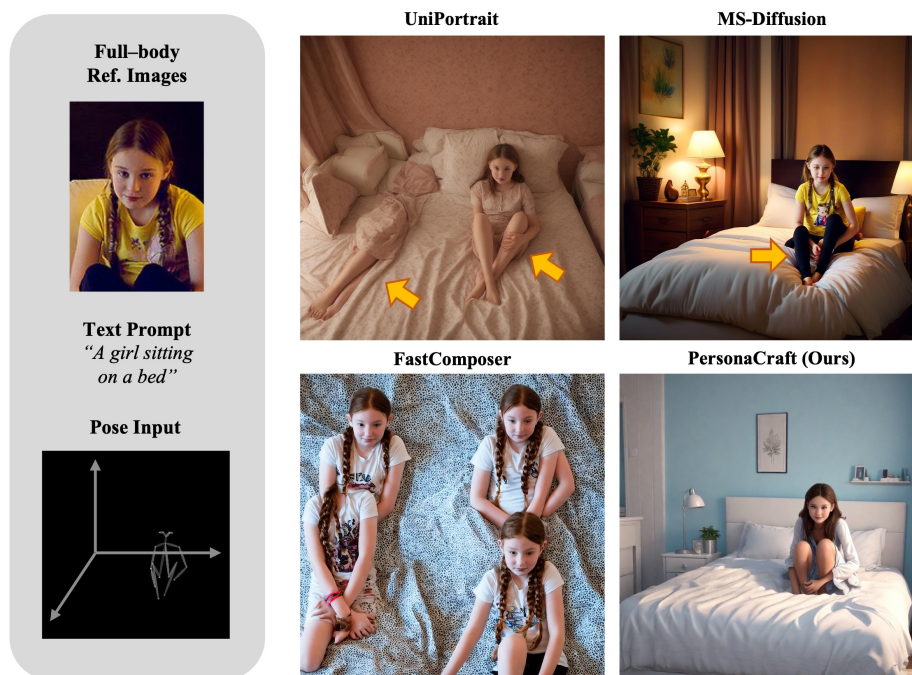


Figure S3. Comparison of PersonaCraft with UniPortrait [S12], MS-Diffusion [S29], and FastComposer [S31]. Yellow arrows show anatomical inconsistencies in poses and occlusions. MS-Diffusion copies clothing without proper adjustment. PersonaCraft excels in face identity, body shape consistency, and naturalness compared to the baselines.

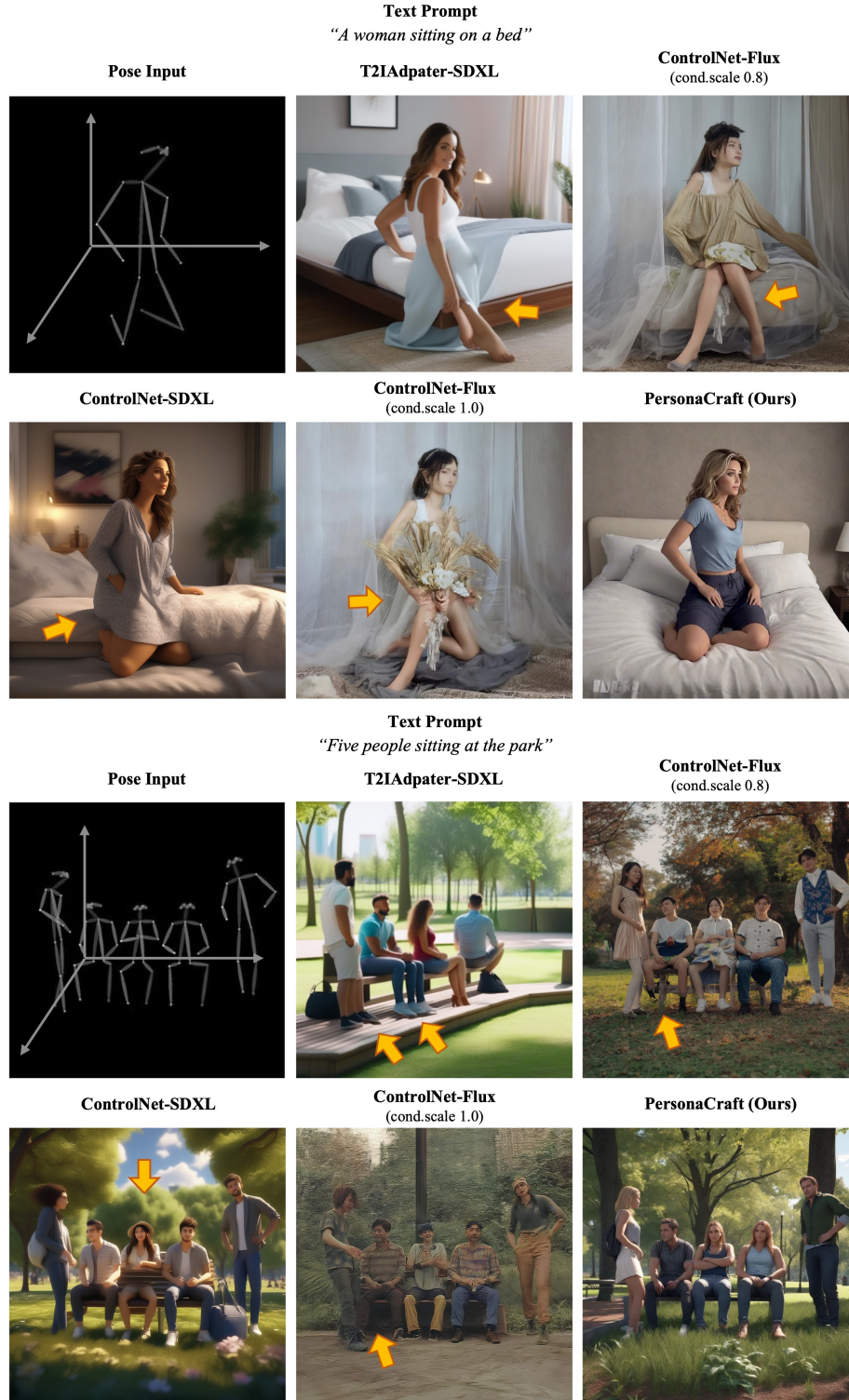


Figure S4. Qualitative comparison of 3D-aware pose conditioning for multi-human generation, covering both single and multi-human scenarios. PersonaCraft achieves superior alignment with the input pose while effectively handling occlusions, allowing for natural human anatomy to be maintained even in complex multi-human interactions. Our method outperforms baselines in preserving identity, body shape, and overall human realism. Yellow arrows highlight unnatural anatomical structures.

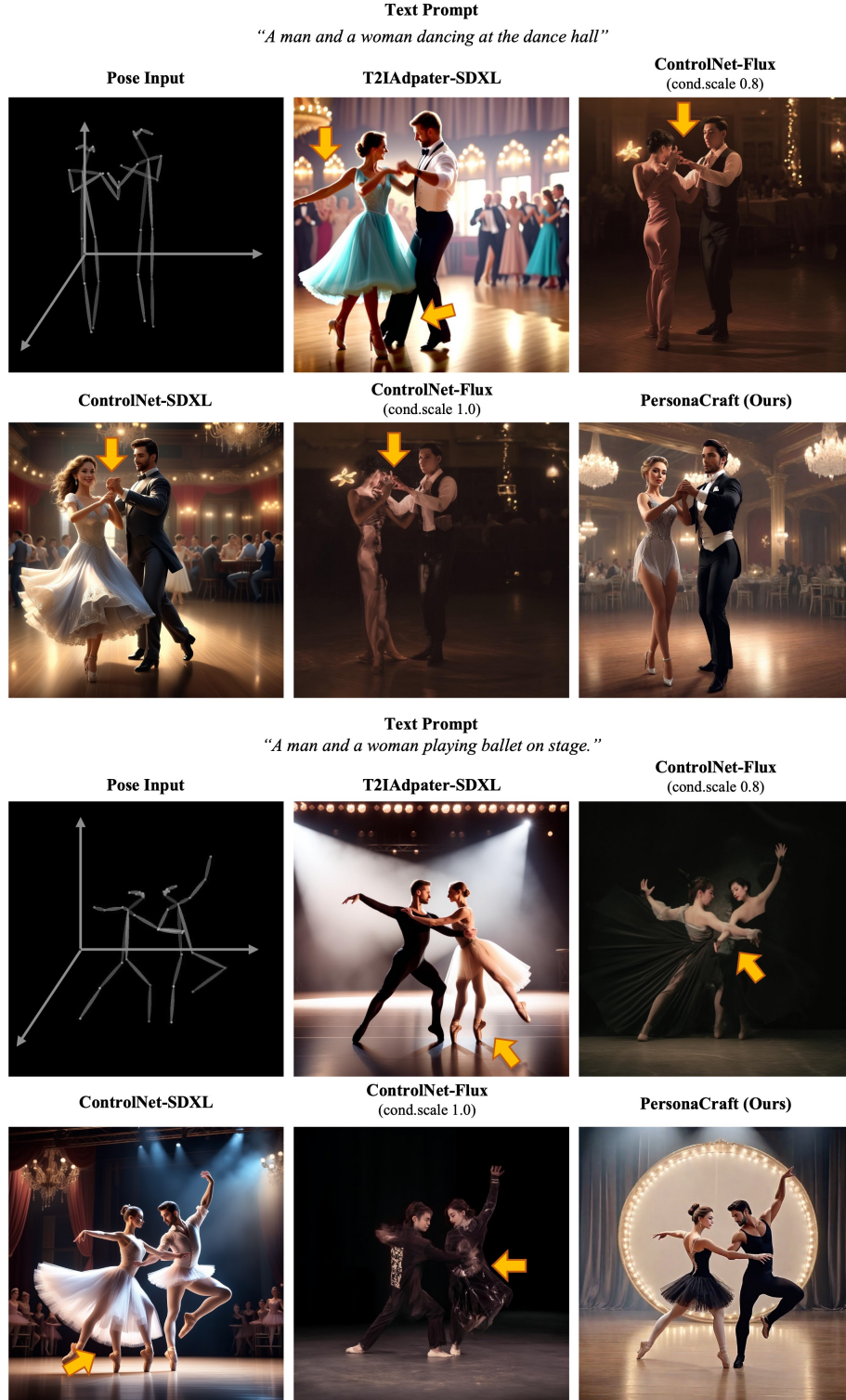


Figure S5. Qualitative comparison of 3D-aware pose conditioning for multi-human generation, covering both single and multi-human scenarios. PersonaCraft achieves superior alignment with the input pose while effectively handling occlusions, allowing for natural human anatomy to be maintained even in complex multi-human interactions. Our method outperforms baselines in preserving identity, body shape, and overall human realism. Yellow arrows highlight unnatural anatomical structures.

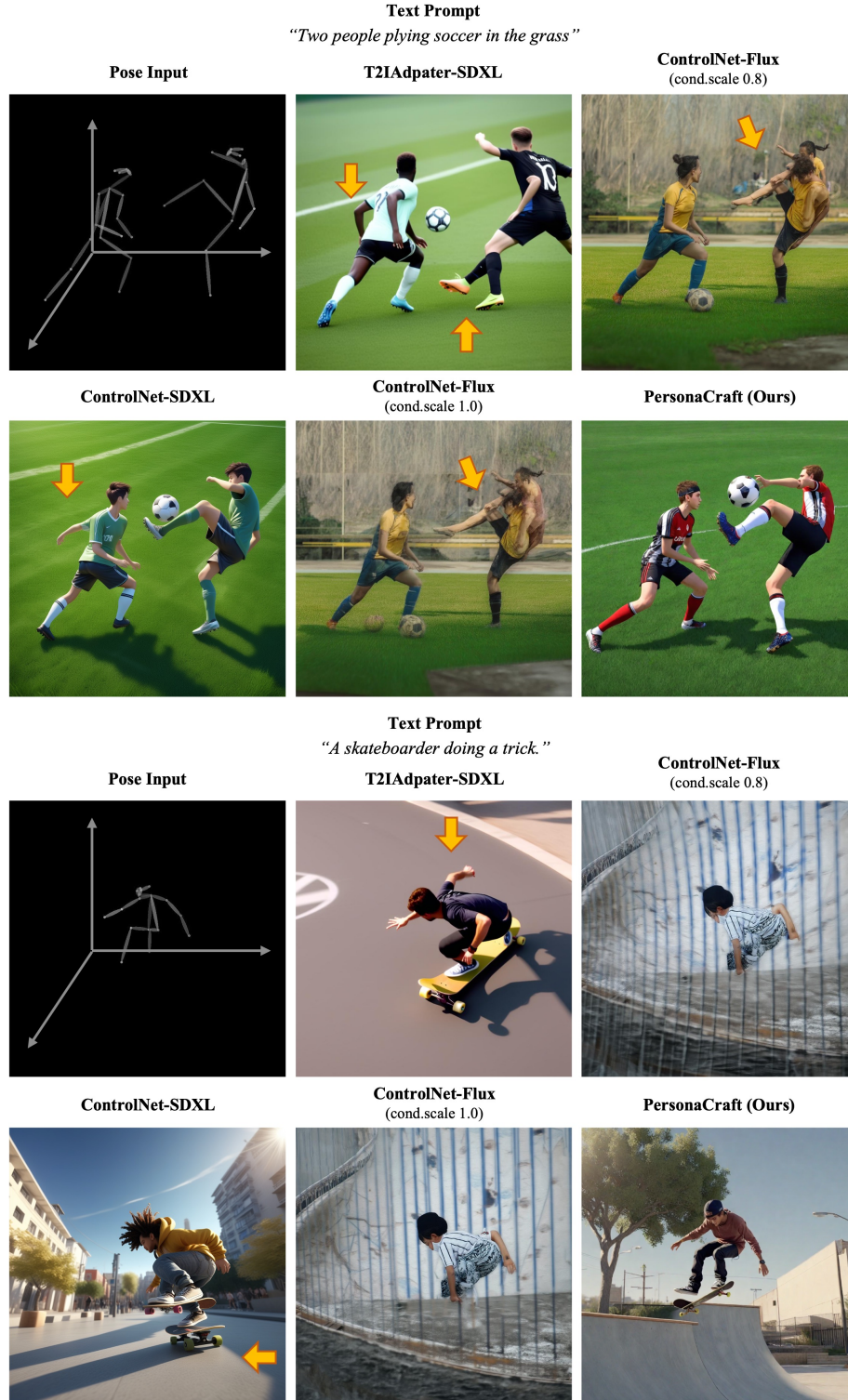


Figure S6. Qualitative comparison of 3D-aware pose conditioning for multi-human generation, covering both single and multi-human scenarios. PersonaCraft achieves superior alignment with the input pose while effectively handling occlusions, allowing for natural human anatomy to be maintained even in complex multi-human interactions. Our method outperforms baselines in preserving identity, body shape, and overall human realism. Yellow arrows highlight unnatural anatomical structures.

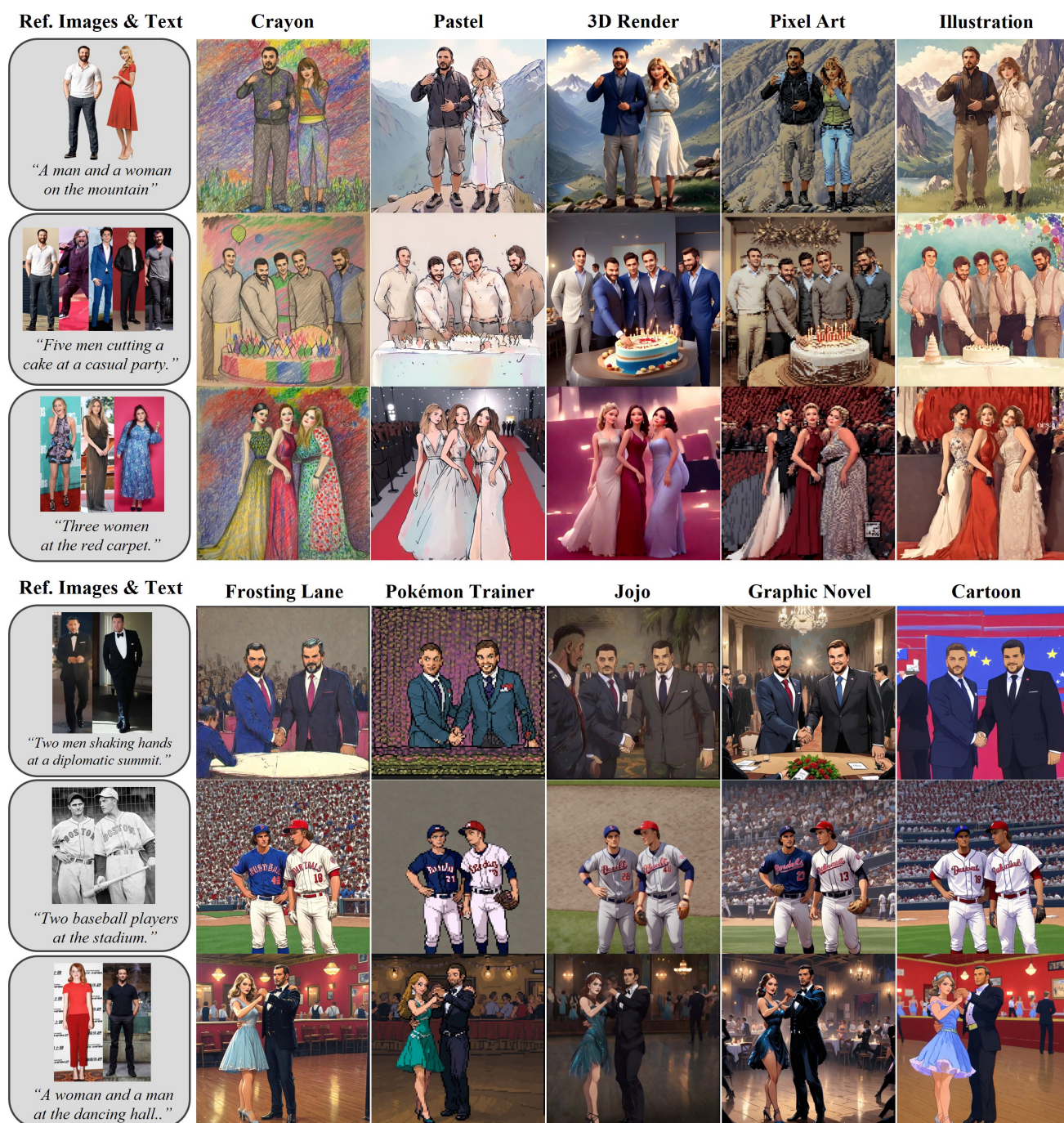


Figure S7. Results of combining PersonaCraft with various style LoRAs, showcasing adaptability to styles like Pastel, JoJo, and Pokémon Trainer. Some styles alter facial and body identities due to their bias, while producing visually impressive outcomes.



Figure S8. Integration of SCNet with face personalization models like InstantID [S28], PhotoMaker V2 [S17] and IPAdapter-Face [S33] achieves robust full-body customization, with slight variations by face module.

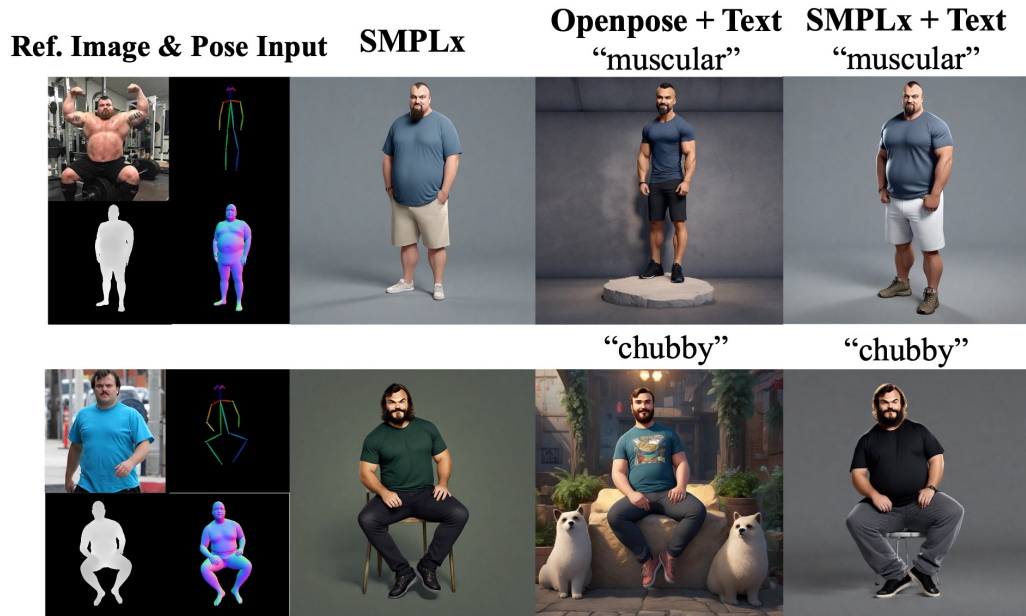


Figure S9. Effectiveness of dual-path body shape control. By combining SMPLx and text, our dual-pathway approach effectively enhances body shape consistency, especially in difficult cases.



Figure S10. Robustness to SMPLx, depth errors. Our model handles pose perturbations (left) and infers occlusions from noisy depth (right), with dual-pathway shape control further improving results.

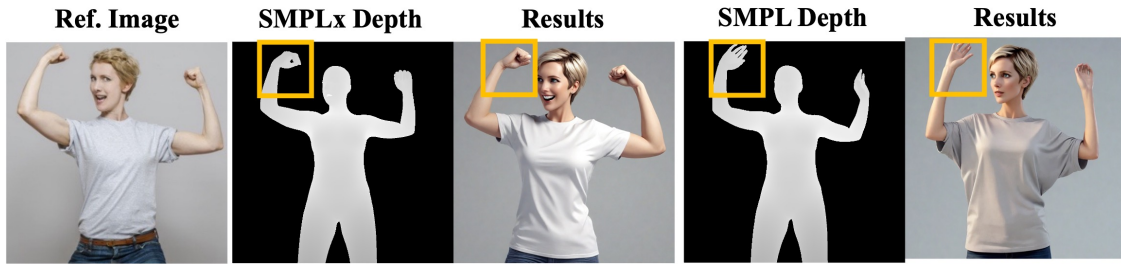


Figure S11. SMPLx control vs SMPL control. With its more detailed representation of hand and face poses, SMPLx offers greater consistency and controllability over SMPL.

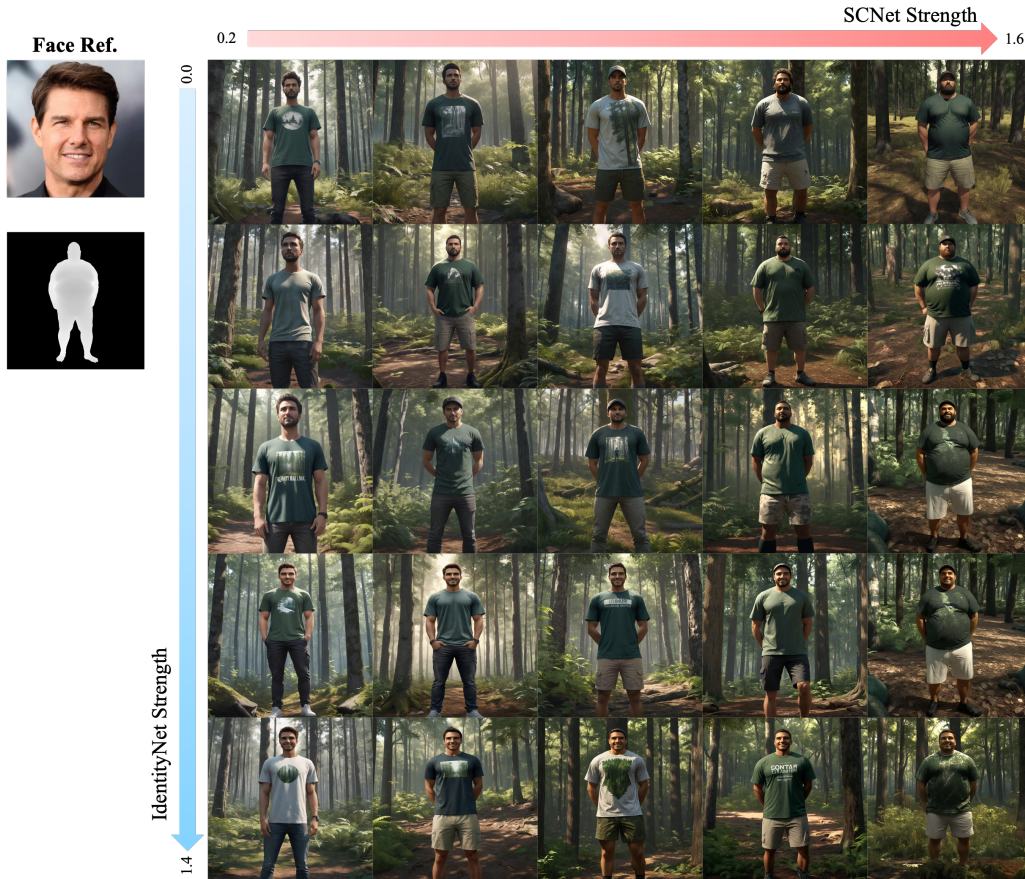


Figure S12. Ablation study on the conditioning scales of IdentityNet and SCNet, demonstrating improved identity preservation for face and body shape as the conditioning scales increase.



Figure S13. Ablation study on occlusion-aware 3D pose & shape conditioning. The combination of depth and normal as conditioning inputs for SCNet achieves the best generation performance in occluded or complex regions. While using SCNet faces issues preserving pose structure in fine-grained occluded regions, adding OccNet and OccCFG effectively addresses these problems.



Figure S14. Effect of our occlusion-aware CFG (OccCFG). Each label refers to the CFG scale applied to specific regions. "Base=3" means the CFG scale is 3 for all regions, while "Base=3, Seg=5" indicates that the CFG scale is 5 for the human segmentation region. "Occ" refers to the occlusion mask region.

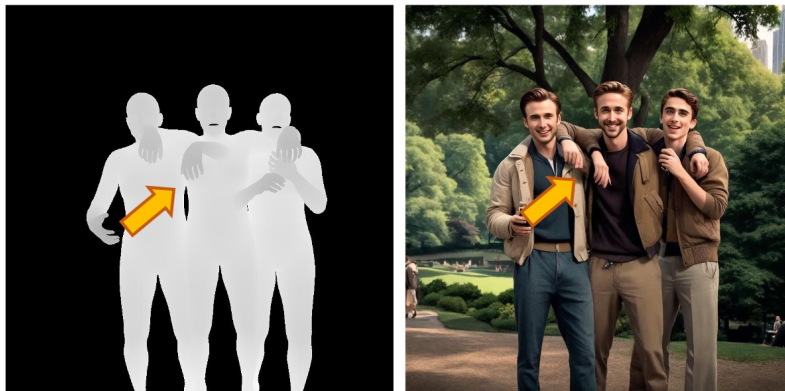


Figure S15. Failure cases. The accuracy of 3D human model fitting depends on the fitting algorithm used, and variations in fitting quality, particularly when the reference data is incomplete or inaccurate, can impact the final output.

B. Additional Experimental Details

B.1. Additional Implementation Details

Additional Details on Training Dataset. To account for occlusion scenarios, we balance the MPII [S6] dataset with a 2:1 ratio of single-person to multi-person images. Depth clipping is applied during depth rendering to retain only values below 5, ensuring consistent quality. After preprocessing, we curate a final dataset of 6,348 image-text-SMPLx-parameter pairs. This carefully curated dataset enables robust model training with diverse 3D human poses, complex interactions, and detailed human parameters such as body shape and pose conditioning.

Details on Training of SCNet and OccNet. As illustrated

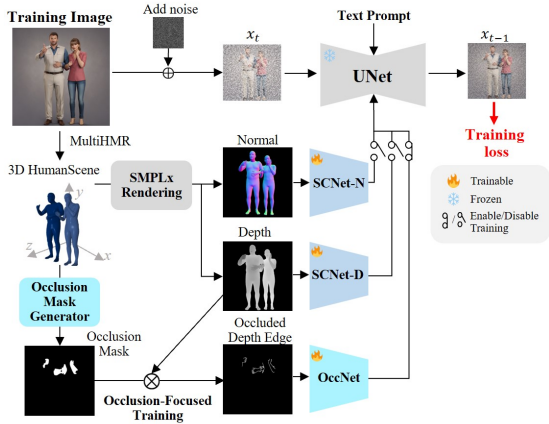


Figure S16. Training of SMPLx-ControlNet (SCNet) and Occlusion Boundary Enhancer Network (OccNet).

In Fig. S16, the networks are trained separately, with SMPLx depth, normal maps, and occlusion masks extracted from training images. The pretrained ControlNet [S34] is fine-tuned with these 3D pose representations.

We base our SCNet on controlnet-union-sd-xl-1.0 [S32] and fine-tune it for SMPLx [S22] depth-normal-occluded edge conditioning. The architecture supports more than 10 control types for high-resolution text-to-image generation, with depth selected as the control type in our implementation.

We utilize 3D poses represented by SMPLx parameters, which include 55 joints (22 body, 1 jaw, 2 eyes, and 30 hands) along with camera parameters (intrinsic and extrinsic). These parameters generate a vertex- and face-based mesh that we render as SMPLx depth maps.

For depth edge extraction, we employed the Canny edge detector for more robust edge extraction instead of thresholding the spatial partial gradient of depth with τ , using a low threshold of 5 and a high threshold of 15.

Details on Full-Body Personalized Image Synthesis. We adopt MultiHMR [S8] as our SMPLx fitting method. MultiHMR is a single-shot model that reconstructs 3D human meshes from a single RGB image, leveraging the SMPLx

parametric model to predict full-body meshes, including hands and facial expressions, with 3D localization in the camera coordinate system. The body shape parameters, β , are represented as 10-dimensional vectors, each scaled by orthonormal shape displacement components.

For facial identity processing, we employ the antelopev2 facial detection and recognition models from InsightFace [S1] to extract 512-dimensional face identity embeddings, f , from human images.

To enhance the visual quality of human-centric scenes, we utilize the YamerMIX-v8 variant of SDXL. For face identity personalization, we incorporate IdentityNet from InstantID [S28], which enables instant, zero-shot, identity-preserving image generation. IdentityNet enforces strong semantic and weak spatial conditions by integrating facial and landmark images with textual prompts to guide the generation process.

Following InstantID, we use five key facial landmarks (two for the eyes, one for the nose, and two for the mouth) as spatial control signals, providing a more generalized constraint than detailed key points.

Details on Dual-Pathway Body Shape Personalization. In this method, a CLIP [S24]-based classifier is employed to extract body shape attributes in the form of text descriptions. These descriptions categorize the body type into various categories, such as "overweight," "muscular," "fat," etc. This is achieved by using a combination of CLIP, which bridges the gap between vision and text, and specific regional prompting techniques.

The body shape information is then used in Regional Diffusion, a concept derived from MultiDiffusion [S7], where each diffusion timestep involves conditioning on both the body pose and a corresponding text description about the individual's body shape. The process operates on each person's instance throughout multiple diffusion timesteps, ensuring that the shape-specific features are incorporated and aligned with the pose dynamics.

Incorporating these shape attributes into the diffusion process allows the model to better represent personalized body shapes in the generation process, resulting in a more accurate and detailed synthesis of human shapes. This integration is achieved through the use of regional prompting, where at each timestep, the model is conditioned on both the body pose and the specific body shape description to refine the body shape in the generated instance. This process is further integrated with SCNet, a network that helps guide the shape refinement and personalization.

User-Defined Body Shape Control. PersonaCraft enables user-defined body shape control, allowing adjustments based on user preferences: 1) *Reference-based control*: The target body shape parameter, β_{target} , is obtained from a reference image via SMPLx fitting. 2) *Interpolation/extrapolation-based control*: Given two ref-

erence body shape parameters, β_1 and β_2 , the target shape is computed as $\beta_{\text{target}} = \gamma\beta_1 + (1 - \gamma)\beta_2$, where γ controls the interpolation/extrapolation ratio. The resulting β_{target} .

B.2. Details on Metrics

Face identity preservation was measured for 1 ~ 5 identities following FastComposer [S30], using FaceNet [S26] for identity similarity within the face mask. The identity similarity score is computed by averaging the non-negative cosine similarity over both the number of humans and the total number of images:

$$S_{\text{face}} = \frac{1}{N_{\text{image}}} \sum_{i=1}^{N_{\text{image}}} \frac{1}{N_{\text{human},i}} \sum_{j=1}^{N_{\text{human},i}} \max \left(0, \frac{\mathbf{f}_{\text{ref},i}^{(j)} \cdot \mathbf{f}_{\text{gen},i}^{(j)}}{\|\mathbf{f}_{\text{ref},i}^{(j)}\| \|\mathbf{f}_{\text{gen},i}^{(j)}\|} \right) \quad (\text{S1})$$

where $\mathbf{f}_{\text{ref},i}^{(j)}$ and $\mathbf{f}_{\text{gen},i}^{(j)}$ are the face embeddings for the j -th reference and generated identity in the i -th image, respectively. N_{image} is the total number of images, and $N_{\text{human},i}$ is the number of humans in the i -th image.

Body shape preservation was evaluated using cosine similarity between the SMPLx body shape parameters β from the reference and generated instances. The score is averaged over both the number of humans and the total number of test images:

$$S_{\text{body}} = \frac{1}{N_{\text{image}}} \sum_{i=1}^{N_{\text{image}}} \frac{1}{N_{\text{human},i}} \sum_{j=1}^{N_{\text{human},i}} \frac{\beta_{\text{ref},i}^{(j)} \cdot \beta_{\text{gen},i}^{(j)}}{\|\beta_{\text{ref},i}^{(j)}\| \|\beta_{\text{gen},i}^{(j)}\|} \quad (\text{S2})$$

where $\beta_{\text{ref},i}^{(j)}$ and $\beta_{\text{gen},i}^{(j)}$ are the body shape parameters for the j -th reference and generated instance in the i -th image, respectively. N_{image} is the total number of images, and $N_{\text{human},i}$ is the number of humans in the i -th image.

CLIP similarity was measured using the CLIP-L/14 model for image-text alignment. Cosine similarity was used to evaluate the alignment between the generated image and the textual description. The CLIP encoders $\mathcal{E}_{\text{image}}$ and $\mathcal{E}_{\text{text}}$ were used for the image and text embeddings, respectively. The alignment score is averaged over all test images:

$$S_{\text{CLIP}} = \frac{1}{N_{\text{image}}} \sum_{i=1}^{N_{\text{image}}} \frac{\mathcal{E}_{\text{image}}(\mathbf{I}_{\text{gen},i}) \cdot \mathcal{E}_{\text{text}}(\mathbf{y}_{\text{ref},i})}{\|\mathcal{E}_{\text{image}}(\mathbf{I}_{\text{gen},i})\| \|\mathcal{E}_{\text{text}}(\mathbf{y}_{\text{ref},i})\|} \quad (\text{S3})$$

where $\mathcal{E}_{\text{image}}(\mathbf{I}_{\text{gen},i})$ is the generated image embedding for the i -th image, and $\mathcal{E}_{\text{text}}(\mathbf{y}_{\text{ref},i})$ is the reference text embedding.

B.3. Details on Baselines

To evaluate PersonaCraft, we compared it with several baselines for single-shot, multi-identity, and pose-controllable image synthesis, all implemented using SDXL [S23]. Key

baselines include OMG [S14] for multi-concept personalization, InstantID/IPAdapter [S28, S33] for single-shot personalization, 2D pose ControlNet [S34], and optimization-based methods like DreamBooth [S25] and Texture Inversion [S10].

OMG + InstantID/IPAdapter/IPA-Face. OMG [S14] introduces a two-stage sampling solution for multi-concept personalization. The first stage handles layout generation and occlusion management, while the second stage integrates concepts using visual comprehension and noise blending. OMG can also be combined with single-concept models like InstantID without additional tuning. For OMG+InstantID, we follow the official inference code from the InstantID repository [S15]. For OMG+InstantID and OMG+IPAdapter/IPA-Face, we replace InstantID with IPAdapter and IPA-Face, respectively, to adapt the framework for different face identity modules.

Textual Inversion. In original Textual Inversion [S10], text embeddings are optimized for user-provided visual concepts, linking them to new pseudo-words that can be seamlessly incorporated into future prompts, effectively performing an inversion into the text-embedding space. To enable single-reference, multi-concept personalization, we optimize a unique text embedding $\mathcal{V}^{(i)}$ for each concept derived from a single reference image. These embeddings are paired with unique identifiers, allowing for the dynamic integration of multiple concepts into prompts during inference, facilitating multi-concept personalization.

DreamBooth. In original DreamBooth [S25], the model is fine-tuned with images and text prompts using a unique identifier. A prior preservation loss is applied to encourage class diversity. For single-reference, multi-concept personalization, we adopt DreamBooth-LoRA [S25, S13], where each reference image is associated with a unique $\mathcal{M}^{(i)}$ and identifier $\mathcal{V}^{(i)}$. These are fine-tuned based on the DreamBooth framework. During inference, both \mathcal{M} and identifiers are used simultaneously, enabling personalized, concept-specific image generation from a single reference.

B.4. Details on User study

Personalized Multi-Human Scene Generation We conducted a user study to assess the naturalness, face identity preservation, body shape preservation, and image-text correspondence of images generated by three baseline methods (one from each group) and our method. Participants ranked the top three methods based on the following criteria: 1) *Text Correspondence*: Rank the images based on how closely they align with the textual description. 2) *Face Identity Preservation*: Rank the images in order of how well they reflect the face identity of the personalized character. 3) *Body Shape Personalization*: Rank the images based on how accurately they reflect the personalized character’s body shape. 4) *Naturalness*: Rank the images in order of

the most natural-looking, considering factors such as physically impossible appearances, illogical features, inconsistencies, or lack of real-world physics and connections. The study collected a total of 18,540 responses from 103 participants across 15 cases, including both custom and COCO-Wholebody scenarios. We present illustrative example images from the user study in Fig. S17.

Pose-Controlled Multi-Human Scene Generation We conducted a user study to assess the text-image correspondence, pose consistency and naturalness of images generated by three baseline methods (one from each group) and our method. Participants ranked the top three methods based on the following criteria:

1) *Text Correspondence*: Rank the images based on how closely they align with the textual description. 2) *Pose Consistency*: Rank the images based on how well they reflect the given pose input. 3) *Naturalness*: Rank the images in order of the most natural-looking, considering factors such as physically impossible appearances, illogical features, inconsistencies, or lack of real-world physics and connections. The study collected a total of 15,390 responses from 114 participants across 15 cases, including both custom and COCO-Wholebody scenarios. We present illustrative example images from the user study in Fig. S18.

Reference Images



Text Prompt

"Two men shaking hands at a diplomatic summit"

A



D



B



E



C



F



1) Rank the images based on how closely they align with the textual description *

	A	B	C	D	E	F
1st	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
2nd	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
3rd	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

3) Rank the images based on how accurately they reflect the personalized character's body shape. *

	A	B	C	D	E	F
1st	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
2nd	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
3rd	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

2) Rank the images in order of how well they reflect the face identity of the personalized * character.

	A	B	C	D	E	F
1st	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
2nd	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
3rd	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

4) Rank the images in order of the most natural-looking, considering factors such as physically impossible appearances, illogical features, inconsistencies, or lack of real-world physics and connections. *

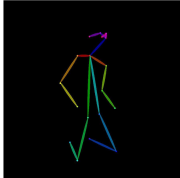
	A	B	C	D	E	F
1st	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
2nd	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
3rd	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Figure S17. Example images from the user study for personalized multi-human scene generation.


Text Prompt

"A woman sitting on a bed"


Pose Input




A



B



C



1) Rank the images based on how closely they align with the textual description. *

	A	B	C
1st	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
2nd	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
3rd	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

2) Rank the images based on how well they reflect the given pose input. *

	A	B	C
1st	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
2nd	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
3rd	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

3) Rank the images based on how natural the figures appear. Consider factors such as the correct number of limbs, proper limb connectivity, and physical plausibility. *

	A	B	C
1st	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
2nd	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
3rd	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Figure S18. Example images from the user study for pose-controlled multi-human scene generation.

References

- [S1] Insightface: 2d and 3d face analysis project. Github, <https://github.com/deepinsight/insightface>. 14
- [S2] Alvdansen. frosting-lane-lora-sd-xl. Hugging Face, <https://huggingface.co/alvdansen/frosting-lane>. 2
- [S3] Alvdansen. gemini-anime-lora-sd-xl. Hugging Face, <https://huggingface.co/alvdansen/geminianime>. 2
- [S4] Alvdansen. painting-light-lora-sd-xl. Hugging Face, <https://huggingface.co/alvdansen/paintinglight/tree/main>. 2
- [S5] Alvdansen. softpastel-anime-lora-sd-xl. Hugging Face, <https://huggingface.co/alvdansen/softpastelanime>. 1
- [S6] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *CVPR*, 2014. 14
- [S7] Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. Multidiffusion: Fusing diffusion paths for controlled image generation. *arXiv preprint arXiv:2302.08113*, 2023. 14
- [S8] Fabien Baradel*, Matthieu Armando, Salma Galaoui, 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. 2, 14
- [S9] Blink7630. graphic-novel-illustration-lora-sd-xl. Hugging Face, <https://huggingface.co/blink7630/graphic-novel-illustration>. 2
- [S10] 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. 3, 15
- [S11] GoofyAI. 3d-render-style-lora-sd-xl. Hugging Face, https://huggingface.co/goofyai/3d_render_style_xl. 1
- [S12] Junjie He, Yifeng Geng, and Liefeng Bo. Unipor-trait: A unified framework for identity-preserving single-and multi-human image personalization. *arXiv preprint arXiv:2408.05939*, 2024. 1, 6
- [S13] 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. 15
- [S14] 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. 1, 3, 4, 5, 15
- [S15] Kongzhe. Inference code for omg + instantid. GitHub, https://github.com/kongzhecn/OMG/blob/master/inference_instantid.py, 2023. Accessed: 2024-11-20. 15
- [S16] Black Forest Labs. Flux. <https://github.com/black-forest-labs/flux>, 2024. 1
- [S17] 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, 11
- [S18] 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. 1
- [S19] Nerijs. pixel-portraits-lora-sd-xl. Hugging Face, <https://huggingface.co/nerijs/pixelportraits192-XL-v1.0>. 2
- [S20] Norod78. jojo-style-lora-sd-xl. Hugging Face, <https://huggingface.co/Norod78/SDXL-JojosoStyle-Lora-v2>. 2
- [S21] Ostris. crayon-style-lora-sd-xl. Hugging Face, https://huggingface.co/ostris/crayon_style_lora_sd-xl. 1
- [S22] 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. 14
- [S23] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sd-xl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 1, 15
- [S24] 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. 14
- [S25] 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. 1, 3, 5, 15
- [S26] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, pages 815–823, 2015. 15
- [S27] sWizad. pokemon-trainer-sprite-pixelart-lora-sd-xl. Hugging Face, <https://huggingface.co/sWizad/pokemon-trainer-sprite-pixelart>. 2
- [S28] 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. 1, 2, 4, 5, 11, 14, 15
- [S29] 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. 1, 6
- [S30] 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. 15
- [S31] 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. 1, 6
- [S32] Xinsir. Controlnet-union-sd-xl-1.0. Hugging Face, <https://huggingface.co/xinsir/controlnet-union-sd-xl-1.0>, 2023. 14
- [S33] Hu Ye, Jun Zhang, Sibio Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-

- image diffusion models. 2023. [1](#), [2](#), [4](#), [5](#), [11](#), [15](#)
- [S34] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *CVPR*, 2023. [1](#), [14](#), [15](#)