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

Supplementary Material

In this supplementary material, we provide additional details and results to complement the main paper. Specifically, we include: the validation of effectiveness of our proposed HOI Alignment metric (Section 7); the integration of ControlNet into the PersonaHOI framework for enhanced pose control in HOI generation (Section 8); extended results demonstrating the effectiveness of our method on General Personalized Face Generation tasks with diverse prompts (Section 9); visualizations combining general personalization with HOI across different scenarios like *Style*, *Context*, and *Accessory* (Section 10); a detailed comparison of image quality metrics such as FID, ImageReward, and Aesthetic Score (Section 11); comparison of the inference time (Section 12); comprehensive ablation studies exploring the impact of Gaussian kernel strategies, identity injection timesteps, and filter configurations (Section 13); and implementation details outlining the models and prompts used in our experiments (Section 14).

7. Validation of HOI Alignment Metric

Unlike prompt consistency which assesses global image-text coherence, our proposed HOI Alignment Metric, *i.e.*, S_{HOI} in Equation 6, provides a fine-grained evaluation of human-object interactions.

User Study. To validate S_{HOI}, we conduct a user study with 20 participants, each rating 20 generated images on a 4-point scale (0: Fail – 3: Excellent) based on their perceived interaction quality. Figure 8 (left) shows a boxplot comparing human ratings and S_{HOI}, revealing a strong positive correlation (Pearson: 0.78, P-value: 5.8×10^{-84}), confirming the reliability of S_{HOI}. In contrast, prompt consistency exhibits a significantly lower correlation with human ratings (Pearson: 0.28, P-value: 2.1×10^{-8}), as it focuses on overall image-text consistency rather than specific HOI.

Sample Visualization. We show S_{HOI} alongside the prompt consistency (measured by CLIP score) for two sets of generated images in Figure 8 (right). As illustrated, images with higher S_{HOI} correspond to more accurate and visually natural interactions following text prompts.

8. Combine PersonaHOI with ControlNet

In this section, we incorporate ControlNet [45] into our framework to improve human-object interactions by enabling precise pose control. Using pose information as an additional input, ControlNet enables enhanced customization for HOI content generation, offering greater flexibility for handling complex scenarios.

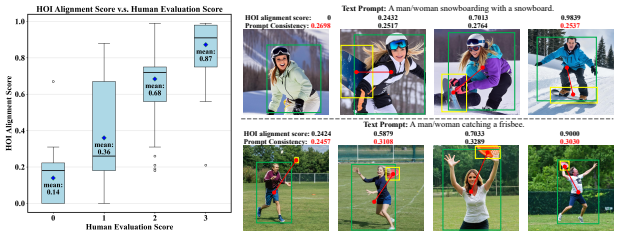

Figure 8. **Left:** A boxplot of human ratings *vs.* HOI alignment scores, showing a strong positive correlation (Pearson = 0.78, P-value = 5.8×10^{-84}). **Right:** Two sets of generated images ranked by increasing HOI alignment scores alongside Prompt Consistency (CLIP score). Red numbers highlight cases where CLIP score misaligns with intuitive interaction correctness.

Framework Modification. We integrate ControlNet [45] into our framework by replacing the StableDiffusion (SD) [26] branch with a ControlNet model. Human Pose images from the V-COCO dataset [9] are used as inputs, providing explicit pose constraints for image generation. Leveraging our scalable architecture, we combine the personalized face generation model, FastComposer [37], with the ControlNet branch using the proposed Cross-Attention Constraint, Latent Fusion, and Residual Fusion strategies. Notably, this integration is **training-free and requires no test-time tuning**, ensuring efficient incorporation of pose-specific controls while preserving identity-specific facial features.

Visualization. Figure 9 showcases examples of integrating ControlNet into our framework. By applying the same pose control to different subjects, our method effectively generates distinct identities with the specified pose, as well as faithfully depicts the human-object interaction described in the given text prompt. The generated human poses align closely with the provided poses, including aspects such as arm positioning, leg placement, and overall body orientation. Furthermore, the results demonstrate high fidelity in preserving facial identity, underscoring the effectiveness of our approach in achieving both pose accuracy and identity consistency across varied scenarios.

This experiment highlights the generalizability and flexibility of our framework. By incorporating ControlNet as an alternative branch, our method achieves fine-grained pose control in personalized face generation, making it adaptable to more complex and detailed HOI scenarios. This integration not only enhances the realism and coherence of the generated content but also broadens the applicability of our

Figure 9. Examples of integrating ControlNet [45] into the baseline FastComposer within our PersonaHOI framework.

approach, particularly in domains like virtual reality, gaming, and digital content creation.

9. More Results on General PFG

As detailed in Section 5 of the main text, we evaluate our method on the General Personalized Face Generation (General PFG) task using 40 test prompts from FastComposer [37]. These prompts encompass a variety of scenarios, including *Style*, *Accessory*, *Context*, and *Action*, enabling a comprehensive assessment of our model's adaptability across diverse conditions.

9.1. More Quantitative Results

Table 4 presents a comparison of our PersonaHOI-enhanced methods with baseline approaches (FastComposer [37], IP-Adapter [40], PhotoMaker [19]) on the General PFG task. Our methods consistently deliver balanced performance across *Identity Preservation* and *Prompt Consistency*, unlike baseline models, which often favor one metric at the expense of the other. Notably, IP-Adapter achieves the highest *Identity Preservation* scores but struggles with *Prompt Consistency*, especially in the *Style* category, where its score drops to just 18.25%. On the other hand, PhotoMaker excels in *Prompt Consistency*; however, it suffers from the lowest *Identity Preservation* score among all baselines (45.31%). In contrast, PersonaHOI achieves a strong balance by consistently ranking among the top two in most metrics. This underscores our capability to preserve identity while adhering to diverse text prompts effectively. Furthermore, our efficient training-free design improves its practicality, making it adaptable to a wide range of scenarios.

9.2. Visualization

Figure 10 illustrates challenging examples from four categories: *Style*, *Accessory*, *Context*, and *Action*, showcasing the comparison between our method and the baselines (FastComposer [37], IP-Adapter [40], PhotoMaker [19]). Our method shows significant improvements, achieving a strong balance between face personalization and prompt adherence. In the first row (*Style*), our approach accurately applies the specified stylization while maintaining the subject's identity, delivering outputs that are coherent and identity-consistent, surpassing the baselines. In the second row (*Accessory*), featuring "a man wearing pink glasses", our method faithfully generates the pink glasses specified in the prompt. By contrast, FastComposer and IP-Adapter misinterpret the prompt, producing outputs with pink cloth-

Method	Accessory (%)	Style (%)	Action (%)	Context (%)	Mean (%)
StableDiffusion v1.5 [26]	NA / 26.70	NA / 27.21	NA / 23.66	NA / 25.86	NA / 25.86
StableDiffusion XL [23]	NA / 27.48	NA / 27.49	NA / 24.57	NA / 26.82	NA / 26.67
FastComposer [37]	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
IP-Adapter [40]	**63.75** / 22.42	**64.16** / 18.25	**63.57** / 22.07	**62.91** / 21.86	**63.60** / 21.15
+ Ours	60.72 / 24.66	51.57 / 23.52	58.17 / **23.88**	60.29 / 23.68	57.69 / 23.94
PhotoMaker [19]	51.69 / **26.26**	27.34 / **26.85**	51.16 / 23.45	51.04 / **25.30**	45.31 / **25.46**
+ Ours	58.97 / 23.84	56.02 / 23.58	57.67 / 22.94	60.29 / 23.46	58.24 / 23.45

Table 4. **Comparison of Our Method with FastComposer [37] on General Personalized Face Generation.** We compare across four categories of text prompts including Accessory, Style, Action, and Context, following [22, 37]. Results are formatted as " Identity Preservation (%) / Prompt Consistency (%)". The best-performing results for each metric are highlighted in **bold**, while the second-best results are underlined.

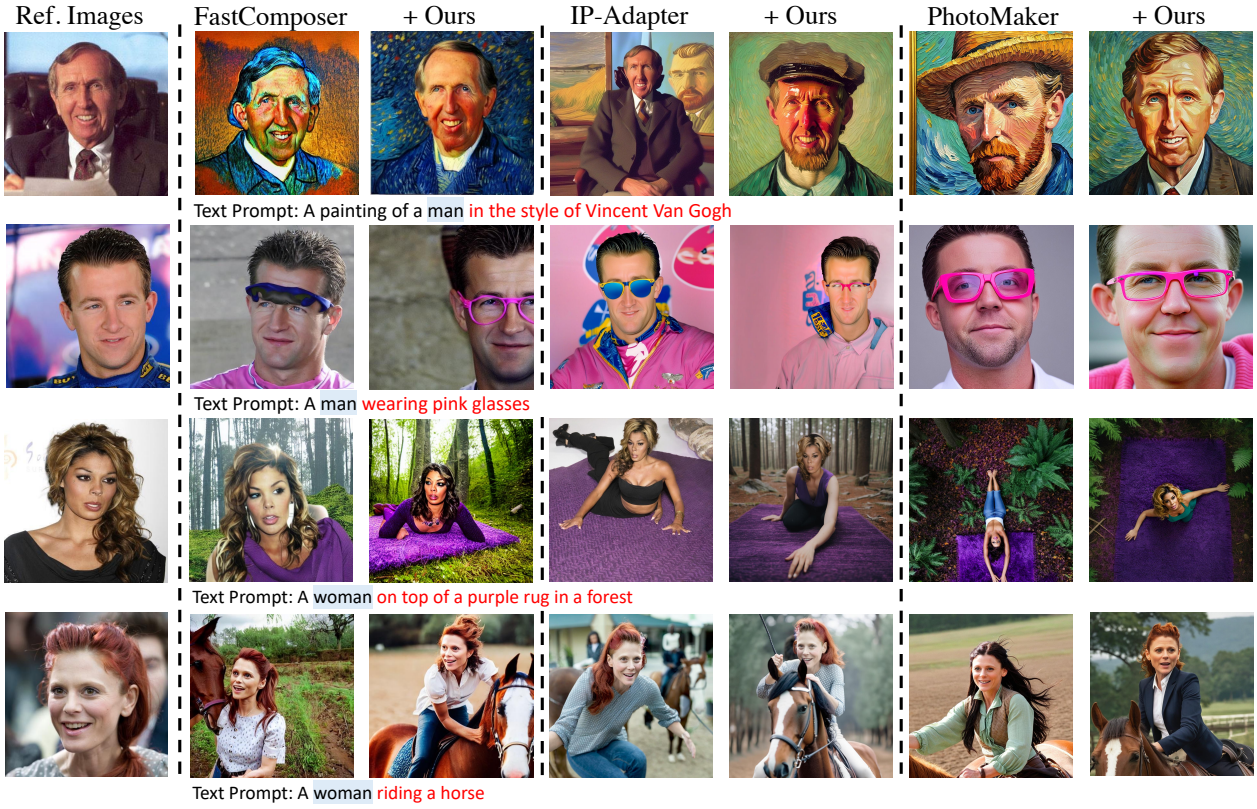

Figure 10. Visualization comparison of our method with baseline approaches (FastComposer [37], IP-Adapter [40], PhotoMaker [19]) across four categories of general personalized face generation: *Style*, *Accessory*, *Context*, and *Action*.

ing or backgrounds instead, illustrating the challenges of precise accessory generation. In the third row (*Context*), depicting "a woman on top of a purple rug in a forest", our method effectively captures the purple rug and forest background while preserving facial details, whereas the baselines fail to maintain scene coherence or facial fidelity. In the fourth row (*Action*), with the prompt "a woman riding a horse", our method captures both the riding action and the subject's facial features, producing realistic and cohesive results. In contrast, the baseline methods struggle with achieving realistic actions or maintaining identity consistency.

Figure 11 presents additional comparisons leveraging the high-quality SD-XL-based PhotoMaker [19]. By incorporating PersonaHOI, PhotoMaker demonstrates significant improvements in adhering to text prompts and preserving

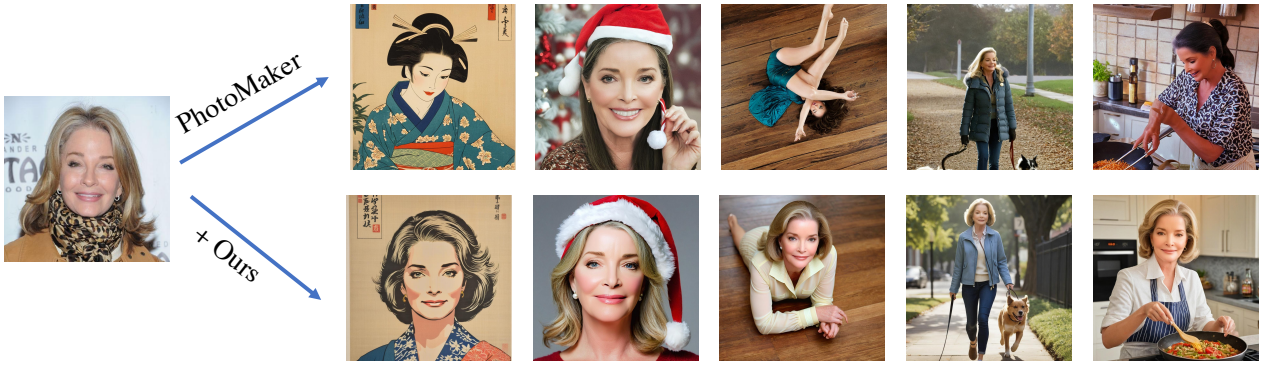

Figure 11. Visualization of our PersonaHOI-enhanced PhotoMaker [19] compared to the baseline. From left to right, the prompts are: "a Japanese woodblock print of a woman", "a woman wearing a Santa hat", "a woman on top of a wooden floor", "a woman walking a dog," and "a woman cooking a meal."

Figure 12. Examples of integrating general personalization with HOI across diverse scenarios. From top to bottom, the rows illustrate *Style+HOI*, *Context+HOI*, and *Accessory+HOI*.

facial features. For instance, given the prompt "a woman on top of a wooden floor", baseline results frequently display distorted facial features and unnatural human poses. In contrast, our method effectively preserves the subject's identity and accurately adheres to the given prompt. These findings underscore the robustness of our approach to maintaining identity personalization while achieving prompt fidelity.

Overall, these results highlight the flexibility and effectiveness of PersonaHOI in handling diverse and complex personalized face-generation tasks. By enhancing existing personalized face generation models, our approach integrates text prompt alignment and identity preservation, offering a versatile solution for advancing general PFG capabilities.

10. Visualization on General PFG + HOI

In this section, we provide additional examples of personalized face generation combining Human-Object Interaction (HOI) with general modifications, complementing Figure 1 from the main text. We focus on scenarios that combine *Context+HOI*, *Style+HOI*, and *Accessory+HOI*, as illustrated in Figure 12.

The results highlight our method's ability to integrate identity preservation with both HOI-specific and general prompt elements in personalized face generation. Unlike baseline models, which struggle to balance these tasks, our approach excels in producing coherent and contextually accurate outputs. In the first row (*Style+HOI*), FastComposer [37] and PhotoMaker [19] fail to generate the bench

properly, and IP-Adapter [40] neglects the stylization requirements, resulting in outputs that lack the desired artistic effect. In the second row (*Context+HOI*), all baseline methods struggle with the natural placement of the umbrella, creating awkward and unrealistic interactions. In the third row (*Accessory+HOI*), baseline methods either omit or generate incomplete frisbee, while our approach captures both the accessory and the interaction comprehensively.

These results highlight the robustness and adaptability of our method in addressing intricate prompts that combine general personalization with realistic human-object interactions. By excelling in both identity preservation and contextual fidelity, our approach offers a unified and effective solution for personalized face generation across diverse and complex scenarios.

11. Comparison of Image Quality

We evaluate the image quality of our method compared to baseline approaches on the task of Personalized Face with HOI Generation. The FID metric, calculated on the V-COCO [9] test set, quantifies the similarity between the distribution of generated images and that of realistic ones. To further assess image quality, we use ImageReward [38] and Aesthetic Score [30], which evaluate human preference alignment and visual appeal, respectively. As shown in Table 5, our method consistently outperforms baselines in both ImageReward and FID, highlighting its capacity to generate high-quality images that align closely with real-world distributions and human preferences. For Aesthetic Score, our approach significantly enhances the results for FastComposer [37] and IP-Adapter [40], emphasizing its effectiveness in improving visual quality. Although a slight decrease is observed for PhotoMaker [19], our method still maintains competitive performance. Overall, these results confirm the capability of our training-free framework to generate identity-preserving, interaction-rich images that balance realism, human preference, and aesthetic quality.

12. Comparison of Running time

Our diffusion models' fusion and head mask generation introduce an additional denoising and fusion pass along with head detection, increasing the computation time per image: FastComposer ($2s \rightarrow 6s$), PhotoMaker ($4s \rightarrow 16s$), IP-Adapter ($2s \rightarrow 9s$), InstantID ($5s \rightarrow 13s$) , and PuLID ($1s \rightarrow 4s$). Inference times are measured on an NVIDIA L40 GPU. We will explore more efficient denoising strategies to accelerate the pipeline in future.

13. Additional Ablation Studies

13.1. Ablation on Gaussian Kernels

We investigate the effect of Gaussian kernel sizes, controlled by the scaling factor α, on Personalized Face with HOI Generation. The kernel size is determined from the head segmentation mask extracted from SD-generated images. Specifically, the area of the head mask is computed and then scaled by taking its square root to derive a base size. This base size is multiplied by α, where larger α values result in broader kernels, emphasizing global context, while smaller α values produce more compact kernels, focusing on fine-grained facial details.

Table 6 illustrates that constant kernel sizes exhibit a trade-off between metrics. Larger kernels (e.g., $\alpha = 3.5$) excel in *Action Alignment* (57.07%) by prioritizing interaction layouts but significantly compromise *Identity Preservation* (23.67%). Conversely, smaller kernels (e.g., $\alpha = 0.5$) preserve identity better (51.58%) but perform worse in *Action Alignment* (54.46%). To address this, we implement dynamic kernel strategies that adapt over timesteps. The decremental kernel ($2.5 \rightarrow 0.5$) achieves the best overall performance, delivering the highest *Identity Preservation* (55.28%) and competitive *Action Alignment* (56.65%). In contrast, the incremental kernel ($0.5 \rightarrow 2.5$) underperforms across all metrics. These findings suggest that starting with larger kernels to capture global interaction layouts and progressively reducing them to refine facial details is the most effective approach. Consequently, we adopt the decremental kernel in all experiments.

13.2. Ablation on Identity Injection Timestep

In the Introduction of the main paper, we discuss that existing methods [19, 22, 37] often adopt a delayed injection strategy, introducing identity embeddings at later diffusion timesteps to balance text alignment and identity preservation. This approach allows text embeddings to dominate early stages, enhancing prompt adherence before incorporating identity-specific details.

In contrast, our PersonaHOI framework integrates StableDiffusion (SD) from the beginning of the generation process, leveraging its robust text alignment capabilities. This enables immediate injection of identity embeddings at timestep 0, ensuring seamless integration of identity-specific details without compromising text alignment or interaction coherence. As shown in Table 7, our method achieves the highest *Identity Preservation* (55.28%) and *Action Alignment* (56.65%) while maintaining strong *Prompt Consistency* (23.16%). Delayed injection strategies, however, significantly diminish *Identity Preservation* (*e.g.*, 6.28% at timestep 50) as the influence of identity information is reduced during denoising. These results confirm that PersonaHOI effectively combines identity preservation and

	FID ↓	Aesthetic Score ↑	Image Reward ↑
FastComposer	85.98	6.02	0.39
+ Ours	82.28 (-3.70)	6.30 (+0.28)	0.88 (+0.49)
PhotoMaker	84.24	6.29	1.22
+ Ours	82.38 (-1.86)	6.20 (-0.09)	1.31 (+0.09)
IP-Adapter	80.59	6.11	0.65
+ Ours	78.41 (-2.18)	6.47 (+0.36)	0.91 (+0.26)

Table 5. Comparison of image quality on the task of Personalized Face with HOI Generation. Metrics include FID (lower is better), ImageReward (higher is better), and Aesthetic Score (higher is better). We use **(green)** scripts to denote the performance improvement and **(red)** scripts for the decrease.

α	Identity Pres. (%)	Prompt Consist. (%)	Action Align. (%)
0	51.34	22.79	54.03
0.5	<u>51.58</u>	22.80	54.46
1.5	51.20	22.87	54.92
2.5	47.25	22.96	56.04
3.5	23.67	**23.39**	**57.07**
$0.5 \rightarrow 2.5$	50.27	22.70	55.46
$2.5 \rightarrow 0.5$	**55.28**	<u>23.16</u>	<u>56.65</u>

Table 6. **Ablation Study on Gaussian Kernel Size.** We evaluate the impact of varying Gaussian kernel sizes with α on the task of Personalized Face with HOI Generation. The best-performing results for each metric are highlighted in **bold**, while the second-best results are <u>underlined</u>.

Timestep	Identity Pres. (%)	Prompt Consist. (%)	Action Align. (%)
	<u>53.57</u>	21.30	35.96
0	6.28	**23.46**	<u>56.73</u>
10	10.05	<u>23.39</u>	56.47
20	22.89	23.24	56.54
30	36.49	23.04	56.32
40	44.92	22.95	55.96
50	**55.28**	23.16	**56.65**

Table 7. **Ablation Study on Identity Injection Timestep.** We analyze the impact of injecting identity embeddings at different timesteps on the task of Personalized Face with HOI Generation. The experiments are conducted on FastComposer [37] with a total of 50 diffusion timesteps. The first row represents the baseline results from FastComposer without our method. The best-performing results for each metric are highlighted in **bold**, while the second-best results are <u>underlined</u>.

text alignment, eliminating the limitations of delayed strategies and ensuring a balanced integration of identity and interaction realism throughout the generation process.

CAC	LM	RM	Identity. (%)	Prompt. (%)	Interaction. (%)
			53.57	21.30	35.96
✓	✓	✓	**55.28**	**23.16**	**56.65**
✓		✓	47.02	22.43	46.62
✓	✓		49.97	23.03	55.80
	✓	✓	45.56	22.76	53.21

Table 8. **Effect of Individual Components.** We evaluate the contributions of Cross-Attention Constraint (CAC), Latent Merge (LM), and Residual Merge (RM) in PersonaHOI by selectively removing each of them. Experiments are conducted with FastComposer on HOI-specific personalized face generation. Red numbers denote the performance lower than FastComposer [37] baseline.

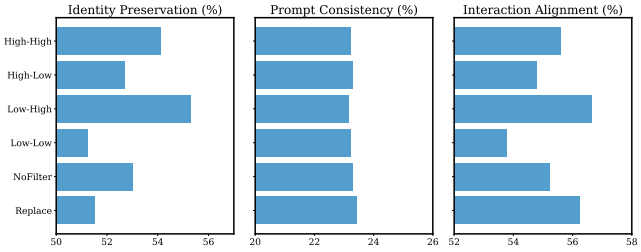

Figure 13. **Effect of Low-Pass and High-Pass Filters in Residual Merge.** We evaluate six configurations: direct replacement (Replace), merge without filter (NoFilter), and combinations of low-pass and high-pass filters applied to SD and PFD branches. Our Low-High configuration, which applies a low-pass filter to SD and a high-pass filter to PFD, achieves the best overall balance, demonstrating its effectiveness as the optimal merging strategy.

13.3. Impact of High-Pass/Low-Pass Filters

We validate the design choice of applying a low-pass filter to the StableDiffusion (SD) branch and a high-pass filter to the personalized face diffusion models (PFD) branch in Residual Merge. To this end, we compare six settings: direct replacement of PFD residuals with SD (Replace), merging without filters (NoFilter), and combinations of low-pass and high-pass filters (*i.e.*, Low-Low, High-High, High-Low, Low-High) for SD and PFD, respectively. As shown in Figure 13, direct replacement and no-filter ap-

Figure 14. **Visual comparison of different filter configurations in Residual Fusion.** The configurations include fusion without filters (*NoFilter*) and different combinations of low-pass and high-pass filters (*Low-Low*, *High-High*, *High-Low*, and *Low-High*) applied to PFD and SD. Experiments are conducted with FastComposer as the backbone. Please zoom in on the images for a clearer comparison.

proaches yield suboptimal results, emphasizing the importance of a balanced merging strategy. Among the configurations, the Low-High design achieves the best overall performance across all metrics, confirming the effectiveness of our Residual Merge design.

To further validate our observations, in Figure 14, we present visualizations of five configurations: fusion without filters (*NoFilter*) and combinations of low-pass and high-pass filters (*Low-Low*, *High-High*, *High-Low*, and *Low-High*) applied to PFD and SD.

The *NoFilter* configuration demonstrates strong initial results due to the inherent robustness of our Residual Fusion, Latent Fusion, and Cross-Attention Constraint. However, certain challenges persist. In the first and second rows, interactions involving the "cup" and "bench" appear distorted, leading to unnatural object dynamics and contextual layouts. Introducing high-pass and low-pass filters effectively mitigates these issues. Among the configurations, *Low-High* proves to be the most effective. It resolves contextual inconsistencies observed with *NoFilter*, producing realistic object placement (e.g., natural positioning of the cup and bench in the first and second rows). Furthermore, as illustrated in the third and fourth rows, *Low-High* enhances accessory placement (*e.g.,* snow glasses) and preserves detailed facial textures, delivering sharper visuals and well-balanced lighting. By contrast, other configurations (*High-High*, *High-Low*, *Low-Low*) show inferior performance, failing to achieve the same balance between global scene coherence and fine-grained details.

Overall, while *NoFilter* establishes a robust baseline, the addition of high-pass and low-pass filters, particularly in the *Low-High* configuration, significantly enhances the fusion process. This approach effectively addresses limitations, delivering the most balanced and realistic results for personalized human-object interaction generation.

14. Implementation Details

14.1. Off-the-Shelf Models

We employ several off-the-shelf models in implementation to ensure robust personalized generation and evaluation. For diffusion methods, we adopt the original configurations from baseline methods: StableDiffusion v1.5 (SD v1.5) [26] for FastComposer [37]; advanced StableDiffusion XL (SD-XL) [23] for IP-Adapter [40] and PhotoMaker [19]. Corresponding SD models are incorporated

into the PersonaHOI framework. For head mask segmentation, we use a pretrained DensePose [25] model (ResNet-50-FPN backbone), enabling precise extraction of head regions for fusion and attention constraints. To evaluate human-object interactions, we employ the pretrained UPT HOI detector [43] (ResNet-101-DC5 bakbone). For the combination of PersonaHOI and ControlNet [45], we utilize a pretrained SD v1.5-based ControlNet conditioned on human pose estimation. The pose control is extracted from V-COCO [9] dataset with Openpose [2] pose estimator.

14.2. Text Prompts for Image Generation

Prompts for General Personalized Face Generation. Following previous works [22, 37], we utilized 40 prompts across four types:

- Accessory:
  "a man/woman wearing a red hat",
  "a man/woman wearing a Santa hat",
  "a man/woman wearing a rainbow scar",
  "a man/woman wearing a black top hat and a monocle",
  "a man/woman in a chef outfit",
  "a man/woman in a firefighter outfit",
  "a man/woman in a police outfit",
  "a man/woman wearing pink glasses",
  "a man/woman wearing a yellow shirt",
  "a man/woman in a purple wizard outfit".
- Style:
  "a painting of a man/woman in the style of Banksy",
  "a painting of a man/woman in the style of Vincent Van Gogh",
  "a colorful graffiti painting of a man/woman",
  "a watercolor painting of a man/woman",
  "a Greek marble sculpture of a man/woman",
  "a street art mural of a man/woman",
  "a black and white photograph of a man/woman",
  "a pointillism painting of a man/woman",
  "a Japanese woodblock print of a man/woman",
  "a street art stencil of a man/woman".
- Context:
  "a man/woman in the jungle",
  "a man/woman in the snow",
  "a man/woman on the beach",
  "a man/woman on a cobblestone street",
  "a man/woman on top of pink fabric",
  "a man/woman on top of a wooden floor",
  "a man/woman with a city in the background",
  "a man/woman with a mountain in the background",
  "a man/woman with a blue house in the background",
  "a man/woman on top of a purple rug in a forest".
- Action:
  "a man/woman riding a horse",
  "a man/woman holding a glass of wine",
  "a man/woman holding a piece of cake",
  "a man/woman giving a lecture",
  "a man/woman reading a book",
  "a man/woman gardening in the backyard",
  "a man/woman cooking a meal",
  "a man/woman working out at the gym",
  "a man/woman walking the dog",
  "a man/woman baking cookies".

Prompts for Personalized Face with HOI Generation. We select 30 human-object-interactions from V-COCO [9] dataset and format them as" a man/woman" + "[verb]-ing"+ object name for personalized face with HOI generation:
"a man/woman surfing with a surfboard",
"a man/woman skateboarding with a skateboard",
"a man/woman jumping with a skateboard",
"a man/woman snowboarding with a snowboard",
"a man/woman sitting on a chair",
"a man/woman skiing with skis",
"a man/woman working on a laptop",
"a man/woman catching a frisbee",
"a man/woman carrying a suitcase",
"a man/woman talking on a cell phone",
"a man/woman hitting a sports ball",
"a man/woman cutting a cake",
"a man/woman riding a motorcycle",
"a man/woman riding a horse",
"a man/woman sitting on a bench",
"a man/woman eating pizza",
"a man/woman reading a book",
"a man/woman holding a cat",
"a man/woman drinking with a cup",
"a man/woman holding a toothbrush",
"a man/woman holding a teddy bear",
"a man/woman looking at a tv",
"a man/woman holding an umbrella",
"a man/woman laying on a bed",
"a man/woman looking at a dog",
"a man/woman carrying a book",
"a man/woman kicking a sports ball",
"a man/woman throwing a frisbee",
"a man/woman cutting with scissors",
"a man/woman riding a car".