This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

PersonaHOI: Effortlessly Improving Face Personalization in Human-Object Interaction Generation

Xinting Hu* xhu@mpi-inf.mpg.de
Haoran Wang* hawang@mpi-inf.mpg.de
Jan Eric Lenssen jlenssen@mpi-inf.mpg.de
Bernt Schiele schiele@mpi-inf.mpg.de
Max Planck Institute for Informatics, Saarland Informatics Campus, Germany



Figure 1. **Examples of Personalized Face with Human-Object Interaction (HOI) Generation.** We present PersonaHOI, a training- and tuning-free framework built on existing diffusion models. Using a single reference image and diverse HOI prompts, PersonaHOI generates identity-consistent human-object interactions compared to FastComposer [35]. PersonaHOI can further seamlessly integrate varied contexts, styles, accessories, and multi-person scenarios, ensuring scalability and practicality for real-world applications.

Abstract

We introduce PersonaHOI, a training- and tuning-free framework that fuses a general StableDiffusion model with a personalized face diffusion (PFD) model to generate identity-consistent human-object interaction (HOI) images. While existing PFD models have advanced significantly, they often overemphasize facial features at the expense of full-body coherence, PersonaHOI introduces an additional StableDiffusion (SD) branch guided by HOI-oriented text inputs. By incorporating cross-attention constraints in the PFD branch and spatial merging at both latent and residual levels, PersonaHOI preserves personalized facial details while ensuring interactive non-facial regions. Experiments, validated by a novel interaction alignment metric, demonstrate the superior realism and scalability of PersonaHOI, establishing a new standard for practical personalized face with HOI generation. Code is available at https://github.com/JoyHuYY1412/PersonaHOI.

*Equal Contribution.



Figure 2. **(a)** The spatial layout of StableDiffusion guides PersonaHOI to generate personalized content with coherent human-object interactions (HOI). **(b)** Analysis of identity injection timing in PFD models. We use FastComposer [35] for diffusion model generation. Injecting face representation at the start of image generation preserves facial details but lacks coherent HOI, while delayed injection continuously deviates from the original identity, resulting in random human features and meaningless human-object interactions.

1. Introduction

Personalized face generation has seen increasing public interest and demand in user-specified digital content creation. Current learning-based personalized face diffusion (PFD) models [18, 21, 29, 35], trained on large-scale face-centric datasets, can incorporate a single user-provided reference facial image into a text-to-image generation process, enabling the rapid creation of images that depict specific persons in diverse scenes, outfits, and styles within seconds.

Although these methods perform well in simpler scenarios, they often struggle with generating full-body depictions involving detailed human-object interactions (HOI). As illustrated in Figure 1, the generated images of FastComposer [35] show missing objects or body parts, resulting in portrait-focused outputs without HOI information. This limitation compromises the overall realism and practical utility of the generated content, especially in immersive or interactive applications.

To diagnose the limitations of PFD models, we examine the results of PFD models by varying the timing of identity injection during the generation process. As seen in the first and last columns of Figure 2(b), early injection effectively preserves facial details from the reference image, but relying solely on text input (last column) results in incoherent HOI outputs. This analysis highlights that the core challenge lies not in maintaining identity fidelity but in generating coherent body movements and interactions driven by text prompts. Notably, pre-trained models used for PFD, such as StableDiffusion [25], can generate satisfactory HOI images due to their extensive and diverse training data. This observation implies that the fine-tuning process on face-centric datasets of PFD models diminishes their ability to follow complex HOI text prompts. Therefore, to advance current PFD methods in HOI tasks, reintroducing the capability to leverage text prompts for natural full-body interactions is essential.

Motivated by this need, we propose a straightforward yet powerful approach: leveraging the generative capabilities of pre-trained diffusion models, such as StableDiffusion (SD), to augment existing PFD frameworks. Our method, PersonaHOI, integrates the strengths of SD without requiring additional training or fine-tuning, thereby restoring the ability to generate realistic, text-driven full-body HOI while preserving identity-specific details from the reference image.

Central to this integration, we incorporate an additional SD branch that aligns identity-specific details from PFD models with the HOI layouts generated by SD. A head segmentation mask derived from the SD output guides the merge of non-facial components from the SD branch with the facial details from the PFD branch. Specifically, in the PFD model branch, we introduce a Cross-Attention Constraint to prevent the overemphasis of identity features across the entire image, ensuring these details are confined to the facial region. To integrate HOI representations from SD, we implement Latent Merge and Residual Merge, merging identity and contextual features at both the latent representation level and via skip connections within the U-Net architecture. This multi-level merging strategy ensures that the generated images maintain realistic human-object interactions while preserving personalized facial details, resulting in cohesive, interaction-rich outputs.

To evaluate the effectiveness of our method in HOI scenarios, we propose a novel "interaction alignment" metric. These metrics leverage HOI detectors to measure alignment between generated images and text prompts, objectively assessing interaction realism. Our contributions can be summarized as follows:

- We present PersonaHOI, a dual-path architecture that integrates general and personalized diffusion models. Our framework generates realistic human-object interactions with specific identities, without requiring additional training or test-time tuning.
- Our method introduces key components, including cross-attention constraints and a combination of latent and residual merge strategies. These innovations enable effective feature integration, ensuring that generated images maintain both interaction coherence and identity fidelity.
- We design a novel evaluation metric for human-object interaction. Our experiments show significant improvements over state-of-the-art face personalization techniques in terms of interaction realism across models, emphasizing the scalability and robustness of our approach.

2. Related Works

Personalized Face Generation Diffusion Models. Personalized face generation aims to create identity-consistent images of individuals from limited reference data [32, 36, 43]. Given pretrained general image generation models [25], traditional optimization-based approaches [5, 26, 27] refine model parameters during inference for each individual but are computationally expensive, making them unsuitable for real-time applications. In contrast, learning-based approaches leverage human-centric datasets [1, 15, 19, 20] to train robust models capable of varying identity inputs without requiring test-time fine-tuning. Facial representations, extracted by models like CLIP [23] or other face recognition models [3, 16, 33], are often fused with [18, 21, 35] or replace [30] textual embeddings of specific words, such as "person", to provide identity conditions for diffusion models and enhance contextual relevance [3, 29]. In IP-Adapter [37] and other works [7, 9, 34], facial features are processed separately through decoupled cross-attention modules. Incorporating human masks as prior, methods can clean the background during data processing; alternatively, masks can be leveraged to construct [2, 18, 21, 35] and adjust [12] loss functions during training. In our work, we extend face personalization to complex human-object interactions (HOI), achieving a higher level of personalization and adaptability for diverse generative tasks.

Human-Object Interaction in Diffusion Models. HOI detection aims to locate human-object pairs and categorize their interactions as triplets, such as (person, playing, football) [4, 38–40], while HOI image synthesis remains relatively under-explored. InteractGAN [6] generates HOI images using human pose templates and reference images, but its reliance on pose-template pools and object references limits flexibility. Recent diffusion-based approaches tackle HOI generation by incorporating layout controls such as bounding boxes [10, 14, 17] or human poses [17, 42], or by leveraging in-context samples with similar interactions [11, 44]. Additionally, ReCorD [14] integrates Latent Diffusion Models with Visual Language Models, introducing modules for reasoning and correcting interactions to enhance cross-modal alignment. In this work, we rely on the standard text-to-image diffusion model (*i.e.,* StableDiffusion [25]) for HOI generation, and fuse the generation process with personalized face diffusion models to generation identity-preserving HOI images.

3. Preliminary

Understanding the fundamentals of StableDiffusion (SD) is essential to leverage its generative capabilities within our PFG framework. SD is a latent diffusion model designed to generate high-quality images based on text or image prompts. It operates in a latent space, where an input image is encoded in a latent representation $\mathbf{z}_0$ using a pretrained variational autoencoder (VAE) encoder, reducing the dimensionality of the image for efficient processing.

During the generation phase, the SD model begins with an initial noisy latent representation $\mathbf{z}_T$ and progressively denoises it over T timesteps to reconstruct $\mathbf{z}_0$. This denoised latent is then decoded by the VAE decoder to produce the final image. At each timestep t, the U-Net architecture θ predicts noise $\epsilon(\mathbf{z}_t, t, \mathbf{C})$, where $\mathbf{C}$ represents the conditioning inputs (text or image embeddings). The iterative denoising process follows:

$$\mathbf{z}_{t-1} = \text{Denoise}(\mathbf{z}_t, \epsilon(\mathbf{z}_t, t, \mathbf{C}); \theta) \tag{1}$$

The U-Net architecture comprises an encoder-decoder structure with skip connections that transfer features between the encoder and the decoder. Within this architecture, SD employs a cross-attention mechanism to incorporate conditioning inputs effectively during denoising. This mechanism projects latent representations and conditioning inputs into query, key, and value spaces:

$$\begin{cases} \mathbf{Q} = \mathbf{W}_q \mathbf{z}, \mathbf{K} = \mathbf{W}_k \mathbf{C}, \mathbf{V} = \mathbf{W}_v \mathbf{C}; \\ \mathbf{A} = \text{Softmax}\left(\dfrac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}}\right), \mathbf{z}_{\text{attn}} = \mathbf{A}\mathbf{V}. \end{cases} \tag{2}$$

Here, $\mathbf{A}$ is the attention map that ensures that relevant conditioning tokens guide denoising at each generation step. This cross-attention mechanism is critical for aligning the generated image with the input prompts, enabling consistency and fidelity in the final output.

4. Method

4.1. Overall Architecture

As illustrated in Figure 3, PersonaHOI augments the PFD model by integrating an additional StableDiffusion (SD) branch, enhancing its capacity to generate personalized images that capture complex human-object interactions (HOI)

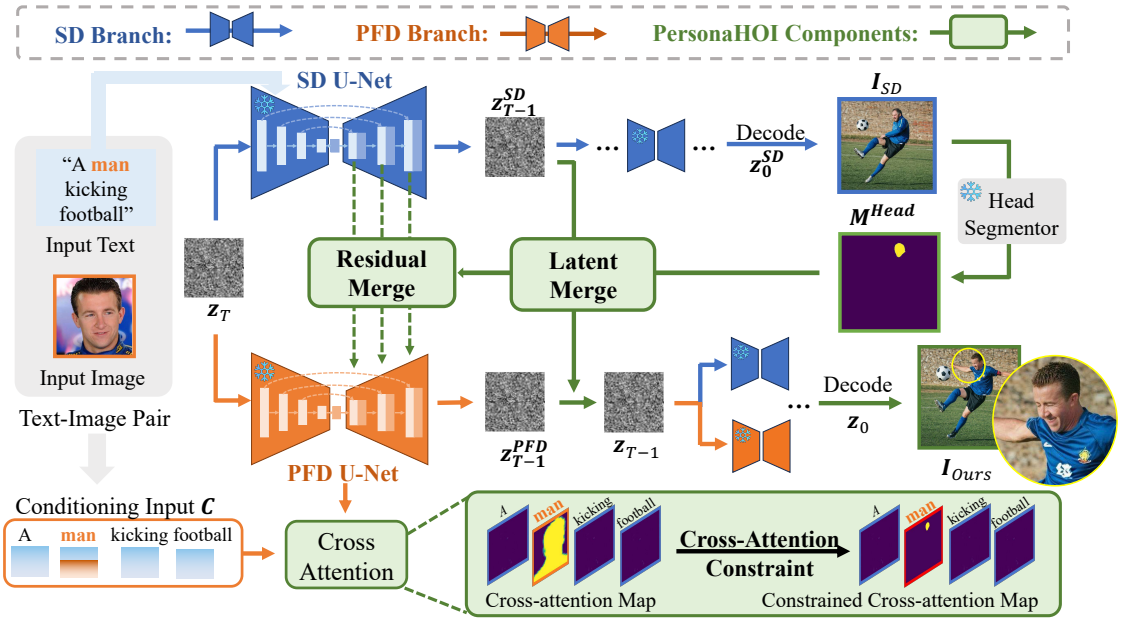


Figure 3. **Overview of Our Proposed Framework, PersonaHOI.** The architecture integrates a personalized face diffusion (PFD) model with an additional StableDiffusion (SD) branch. First, SD generates an image (I_{SD}) from a text prompt and noisy latent representation (z_T), which is decoded and segmented to produce a head mask. Next, SD and PFD run in parallel from the same z_T. At every timestep t, the head mask guides the Cross-Attention Constraint in PFD and merging modules (Latent Merge and Residual Merge) to merge interaction-relevant features from SD with identity-specific details from PFD. Iteratively, this process introduces HOI context to personalized face generation in a training&tuning-free manner.

in a training-free manner. Given an input comprising a user-provided reference image and a text prompt specifying the interaction (e.g., "a person kicking football"), our framework produces a cohesive output image that retains the subject's identity while accurately depicting the specified HOI.

To effectively merge identity from PFD and interaction features from SD, we introduce three core strategies: Cross-Attention Constraint, which regulates attention to reference image features within the PFD model (detailed in Sec. 4.2); Latent Merge, which combines features in the latent space at each generation timestep (Sec. 4.3); and Residual Merge, which integrates identity details through skip connections in the U-Net architecture (Sec. 4.4). Together, these strategies enable PersonaHOI to align generated content with interaction-driven text prompts while maintaining identity fidelity. Furthermore, this design allows PersonaHOI adaptable to various PFD models without the need for model retraining or test-time tuning.

4.2. Cross-Attention Constraint (CAC)

In the cross-attention layers of the PFD model, conditioning inputs $\mathbf{C} \in \mathbb{R}^{N \times D}$ include both text and image embeddings to guide the generation process, where N is the number of tokens, and D is the embedding dimension. Following established PFD model approaches [18, 21, 35], the n_{img}-th token in $\mathbf{C}$ incorporates the image embedding, encapsulating identity-specific details essential for facial preservation, while the remaining $N-1$ tokens correspond to text embeddings. In standard PFD models, attention to the n_{img}-th token often spreads across the image, overemphasizing facial features thus decreasing spatial capacity for representing body movements and object interactions.

To address this, we introduce Cross-Attention Constraint (CAC) within the cross-attention layers of the PFD model. CAC restricts identity features to specific facial regions, ensuring sufficient spatial capacity for body and object interactions. Different from the attention localization loss used in training [21, 35], our CAC utilizes fine-grained head mask derived from the SD-generated image I_{SD} in image generation process, ensuring the attention constraint in PFD is consistent with the HOI layout generated by SD. The head mask is segmented by an off-the-shelf head segmentor [24], then resized to match the spatial dimensions $(H \times W)$ of the latent representation $\mathbf{z}$, denoted as $\mathbf{M}^{\text{head}} \in [0,1]^{H \times W}$ (1 for the head region and 0 for other regions). We apply $\mathbf{M}^{\text{head}}$ to the attention map $\mathbf{A} \in \mathbb{R}^{(H \times W) \times N}$ as follows:

$$\mathbf{M}^{CAC}_{i \in [0,1,\ldots,N-1]} = \begin{cases} \text{flatten}(\mathbf{M}^{head}), & \text{if } i = img, \\ 1, & \text{otherwise} \end{cases} \quad (3)$$

$$\mathbf{A}^{CAC} = \mathbf{A} \odot (\mathbf{M}^{CAC})^{\intercal}.$$

$\mathbf{M}^{CAC}$ modifies the attention weights for the n_{img} token

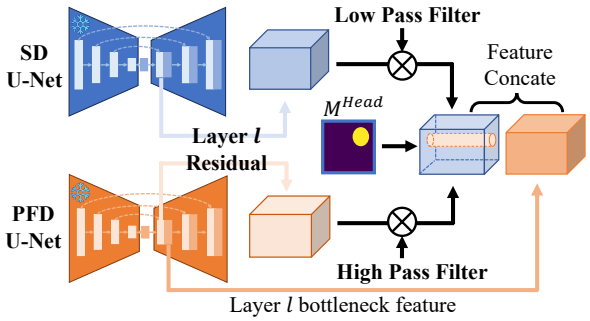


Figure 4. **Illustration of Residual Merge.** In each residual layer, Residual Merge operates within the U-Net skip connections, utilizing a head mask to guide the integration of high-frequency identity details from PFD residuals and low-frequency interaction layouts from SD residuals. The merged residuals are then concatenated to the corresponding bottleneck features from PFD.

by setting the weights to 0 in regions outside the head, preventing the influence of human identity on non-facial areas. Since the head mask is derived from SD-generated outputs, CAC maintains facial layout coherence with SD. When further combined with Latent and Residual Merge, CAC helps blend facial features from PFD and non-facial interaction features from SD without mutual interference.

4.3. Latent Merge (LM)

In this section, we introduce Latent Merge, which directly merges the latent representation within U-Net from PFD and SD models. After applying Cross-Attention Constraint (CAC) in Sec. 4.2, the facial region generated by the PFD model becomes spatially aligned with the corresponding region in the broader HOI layout produced by the SD model. This alignment facilitates spatial merge, blending facial identity details from PFD with interaction contexts from SD in latent space.

At each diffusion timestep t of PFD and SD, we implement the latent space merging with $\mathbf{M}^{head}$ as follows:

$$\mathbf{z}_t = \mathbf{M}^{head} \odot \mathbf{z}_t^{\text{PFD}} + (1 - \mathbf{M}^{head}) \odot \mathbf{z}_t^{\text{SD}}. \quad (4)$$

Here, $\mathbf{z}_t^{\text{PFD}}$ and $\mathbf{z}_t^{\text{SD}}$ represents the latent features from PFD and SD at timestep t, respectively. The merged latent $\mathbf{z}_t$ then serves as input for both SD and PFD models in the next denoising timestep $t-1$. This merging strategy allows identity details retained within facial regions, while non-facial regions maintain the HOI context generated by SD.

4.4. Residual Merge (RM)

While Latent Merge enables interactions between the SD and PFD at the latent level, it lacks integration at the intermediate feature stage within the U-Net. To address this, we introduce Residual Merge, which enhances feature integration within the skip connections. In the U-Net, skip connections transmit residual features containing rich detail from the encoder to the decoder, significantly impacting the content and quality of the generated images [13, 31].

Our Residual Merge is applied to each skip connection layer l ($l = 1, \ldots, L$), as shown in Figure 4. First, residual features are extracted from both the PFD and SD U-Net backbones. A low-pass filter and high-pass filter is applied to the SD and PFD residuals, respectively. Low-pass and high-pass filters are applied to these residuals to balance global coherence and local precision: low-pass filtering SD residuals suppresses high-frequency noise to retain the broader scene layout, while high-pass filtering PFD residuals emphasizes fine-grained identity-specific details. To achieve precise integration between facial and non-facial regions, we apply the head segmentation mask $\mathbf{M}^{\mathbf{head}}$, resized to match the residual resolution at layer l, denoted as $\mathbf{M}_{\text{R}}^{l}$. The fusion of residual features from both paths is then implemented for each layer l as follows:

$$\mathbf{R}^{l}_{\text{merged}} = \mathbf{M}^{l}_{R} \odot \text{HP}(\mathbf{R}^{l}_{\text{PFD}}) + (1 - \mathbf{M}^{l}_{R}) \odot \text{LP}(\mathbf{R}^{l}_{\text{SD}}), \quad (5)$$

where $\text{HP}(\cdot)$ and $\text{LP}(\cdot)$ denote the high pass and low pass filter, $\mathbf{R}_{\text{PFD}}$ and $\mathbf{R}_{\text{SD}}$ denote the residual features from the PFD and SD paths, respectively. The merged residual $\mathbf{R}_{\text{merged}}$ is then concatenated with the bottom-up features in the PFD U-Net for further processing.

The Residual Merge strategy enables fine-grained feature integration within the U-Net. Through this, facial regions retain high-frequency identity details from PFD while non-facial regions incorporate interaction layouts from SD, ensuring cohesive and contextually aligned outputs.

5. Experiments

5.1. Setup

Evaluation Data. We use two test sets for evaluation:
1) *HOI-Focused Personalized Face Generation:* To test complex human-object interactions (HOI), we use interaction labels from the widely-used V-COCO [8] dataset. V-COCO provides diverse HOI scenarios for robust evaluation of interaction fidelity. We construct 30 prompts in the "subject + interaction + object" format (*e.g.*, "man kicking a ball") with the same 15 reference subjects as in 1).
2) *General Personalized Face Generation:* Following prior studies [21, 35], we compile a benchmark with 15 reference subjects, each paired with 40 prompts spanning diverse scenarios like context, stylization, accessory, and actions. These prompts aim to assess both identity retention and the model's ability to adapt to various text-guided settings.
Full prompt details are provided in the Appendix **??**.

Metrics. We employ the following evaluation metrics:

Method	Architecture	Identity Preservation (%)	Prompt Consistency (%)	HOI Alignment (%)
StableDiffusion v1.5 [25]	SD1.5	NA	24.96	60.23
StableDiffusion XL [22]	SDXL	NA	25.32	70.78
FastComposer [35]	SD1.5	53.57	21.30	35.96
+ Ours	SD1.5	55.28 (+1.71)	23.16 (+1.86)	56.65 (+20.69)
IP-Adapter [37]	SDXL	**62.86**	22.73	49.83
+ Ours	SDXL	55.74 (-7.11)	24.20 (+1.47)	68.30 (+18.47)
PhotoMaker [18]	SDXL	50.54	23.96	56.77
+ Ours	SDXL	51.34 (+0.80)	**24.34** (+0.38)	**76.01** (+19.24)

Table 1. **Comparison of Our Method with Baseline Approaches on HOI-Specific Personalized Face Generation.** StableDiffusion serves as the text-only baseline without subject conditioning. PersonaHOI seamlessly incorporates existing Personalized Face Diffusion models (FastComposer [35], IP-Adapter [37], PhotoMaker [18]) with their corresponding StableDiffusion architectures. We **bold** the higher number for each pair of comparison.



Figure 5. **Qualitative Examples of PersonaHOI and Baseline Models.** Comparison of baseline models (FastComposer [35], IP-Adapter [37], PhotoMaker [18]) and their PersonaHOI-enhanced results for diverse human-object interaction prompts.

1) *Identity Preservation:* Measures how well the generated images retain the identity of the reference image. Face detection is performed using MTCNN [41], and identity similarity is calculated with FaceNet [28], following [21, 35].
2) *Prompt Consistency:* Evaluates text-image alignment using CLIP-based [23] scores, where higher values indicate better prompt adherence.
3) *HOI Alignment (Ours):* Assesses how well the generated images depict specified human-object interactions. We adopt UPT [40], an HOI detector pre-trained on V-COCO dataset, to identify the interaction specified in text prompts for generated images with HOI alignment score S_{HOI}. Specifically, UPT provides confidence scores for detected human and object classes (S_{Human}, S_{Object}) from its object detection head, as well as interaction logits (L_{Inter}) from its interaction head. With these outputs, HOI align-

ment score S_{HOI} is defined as:

$$S_{\text{HOI}} = (S_{\text{Human}} \cdot S_{\text{Object}})^{\lambda} \cdot \sigma(L_{\text{Inter}}) \quad (6)$$

where σ is the sigmoid activation function and $\lambda = 2.8$. We validate the human alignment and effectiveness for our proposed HOI Alignment metric in Appendix **??**.

Compared Methods and Implementation Details. We compare our method with state-of-the-art learning-based personalized face diffusion models, including FastComposer [35], IP-Adapter [37], and PhotoMaker [18]. Our PersonaHOI used their pre-trained diffusion backbones as the additional branch, *i.e.*, StableDiffusion v1.5 for FastComposer and StableDiffusion XL for PhotoMaker, with output resolutions set to 512×512 and 1024×1024, respectively. Head segmentation for SD generated image is performed with DensePose [24] due to its robust segmentation accuracy. In our generation process, no additional training or test-time tuning is applied to any compared methods.

5.2. Personalized Face with HOI Generation

Qualitative Comparison. Figure 5 illustrates qualitative comparisons between baseline models (FastComposer [35], IP-Adapter [37], PhotoMaker [18]) and their PersonaHOI-enhanced counterparts for diverse human-object interaction prompts. Baseline methods often fail to produce meaningful interactions, leading to missing or poorly integrated objects and unnatural human-object dynamics. For instance, in the prompt "A man skateboarding with skateboard" (first row of Figure 5), FastComposer [35] and PhotoMaker omit the skateboard entirely and IP-Adapter fails to generate the accurate interaction ("holding" the skateboard instead), resulting in inconsistent outputs with the text prompt. In contrast, PersonaHOI-enhanced models generate realistic skateboard placement and natural human-object dynamics. Moreover, PersonaHOI also ensures high-fidelity preservation of facial identity given in reference images.

Quantitative Comparison. Table **??** compares the performance of PersonaHOI across various baseline models. For FastComposer [35] and PhotoMaker [18], PersonaHOI achieves consistent improvements across all metrics. Notably, *Interaction Alignment* improves by 20.69% for FastComposer [35] and 19.24% for PhotoMaker [18], demonstrating superior human-object interaction coherence. Gains in *Identity Preservation* and *Prompt Consistency* further validate PersonaHOI's ability to retain identity while enhancing text-image alignment. For IP-Adapter [37], a decrease in *Identity Preservation* (from 62.86% to 55.74%) is observed. This drop is attributed to IP-Adapter's inherent bias toward generating face-centric images, which inflates identity scores at the cost of interaction fidelity. PersonaHOI mitigates this trade-off by significantly improving *Interaction Alignment* (+18.47%) and *Prompt Consistency* (+1.47%), balancing identity preservation with realistic interaction synthesis. Despite the decrease, the identity preservation score of 55.74% remains competitive for personalized face generation.



Figure 6. **Qualitative results for multi-subject HOI generation.** We compare generation outputs from SD v1.5 [25], FastComposer [35], and our PersonaHOI based on FastComposer [35] with different multi-subject interaction prompts. PersonaHOI not only preserves distinct identities but also generates coherent human-object interactions aligned with HOI layouts produced by SD v1.5.

Overall, PersonaHOI bridges the gap between personalized face generation and interaction realism. Its training- and tuning-free design enables seamless integration into different SD-based personalized face diffusion models, providing scalability and adaptability across diverse architectures and application scenarios.

5.3. Personalized Multi-Subject HOI Generation

Generating images with multiple subjects engaged in human-object interactions (HOI) presents significant challenges in preserving distinct identities and accurately depicting interactions. As shown in Figure 6, our method enhances FastComposer [35] using its pre-trained SD v1.5 model, maintaining identity fidelity while generating coherent interactions. For instance, in the challenging "cycling together" (second line) scenario, PersonaHOI successfully preserves the unique identities of both subjects and generates two bicycles alongside a natural background. In contrast, FastComposer [35] produces two adjacent faces with minimal interaction details. These results highlight the potential of PersonaHOI in handling complex multi-subject HOI scenarios, ensuring both identity preservation and interaction realism.

5.4. More Results

General Personalized Face Generation Performance. We evaluate PersonaHOI on the task of general personalized face generation [21, 35] across four categories of text prompts: *Accessory*, *Style*, *Action*, and *Context*. As shown in Table 2, the scores across all categories for the PersonaHOI-enhanced method are consistently comparable to or slightly better than the baseline FastComposer [35].

Method	Accessory (%)	Style (%)	Action (%)	Context (%)	Mean (%)
StableDiffusion v1.5 [25]	NA / 26.70	NA / 27.21	NA / 23.66	NA / 25.86	NA / 25.86
FastComposer [35]	54.65 / 24.22	41.13 / 24.01	**55.35** / 21.30	52.70 / 22.31	50.95 / 22.96
+ Ours	**56.43 / 24.25**	**46.07 / 23.97**	55.09 / **22.21**	**53.77 / 22.59**	**52.84 / 23.26**

Table 2. **Comparison of Our Method with FastComposer [35] on General Personalized Face Generation.** We compare across four categories of text prompts including Accessory, Style, Action, and Context, following [21, 35]. Results are formatted as " Identity Preservation (%) / Prompt Consistency (%)" and we **bold** the higher number for each pair of comparison.

These results demonstrate that while our method is specifically designed for personalized HOI scenarios, it effectively preserves the generative capabilities of the base models and does not degrade their performance on general text prompts, showcasing its adaptability and robustness.

Effectiveness of Individual Component. We conducted an ablation study on FastComposer [35] to evaluate the contributions of the Cross-Attention Constraint (CAC), Latent Merge (LM), and Residual Merge (RM) in PersonaHOI. As shown in Table **??**, the full model, integrating all three components, achieves the best performance across all metrics, underscoring the complementary roles of our proposed modules. Removing LM or RM results in significant declines in both *Identity Preservation* and *Interaction Alignment*, highlighting the critical role of these merging strategies. The absence of CAC has a relatively smaller effect on *Interaction Alignment*, but it substantially impacts *Identity Preservation*. This is because without CAC, faces generated by PFD fail to align with the spatial layout provided by SD, disrupting the coherence between PFD and SD branches and leading to diminished alignment and identity fidelity in facial regions. In summary, the ablation results confirm the distinct and critical functionalities of each module. CAC ensures precise spatial alignment of faces between PFD and SD, while LM and RM are indispensable for consistently merging non-facial HOI-relevant features from SD with facial identity details from PFD.

Results on Additional Baselines. To further validate the efficacy of our approach, we conduct additional evaluations using InstantID [34] and PuLID [7]. Results in Table 3 and Figure 7 indicate PersonaHOI can benefit both PFD methods, notably improving HOI alignment and image-text consistency. InstantID shares design similarities with IP-Adapter, prioritizing portrait generation, which improves identity preservation but compromises interaction fidelity. The fusion mechanism of PersonaHOI significantly enhances HOI alignment and overall image-text consistency. PuLID, on the other hand, balances personalization against text-guided image synthesis capability during training. However, it can struggle to correctly place the object in some cases (second row in Figure 7). By incorporating PersonaHOI, we achieve consistent improvements across identity preservation, image-prompt consistency, and interaction alignment, demonstrating the versatility and robustness of our method.

Method	Identity (%)	Prompt (%)	HOI (%)
InstantID	76.21	22.93	45.73
+ Ours	69.93 (-9.28)	23.80 (+0.87)	61.82 (+16.09)
PuLID	55.93	24.50	72.10
+ Ours	62.56 (+6.63)	25.21 (+0.71)	72.18 (+0.08)

Table 3. Comparison of our method with InstantID and PuLID on HOI-Specific Personalized Face Generation.



Figure 7. Visual comparison of InstantID [34] and PuLID [7] with our method.

6. Conclusions

In this work, we propose a training&tuning-free framework that combines the strengths of Personalized Face Diffusion (PFD) models and StableDiffusion (SD) models to generate personalized images featuring realistic and complex human-object interactions (HOI). By introducing Cross-Attention Constraint, Latent Merge, and Residual Merge, our approach achieves seamless integration of identity-specific facial features from PFD with contextual interaction details from SD, ensuring both spatial alignment and contextual coherence. The flexibility of our framework allows adaptation to various PFD and SD architectures, demonstrating its robustness in handling diverse interaction scenarios. Without requiring additional fine-tuning or task-specific datasets, our method significantly improves identity retention and interaction realism compared to state-of-the-art approaches. This highlights the potential of leveraging pre-trained models to enable high-quality, personalized content generation across diverse applications.

References

[1] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 67–74. IEEE, 2018. 3

[2] Li Chen, Mengyi Zhao, Yiheng Liu, Mingxu Ding, Yangyang Song, Shizun Wang, Xu Wang, Hao Yang, Jing Liu, Kang Du, et al. Photoverse: Tuning-free image customization with text-to-image diffusion models, 2023. 3

[3] Zhuowei Chen, Shancheng Fang, Wei Liu, Qian He, Mengqi Huang, Yongdong Zhang, and Zhendong Mao. Dreamidentity: Improved editability for efficient face-identity preserved image generation. *arXiv preprint arXiv:2307.00300*, 2023. 3

[4] Dylan Campbell Frederic Z. Zhang and Stephen Gould. Spatially Conditioned Graphs for Detecting Human–Object Interactions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13319–13327, 2021. 3

[5] 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. *arXiv preprint arXiv:2208.01618*, 2022. 3

[6] Chen Gao, Si Liu, Defa Zhu, Quan Liu, Jie Cao, Haoqian He, Ran He, and Shuicheng Yan. InteractGAN: Learning to Generate Human-Object Interaction. In *Association for Computing Machinery*, 2020. 3

[7] Zinan Guo, Yanze Wu, Zhuowei Chen, Lang Chen, Peng Zhang, and Qian He. PuLID: Pure and Lightning ID Customization via Contrastive Alignment. In *Advances in Neural Information Processing Systems*, 2024. 3, 8

[8] Saurabh Gupta and Jitendra Malik. Visual Semantic Role Labeling. *arXiv preprint arXiv:1505.04474*, 2015. 5

[9] 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. 3

[10] Jiun Tian Hoe, Xudong Jiang, Chee Seng Chan, Yap-Peng Tan, and Weipeng Hu. InteractDiffusion: Interaction Control in Text-to-Image Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 3

[11] Siteng Huang, Biao Gong, Yutong Feng, Xi Chen, Yuqian Fu, Yu Liu, and Donglin Wang. Learning disentangled identifiers for action-customized text-to-image generation. *arXiv preprint arXiv:2311.15841*, 2023. 3

[12] Junha Hyung, Jaeyo Shin, and Jaegul Choo. Magicapture: High-resolution multi-concept portrait customization. *arXiv preprint arXiv:2309.06895*, 2023. 3

[13] Zeyinzi Jiang, Chaojie Mao, Yulin Pan, Zhen Han, and Jingfeng Zhang. SCEdit: Efficient and Controllable Image Diffusion Generation via Skip Connection Editing. *arXiv preprint arXiv:2312.11392*, 2023. 5

[14] Jian-Yu Jiang-Lin, Kang-Yang Huang, Ling Lo, Yi-Ning Huang, Terence Lin, Jhih-Ciang Wu, Hong-Han Shuai, and Wen-Huang Cheng. ReCorD: Reasoning and Correcting Diffusion for HOI Generation. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, 2024. 3

[15] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3

[16] Xiaoming Li, Xinyu Hou, and Chen Change Loy. When stylegan meets stable diffusion: a w+ adapter for personalized image generation. *arXiv preprint arXiv:2311.17461*, 2023. 3

[17] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. GLIGEN: Open-Set Grounded Text-to-Image Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

[18] Zhen Li, Mingdeng Cao, Xintao Wang, Zhongang Qi, Ming-Ming Cheng, and Ying Shan. PhotoMaker: Customizing Realistic Human Photos via Stacked ID Embedding. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2, 3, 4, 6, 7

[19] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep Learning Face Attributes in the Wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2015. 3

[20] A. Nagrani, J. S. Chung, and A. Zisserman. Voxceleb: a large-scale speaker identification dataset. In *INTERSPEECH*, 2017. 3

[21] Xu Peng, Junwei Zhu, Boyuan Jiang, Ying Tai, Donghao Luo, Jiangning Zhang, Wei Lin, Taisong Jin, Chengjie Wang, and Rongrong Ji. PortraitBooth: A Versatile Portrait Model for Fast Identity-preserved Personalization. *arXiv preprint arXiv:2312.06354*, 2023. 2, 3, 4, 5, 6, 7, 8

[22] 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. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024. 6

[23] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2021. 3, 6

[24] Iasonas Kokkinos Riza Alp Güler, Natalia Neverova. DensePose: Dense Human Pose Estimation In The Wild. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 4, 7

[25] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-Resolution Image Synthesis with Latent Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2, 3, 6, 7, 8

[26] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22500–22510, 2023. 3

[27] 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. 3

[28] Florian Schroff, Dmitry Kalenichenko, and James Philbin. FaceNet: A Unified Embedding for Face Recognition and Clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 6

[29] 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, 3

[30] Kaede Shiohara and Toshihiko Yamasaki. Face2diffusion for fast and editable face personalization. *arXiv preprint arXiv:2403.05094*, 2024. 3

[31] Chenyang Si, Ziqi Huang, Yuming Jiang, and Ziwei Liu. FreeU: Free Lunch in Diffusion U-Net. In *CVPR*, 2024. 5

[32] Yu-Chuan Su, Kelvin CK Chan, Yandong Li, Yang Zhao, Han Zhang, Boqing Gong, Huisheng Wang, and Xuhui Jia. Identity encoder for personalized diffusion. *arXiv preprint arXiv:2304.07429*, 2023. 3

[33] Dani Valevski, Danny Lumen, Yossi Matias, and Yaniv Leviathan. Face0: Instantaneously conditioning a text-to-image model on a face. In *SIGGRAPH Asia 2023 Conference Papers*, pages 1–10, 2023. 3

[34] 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. 3, 8

[35] Guangxuan Xiao, Tianwei Yin, William T Freeman, Frédo Durand, and Song Han. FastComposer: Tuning-Free Multi-Subject Image Generation with Localized Attention. In *International Journal of Computer Vision*, 2024. 1, 2, 3, 4, 5, 6, 7, 8

[36] 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. 3

[37] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-Adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023. 3, 6, 7

[38] Hangjie Yuan, Jianwen Jiang, Samuel Albanie, Tao Feng, Ziyuan Huang, Dong Ni, and Mingqian Tang. RLIP: Relational Language-Image Pre-training for Human-Object Interaction Detection. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 3

[39] Hangjie Yuan, Shiwei Zhang, Xiang Wang, Samuel Albanie, Yining Pan, Tao Feng, Jianwen Jiang, Dong Ni, Yingya Zhang, and Deli Zhao. RLIPv2: Fast Scaling of Relational Language-Image Pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

[40] Frederic Z. Zhang, Dylan Campbell, and Stephen Gould. Efficient Two-Stage Detection of Human-Object Interactions with a Novel Unary-Pairwise Transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20104–20112, 2022. 3, 6

[41] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2016. 6

[42] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding Conditional Control to Text-to-Image Diffusion Models. In *IEEE International Conference on Computer Vision (ICCV)*, 2023. 3

[43] Xulu Zhang, Xiao-Yong Wei, Wengyu Zhang, Jinlin Wu, Zhaoxiang Zhang, Zhen Lei, and Qing Li. A Survey on Personalized Content Synthesis with Diffusion Models. *arXiv preprint arXiv:2405.05538*, 2024. https://arxiv.org/abs/2405.05538. 3

[44] Yuxin Zhang, Fan Tang, Nisha Huang, Haibin Huang, Chongyang Ma, Weiming Dong, and Changsheng Xu. MotionCrafter: One-Shot Motion Customization of Diffusion Models. *arXiv preprint arXiv:2312.05288*, 2023. 3