Foundation Cures Personalization: Improving Personalized Models' Prompt Consistency via Hidden Foundation Knowledge

Yiyang Cai[1], **Zhengkai Jiang**[2], **Yulong Liu**[1], **Chunyang Jiang**[1]
Wei Xue[1], **Yike Guo**[1], **Wenhan Luo**[1*]
[1] Hong Kong University of Science and Technology (HKUST)
[2] Tencent Hunyuan
https://yiyangcai.github.io/freecure-aigc.github.io/

Abstract

Facial personalization faces challenges to maintain identity fidelity without disrupting the foundation model's prompt consistency. The mainstream personalization models employ identity embedding to integrate identity information within the attention mechanisms. However, our preliminary findings reveal that identity embeddings compromise the effectiveness of other tokens in the prompt, thereby limiting high prompt consistency and attribute-level controllability. Moreover, by deactivating identity embedding, personalization models still demonstrate the underlying foundation models' ability to control facial attributes precisely. It suggests that such foundation models' knowledge can be leveraged to **cure** the ill-aligned prompt consistency of personalization models. Building upon these insights, we propose **FreeCure**, a framework that improves the prompt consistency of personalization models with their latent foundation models' knowledge. First, by setting a dual inference paradigm with/without identity embedding, we identify attributes (*e.g.*, hair, accessories, etc.) for enhancements. Second, we introduce a novel foundation-aware self-attention module, coupled with an inversion-based process to bring well-aligned attribute information to the personalization process. Our approach is **training-free**, and can effectively enhance a wide array of facial attributes; and it can be seamlessly integrated into existing popular personalization models based on both Stable Diffusion and FLUX. FreeCure has consistently shown significant improvements in prompt consistency across these facial personalization models while maintaining the integrity of their original identity fidelity.

1 Introduction

Human face-centric personalization represents compelling downstream applications of art creation, advertising, and entertainment in the realm of text-to-image synthesis [23, 50]. Given a limited number of images that depict particular identities, facial personalization models generate novel content that reflects these identities through diverse conditions [14, 46, 69, 38]. However, this aspiration is hindered by a persistent challenge: the necessity to maintain high fidelity to the identity while ensuring the controllability of the generated content, also referred to as prompt consistency [72, 26]. This challenge is pronounced in facial personalization since any imperfections or misalignment in the generated faces are particularly salient to humans. Therefore, compared to common object personalization, human face-centric personalization mandates dedicated attention and research efforts.

*Corresponding author

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

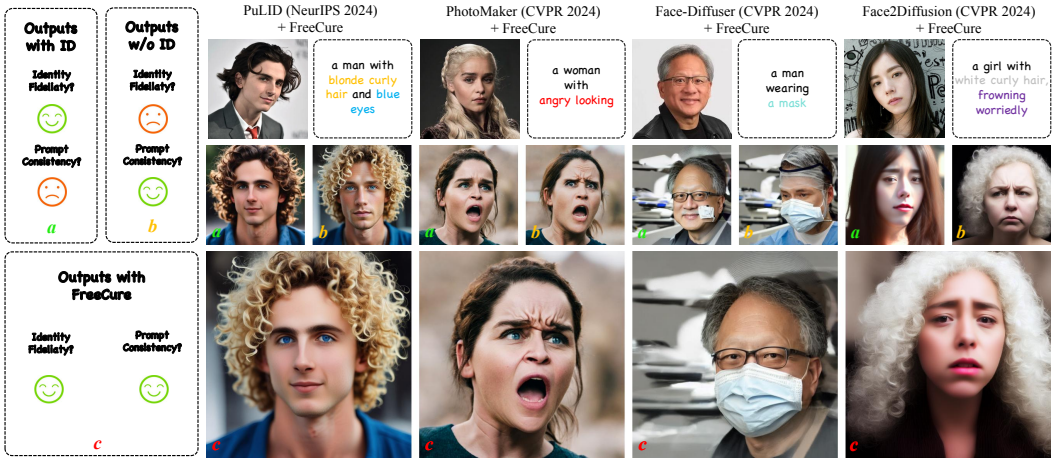


Figure 1: Personalization models (a) demonstrate strong capability in preserving identity fidelity, albeit at the cost of reduced prompt consistency. A prevalent feature in most personalization models is that when their identity embedding inputs are deactivated, they regain the ability to exhibit highly accurate prompt consistency with respect to facial attributes (b), a characteristic closely aligned to their foundation models. Our proposed ***FreeCure*** effectively leverages the latent foundational knowledge inherent in personalized models, enhancing prompt consistency in scenarios involving complex facial attribute control while preserving the identity fidelity (c).

Previous research of facial personalization aims to integrate identity-specific information into the cross-attention mechanisms, with either fine-tuning based strategy [46, 14] or tuning-free paradigm with an identity encoder [64, 34, 18, 59, 27]. However, the dual objectives of maintaining prompt consistency and identity fidelity remain unresolved. Despite the aforementioned advancements, state-of-the-art personalization techniques still struggle to enhance identity fidelity without sacrificing prompt consistency.

Our preliminary experiments indicate that, under identical experimental conditions (*e.g.*, prompts and initial random noise), personalization models' outputs exhibit a significant decline in prompt consistency compared to generated results without identity embedding. For instance, as shown in the first two examples in Fig.1, personalization models fail to generate *"blonde curly hair"* and *"angry"* faces accurately, whereas their counterparts without identity embedding can handle these prompts in a highly faithful manner. Plus, when handling more complex prompts that consist of multiple attributes, the personalization model generates even poorer results. These findings show a fact overlooked by most previous works: while personalization models show their degradation in prompt consistency, the ability of their original foundation models **is unharmed but overridden**. Based on this observation, we further find that identity embeddings can significantly undermine prompt consistency by impeding normal representation of other attribute-related tokens within the prompt through cross-attention mechanisms. This, in turn, adversely affects their effective expression in the latent space of the U-Net. However, given the compelling zero-shot identity extraction capability of identity embedding, directly modifying personalization models' well-trained cross-attention modules can destroy their ability to capture precise identity information. Therefore, our core pursuit is clear: ***Is it feasible to mitigate the erosion of prompt consistency in personalization models while keeping their trained cross-attention modules unaffected?***

Motivated by this objective, we propose FreeCure, a framework enhancing the prompt consistency of personalization models through the guidance of their latent foundation knowledge. While keeping cross-attention modules intact, we propose a novel foundation-aware self-attention (FASA), enabling attributes with high prompt consistency to replace those that are ill-aligned during personalization generation. To protect the identity unharmed, this strategy also leverages semantic segmentation models to generate the scaling masks of these attributes, therefore making such replacement happen in a highly localized and harmonious manner. Furthermore, we use a simple but effective approach called asymmetric prompt guidance (APG) to restore abstract attributes such as expression. Through comprehensive experiments, FreeCure has been verified to effectively restore a wide variety of misaligned attributes produced by various state-of-the-art personalization models.

In summary, the contributions of our paper are threefold:

- We identify the limitations in prompt consistency prevalent across existing face-centric personalization models. Building upon this, we explore the negative effects of identity embedding and elucidate the fundamental reasons for its adverse impacts.
- We propose a training-free framework that leverages high prompt consistency information from foundation models to enhance multiple weakened attributes generated by personalization models. The enhancement of various attributes is achieved without mutual interference.
- Our framework can be seamlessly integrated into widely used personalization models and diverse foundational models, including Stable Diffusion and FLUX. Comprehensive experiments show that our approach improves prompt consistency while maintaining the well-trained ability for identity preservation.

2 Related Work

Identity-preserving Generation. Identity-preserving generation can be broadly categorized into fine-tuning-based and encoder-based approaches. Fine-tuning-based methods [14, 57, 3, 19, 73, 12, 67, 40] either optimize a vector within a textual embedding to encode identity, modify specific weights in the model [46, 31, 7, 25, 4, 17, 47, 52] or use LoRA techniques [24]. However, these methods require training a distinct model for each identity, which compromises scalability and increases susceptibility to overfitting. In contrast, encoder-based methods utilize large-scale pretraining to automatically derive identity representations aligning with textual embeddings, enabling zero-shot personalization with novel identity references. Some methods [63, 9, 64, 49, 62, 34, 43] only train a mapping network or modify cross-attention weights to embed identity information, while others [65, 15, 60, 36, 59, 56, 33, 71, 10] incorporate cross-attention adapters during training. Recent works [18, 27] also leverages Diffusion Transformers (DiT) [1, 42, 13] to reach more impressive performance. In this work, we focus primarily on encoder-based methods, as they represent the state-of-the-art paradigm for personalization and demonstrate particular efficacy in facial performance.

Attention in Diffusion Models. Attention mechanisms serve as foundational components in text-to-image diffusion models. Prior research [20, 16, 74, 30, 55, 6, 35, 29, 37, 41] has leveraged semantic information from cross-attention maps to facilitate object-level editing, while other methods [54, 2, 39, 8, 61, 21, 11, 53, 51, 58] have employed spatial features from self-attention layers to achieve more precise modifications or style transfer. However, the positional information derived from attention maps are inherently constrained to object-level manipulation, making them less robust for fine-grained facial attribute generation. Plus, the function of face-centric embeddings within attention mechanisms remains insufficiently explored, especially in facial personalization models.

3 Revisit ID Embedding in Personalization

We conduct a comprehensive analysis to justify the limitations of current personalization methods in maintaining prompt consistency. This investigation highlights the challenges inherent in existing approaches provides a foundational basis and critical insights for our proposed methodology.

3.1 Dual Denoising to Study Prompt Consistency

To elucidate current personalization methods' limitations, we devise a comparative experiment as an initial exploration. We maintain a fixed noisy latent code z_T and implement two parallel denoising procedures [5]. The only difference is that one denoising procedure incorporates textual embeddings c and identity embedding c_{id}, while the other set c_{id} into a zero tensor $\tilde{c_{id}}$. For clarity, we will refer to the denoising procedure with identity embedding (personalization denoising) as **PD**, and the denoising process that excludes identity embedding (foundation denoising) as **FD** hereafter:

$$\textbf{PD}: \epsilon_p = \epsilon_\theta(z_t, t, c, c_{id}); \textbf{FD}: \epsilon_f = \epsilon_\theta(z_t, t, c, \tilde{c_{id}}) \tag{1}$$

where ϵ_p and ϵ_f denote the predicted noise from PD and FD, respectively. We have conducted these experiments using two facial personalization methods [64, 34] and results are presented in Column

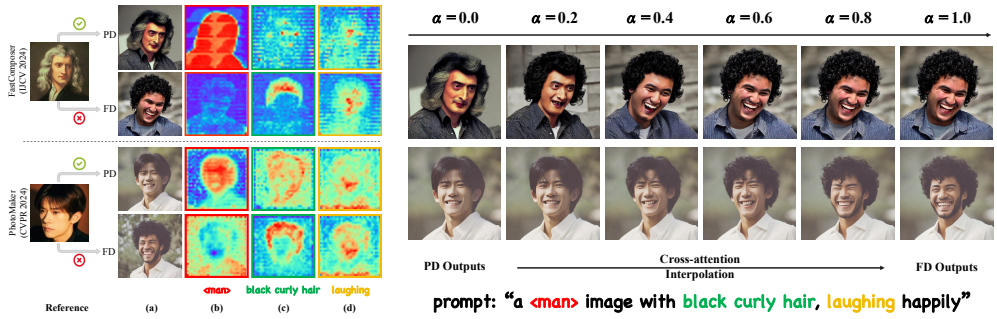


Figure 2: **Analysis on cross-attention maps of facial personalization models. Left**: token-wise attention map visualization. **Right**: interpolation experiment on PD and FD's cross-attention maps.

(a) of Fig.2. It is evident that the output from PD exhibits a reduced ability to match the specific facial attributes compared to the counterpart without identity embedding. For instance: 1) The hair features do not align with the prompt's specified "*black curly*" as expected; 2) The expression "*laughing*" is also restricted. Conversely, the results from FD show greater prompt consistency. These findings indicate that current identity embedding methods may not effectively address "the balance between identity fidelity and prompt consistency".

3.2 ID Embedding's Effect on Attention Layers

OBSERVATION 1: ID embeddings disrupt cross-attention layers. To investigate the underlying causes of this phenomenon, we conduct a visualization experiment inspired by [20, 2, 54], because identity embedding is primarily fused into cross-attention layers. Columns (b-d) of Fig.2's left part present the visualization results of cross-attention maps for both PD and FD processes, focusing on identity embeddings and the attribute-specific token embeddings for "*black curly hair*" and "*laughing*". It reveals that the identity embedding significantly amplifies activation in facial regions while simultaneously reducing activation for other tokens and disrupting their typical interactions within the cross-attention mechanism. Additionally, we observe that during the FD process, prompt consistency remains preserved. This observation underscores an insight: although most personalization models are well fine-tuned, their intrinsic capability to generate faces with high prompt consistency is still preserved but unexpectedly overridden by external identity embeddings. Notably, this latent capability can be effectively activated via the FD procedure, as its name "Foundation Denoising" indicates.

OBSERVATION 2: Personalized cross-attention layers are highly susceptible. We conduct another experiment on cross-attention layers, which involves incorporating a portion of the cross-attention maps from the FD process into those of the PD process. The formula for this approach is

$$A^p_{:,:,m} \leftarrow Softmax(\alpha A^f_{:,:,n} + (1-\alpha)A^p_{:,:,m}), \tag{2}$$

where $A^p_{:,:,m}$ represents the cross-attention map of PD's identity embedding, and $A^f_{:,:,n}$ represents the cross-attention map of the FD's zero-valued embedding. The parameter α controls the weight of the FD's cross-attention map that is injected. The results are shown on the right part of Fig.2. It is observed that, as the weight of the map from the FD process increases, the model quickly loses identity fidelity, even though it regains the facial attributes that were missing before. This experiment demonstrates that cross-attention layers are rather susceptible. To preserve the models' capacity for identity preservation, it is better to leave these well-trained cross-attention modules unaltered.

Our preliminary findings demonstrate that identity embeddings in cross-attention layers are disrupting personalization models' prompt consistency. In contrast, its strong ability of identity extraction makes it rather challenging to be modified. Given this "dilemma", we aim to explore a new approach from the perspective of self-attention, inspired by [16, 52]. Since most personalization models introduce minimal modifications to self-attention layers, it is reasonable to assume that the aforementioned hidden foundation knowledge is preserved within them. By enhancing the self-attention layers in personalization models while keeping cross-attention layers intact, we anticipate achieving better alignment of facial attributes in personalized generation.

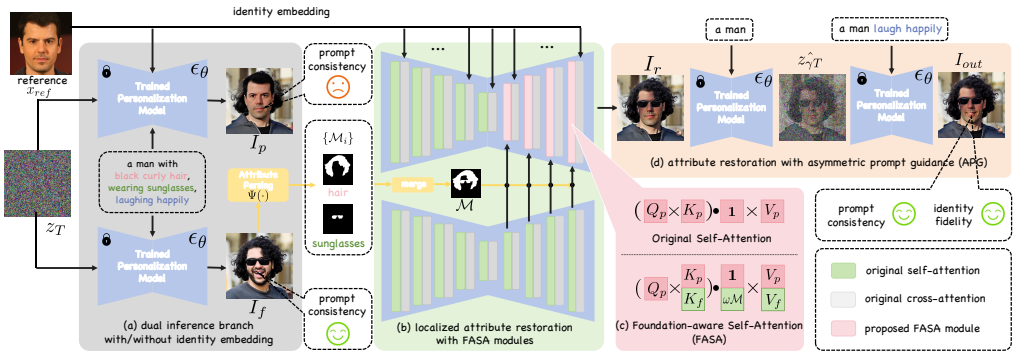


Figure 3: **Overview of FreeCure.** For a personalization model ϵ_θ, we first introduce **(a)**: dual inference paradigm to generate faces with/without identity (I_p and I_f), where I_f without identity embedding shows better prompt consistency. Next, we leverage a segmentation model $\Psi(\cdot)$ to derive related masks of target attributes with clear spatial information (hair, sunglasses, etc.) and merge them into a mask $\mathcal{M}$. In **(b)**: we modify the original self-attention modules with our proposed FASA **(c)**, which concatenates key and value matrices of FD process and PD process, together with a scaling mask to achieve the attribute injection. Finally, we utilize a simple yet effective strategy **(d)**: asymmetric prompt guidance (APG) to restore abstract attributes (*e.g.*, expressions).

4 Methodology

Building upon our investigation in Sec.3, we propose FreeCure, a training-free framework designed to improve the prompt consistency of facial personalization models (see Fig.3). Initially, we develop a foundation-aware self-attention (FASA) mechanism to integrate localized attributes from the FD process into the PD process (Sec.4.1). Subsequently, we apply an asymmetric prompt guidance (APG) strategy to reconstruct more abstract attributes (Sec.4.2).

4.1 Foundation-Aware Self-Attention

Given a reference image x_{ref} that provides the target identity, a well-trained personalization model ϵ_θ, a user-defined prompt c that contains a sequence of facial attributes $\mathcal{A} = \{A_1, A_2, \cdots A_n\}$, our goal is to employ knowledge in FD to enhance prompt consistency of PD's output.

Obtain Masks of Spatial Localized Attributes. As shown in Fig.3 (a), we adapt the dual inference branches which are identical to the FD/PD process introduced in Sec.3 for personalization models to generate the personalized face I_p with unsatisfactory prompt consistency as well as the foundation face I_f with high prompt consistency. Next, we utilize an external face parsing model, denoted as $\Psi(\cdot)$, to extract the binary mask M_i corresponding to A_i from the foundation outputs I_f. Generally, when restoring multiple attributes, we simply need to compute the different masks respectively and merge them $\mathcal{M} = \bigcup\{M_i\}$. $\mathcal{M}$ contains the spatial information of attributes that align with the target prompt, which will play an important role in the attribute restoration process that will be mentioned in the next part.

Figure 4: **Fine-grained attribute enhancement via masks.** Extracting masks from the FD results makes the FASA module only focus on enhancement for target attributes, minimizing its negative effect on identity fidelity.

FASA Mechanism. FASA is the core module that links the information between FD and PD processes and enables personalization models to restore the attributes via their foundation knowledge. As shown in Fig.3(c), we at first identify the self-attention modules in the model. Specifically, in each timestep t and attention layer l, we denote PD and FD's key, query, and value matrices as $\mathcal{KQV}_p^{tl} = \{K_p^{tl}, Q_p^{tl}, V_p^{tl}\}$ and $\mathcal{KQV}_f^{tl} = \{K_f^{tl}, Q_f^{tl}, V_f^{tl}\}$. Second, we concatenate the key and

value matrices of FD process to those of PD process: $\hat{K^{tl}} = [K_p^{tl}, K_f^{tl}]$, $\hat{V^{tl}} = [V_p^{tl}, V_f^{tl}]$. Thus, by omitting the labels of t and l for simplicity, the operation of FASA is

$$\text{FASA}(\mathcal{KQV}_p, \mathcal{KQV}_f) = \text{Softmax}(\frac{Q_p\hat{K}^T}{\sqrt{d}})\hat{V}. \tag{3}$$

Fine-grained Restoration with Scaling Masks. The approach described in Eq.3 allows the PD process to obtain information from the FD process. However, we observe that the method also retains a substantial degree of unrelated features, leading to a notable loss of identity fidelity, as shown in Fig.4. To constrain attribute restoration in a specific region without disrupting identity information, we apply pre-calculated masks $\mathcal{M}$ of different attributes to the similarity map. Additionally, to further enhance FASA's performance, we introduce an additional scaling factor, denoted as ω, to control the magnitude of injecting attribute features from FD to PD. Therefore, the enhanced FASA mechanism can be written as

$$\text{FASA}(\mathcal{KQV}_p, \mathcal{KQV}_f) = \text{Softmax}(\frac{[\mathbf{1}, \omega\mathcal{M}] \odot Q_p\hat{K}^T}{\sqrt{d}})\hat{V}. \tag{4}$$

Where $\mathbf{1}$ denotes a matrix with all elements equal to 1, designed to preserve attention to the original features within the PD process. $\odot$ represents the Hadamard product.

In the full-attention layers of Diffusion Transformers (DiT) such as FLUX, visual and textual information is integrated into a unified sequence representation $[X; C]$. Within this architecture, the FASA mechanism operates in a similar manner, with a key distinction: the attribute mask is applied solely to the component derived from the visual queries of the PD branch (Q_p^X) and the visual keys of the FD branch (K_f^X). This selective masking strategy preserves the original cross-modal attention patterns between visual and textual elements, resembling OminiControl [51]:

$$\text{FASA}_{flux}(\mathcal{KQV}_p, \mathcal{KQV}_f) = \text{Softmax}(\frac{\mathcal{M}(\omega)_{flux} \odot Q_p\hat{K}^T}{\sqrt{d}})\hat{V}, \tag{5}$$

$$\mathcal{M}(\omega)_{flux} = \begin{pmatrix} \mathbf{1}_{l_1\times l_1} & \mathbf{1}_{l_1\times l_2} & \omega\mathcal{M}_{l_1\times l_1} \\ \mathbf{1}_{l_2\times l_1} & \mathbf{1}_{l_2\times l_2} & \mathbf{0}_{l_2\times l_1} \end{pmatrix} \tag{6}$$

Where $Q_p = [Q_p^X; Q_p^C]$ is the original query matrix of PD branch, and $\hat{K} = [K_p^X; K_p^C; K_f^X]$, $\hat{V} = [V_p^X; V_p^C; V_f^X]$ are FASA's enhanced key and value matrices which integrates information from both PD and FD branches. With such enhancement, attributes of the FD process can be successfully extracted and precisely injected into the PD process without disrupting the personalization model's ability of maintaining identity fidelity.

4.2 Asymmetric Prompt Guidance

Following attributes with clear location restored via FASA, to enhance more abstract attributes such as expressions, we introduce a simple yet effective method called *asymmetric prompt guidance* (APG). This strategy is based on the diffusion model's inversion [50]. During the inversion phase, the model accepts only a template prompt (e.g. *"a man"*) that does not include any target attributes. In the denoising process, we add the target attributes' tokens back to this template prompt (e.g. *"a man"* $\rightarrow$ *"a man laughing"*). By leveraging the pretrained controllability of the foundation model, this approach enhances such attributes, resulting in the final refined image I_{out}. Throughout the denoising process, we use only pure textual prompts without identity embeddings, thereby avoiding their potential influence on the tokens related to the target attributes, as discussed in Sec.3. Furthermore, to better preserve the identity, we start the denoising process directly from an intermediate latent code $z_{\hat{\gamma T}}, \gamma \in [0, 1]$, where the high-level identity information has already been established.

5 Experiments

Evaluation Datasets. We collect an extensive dataset including 50 identities, with 30 derived from the CelebA-HQ [32] and the other 20 non-celebrity identities curated by our team. Each identity is represented by a single image, with a spectrum of facial characteristics. The prompt set consists of 20 prompts containing different facial attributes. For each (*identity, prompt*) pair, we produce 20 images. Detailed information can be found in Appendix.A.2.

Baselines. We evaluate FreeCure using several representative facial personalization methods: FastComposer [64], Face-diffuser [62], Face2Diffusion [49], InstantID [59], PhotoMaker [34], PuLID [18], and InfiniteYou [27]. Among these, Face-diffuser, Face2Diffusion, and FastComposer are implemented using SDv1-5 [45], whereas InstantID, PhotoMaker, and PuLID are based on SD-XL [44]. PuLID and InfiniteYou employ FLUX.1-dev [1] as their foundation models.

Evaluation Metrics. We adopt CLIP-T [22] to calculate prompt consistency (**PC**). To calculate identity fidelity (**IF**), we use MTCNN [68] and FaceNet [48] to extract the embedding of the generated/reference faces and compute the cosine similarity. Following PhotoMaker, we adopt the face diversity (**Face Div.**) metric which calculates LPIPS [70] between facial areas. Lastly, following Face2Diffusion, we adopt **PC** $\times$ **IF** score, since it reflects the overall balance of prompt consistency and identity fidelity. We compute the harmonic mean (hMean) of PC and IF.

FreeCure Settings. We set ω in FASA to 2.0, to ensure that attribute information from FD can be sufficiently integrated into PD. The γ in APG is set to 0.5 to maintain the balance between identity fidelity and prompt consistency. For attribute segmentation, we leverage BiSeNet [66] and Segment-Anything [28] for different facial attributes.

5.1 Main Results

Table 1: **Main quantitative evaluation results.** With FreeCure, the mainstream personalization models' prompt consistency is highly enhanced on critical quantitative metrics.

Method	PC(%) ↑	IF(%) ↑	Face Div. (%) ↑	PC × IF (hMean) ↑
FastComposer	18.14	**43.19**	38.92	25.55
FastComposer + **FreeCure**	**21.02** (+15.91%)	41.02 (-5.02%)	**41.01**(+5.37%)	**27.80** (+8.82%)
Face-Diffuser	20.67	**58.34**	40.82	30.52
Face-Diffuser + **FreeCure**	**22.48** (+8.76%)	57.51 (-1.42%)	**41.95**(+2.77%)	**32.32** (+5.90%)
Face2Diffusion	21.92	**39.98**	43.51	28.31
Face2Diffusion + **FreeCure**	**23.26** (+6.12%)	39.23 (-1.88%)	**44.29**(+1.79%)	**29.20** (+3.15%)
InstantID	21.89	**63.94**	48.98	32.61
InstantID + **FreeCure**	**23.62** (+7.90%)	62.01(-3.02%)	**51.82** (+5.80%)	**34.21** (+4.91%)
PhotoMaker	23.04	**51.84**	47.29	31.90
PhotoMaker + **FreeCure**	**24.91** (+8.11%)	50.15 (-3.26%)	**48.52** (+2.60%)	**33.28** (+4.34%)
PuLID (SDXL)	25.16	**58.23**	42.12	35.14
PuLID (SDXL) + **FreeCure**	**26.05** (+3.55%)	56.95 (-2.20%)	**43.72** (+3.80%)	**35.74** (+1.74%)
PuLID (FLUX)	22.42	**74.97**	43.91	34.52
PuLID (FLUX) + **FreeCure**	**24.78** (+10.53%)	72.61 (-3.15%)	**46.09** (+4.96%)	**36.95** (+7.04%)
InfiniteYou	23.77	**79.71**	44.28	36.62
InfiniteYou + **FreeCure**	**25.25** (+6.23%)	77.13 (-3.24%)	**46.82** (+5.74%)	**38.05** (+3.90%)

Overall Performance Comparison. Table 1 and Fig. 5 & 6 present a comparative analysis of our proposed method with the baselines, examining both quantitative and qualitative aspects. It is easy to notice that baselines often fail to accurately reflect key facial attributes mentioned in the prompts. For instance, these baselines often generate faces with identical expression (row 2, 4 of Fig.5 and row 2 of Fig.6). For attributes with large areas (e.g., hair, sunglasses), baselines often cannot generate them harmoniously (row 3, 4 and 5 of Fig.5 and row 1 of Fig.6). Conversely, our approach shows a remarkable ability to enhance absent or faint attributes, significantly improving prompt consistency for these baselines. FreeCure can even tackle subtle facial attributes (e.g., eye color and earrings, see row 5, 6 of Fig.5). Notably, FreeCure achieves significant performance improvements across all foundation models' personalization methods, demonstrating its strong generalizability. Secondly, we notice that FreeCure leads to a slight decline in identity fidelity, which can be attributed to positively improved facial diversity. Since baselines tend to produce faces closely resembling their references, they inherently score higher in identity fidelity. Ultimately, in terms of the PC $\times$ IF metric, our method also shows considerable improvement over all baselines. In conclusion, both quantitatively and qualitatively, FreeCure demonstrates a positive balance by enhancing prompt consistency while keeping the reduction in identity fidelity to a minimum. Additional comparisons and results on more references (including non-celebrities) are available in Appendix. A.3.1. We also

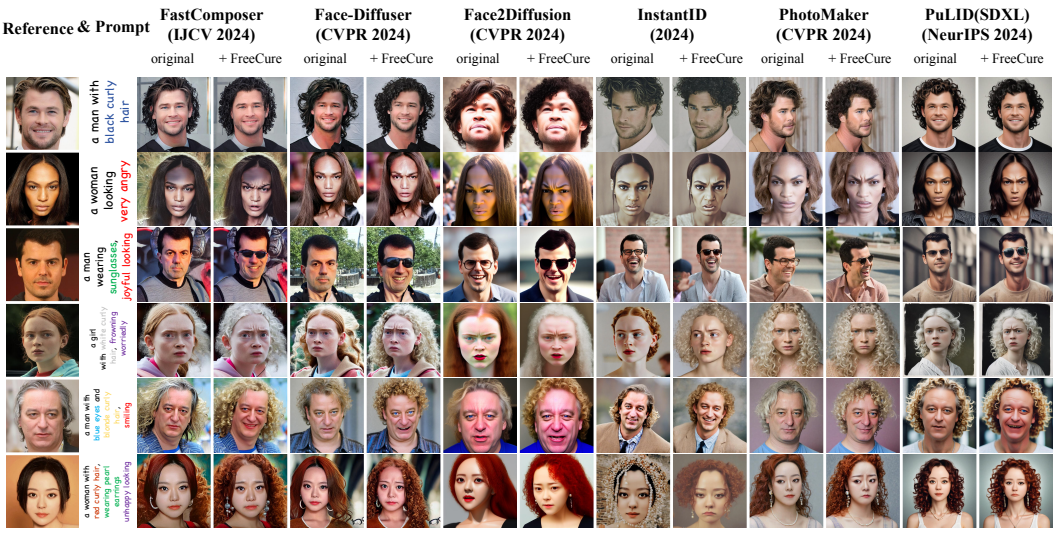


Figure 5: **Qualitative comparison with facial personalization baselines (including baselines built upon SDv1.5 and SDXL).** Different attributes in prompts are highlighted in various colors. Comparison of corresponding FD outputs is provided in the Appendix.A.3.2.

Figure 6: **Qualitative comparison with FLUX-based facial personalization baselines.** Different attributes in prompts are highlighted in various colors.

provide corresponding FD outputs for Fig.5 to show the effectiveness of FreeCure to transfer correct attributes from the FD to PD process in Appendix. A.3.2

Prompt Consistency with Multiple Attributes. Prompts involving multiple facial attributes pose a greater challenge to personalization's prompt consistency but better reflect practical user needs. Table 2 illustrates FreeCure's improvements on baselines with prompts including multiple facial attributes. As the number of attributes increases, the PC values of baselines tend to decrease. In contrast, FreeCure's improvement in PC becomes more significant as the complexity of the prompt increases. More analysis of FreeCure's non-disruption manner in multiple prompt personalization is available in Appendix.A.4.1. In summary, through integration with FreeCure, personalization models can effectively address more complex and realistic prompt instructions.

5.2 Robustness Justification

Robust Performance with Different Initial Noises. We observe that, with identical initial noise, all baselines' FD and PD processes generate faces with similar attribute locations. It is important to validate that FASA's robust performance without these condition. Thus, we relax this condition and regenerate faces with different initial noises. Fig.7 shows that results under the two settings are comparable, which confirms that FASA can effectively enhance the generated results of PD, even when its FD counterpart produces faces with variable spatial structures.

Visualization of FASA maps. Fig.8 visualizes the FASA map $A \in \mathbb{R}^{H\times(2\times W)}$, whose size is doubled according to Eq.4. For a given query point q_i in Q_p, its corresponding scores $A_{i,:}$ are extracted from both K_p and K_f. When q_i falls into areas associated with target attributes (*e.g.*,

Table 2: **Quantitative comparison of prompt consistency with different number of attributes.** After calculating the metrics for each method, we compute their mean values based on the corresponding foundation model type. For metrics of each baseline, please refer to Appendix.A.3.3.

Foundation Model	PC (1 Attr.) ↑	PC (2 Attr.) ↑	PC (3 Attr.) ↑
SDv1.5	21.01	20.34	18.49
SDv1.5 + **FreeCure**	**22.70** (+8.04%)	**22.34** (+9.87%)	**21.16** (+14.44%)
SDXL	23.83	23.31	22.65
SDXL + **FreeCure**	**25.11** (+5.34%)	**24.80** (+6.36%)	**24.49** (+8.14%)
FLUX	24.15	22.64	21.88
FLUX + **FreeCure**	**25.75** (+6.63%)	**24.71** (+9.15%)	**24.16** (+10.45%)

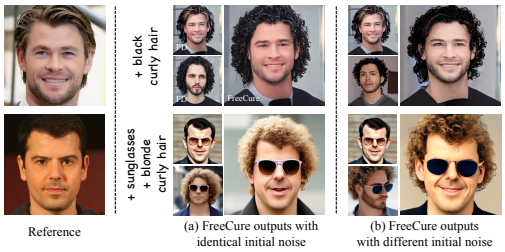


Figure 7: **Performance of FASA w/ and w/o identical initial noises**. FASA can precisely enhance attributes even if PD and FD produce faces with different locations, sizes, and angles.

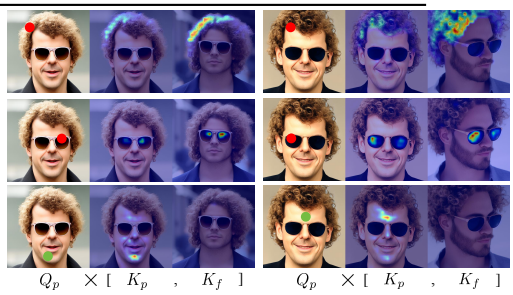


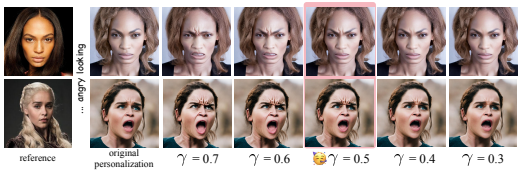
Figure 8: **Visualization of the FASA maps** for attribute related area (red points) and non-attribute related area (green points).

hair, sunglasses), the FASA map exhibits greater attention to information from K_f, corresponding to FD. Conversely, in regions unrelated to attributes, the FASA map retains its attention on PD. This visualization substantiates the role of FASA in transferring fine-grained attribute information from FD to PD. Additionally, FASA preserves its original attention pattern in regions unrelated to attributes, thereby ensuring the performance of personalization models' identity fidelity. More detailed visualization analyses of FASA are available in Appendix.A.4.6.

5.3 Ablation Study

Overall Analysis of Each Component. Table 3 presents the individual performance of FASA and APG in enhancing attributes across all baselines. Both components demonstrate positive effects compared to the baseline metrics in Table 1, with FASA showing more noticeable improvements. We attribute this to FASA's ability to effectively handle attributes covering larger areas, such as hair and sunglasses, resulting in more observable enhancements.

(a) Ablation of FASA scaling mask strategy.



(b) Ablation of APG intermediate timesteps.

Figure 9: Ablation Studies.

Scaling Mask Strategy. Figure 9a illustrates the evaluation of the scaling mask strategy implemented in FASA. Low scaling values critically hinder FASA's capacity to effectively transfer attribute information to the personalization denoising. On the contrary, high scaling factors may negatively impact the overall quality of the generated faces. Generally, the optimal value of ω should be around 2.0. More ablation studies of FASA are available in Appendix.A.4.5.

Effect of Inversion's Intermediate Timesteps. Figure 9b demonstrates the effects of different starting timesteps in APG, represented by the parameter γ in Section 4.2. To achieve a noticeable improvement in attributes while maintaining identity fidelity, an optimal γ should be set as 0.5.

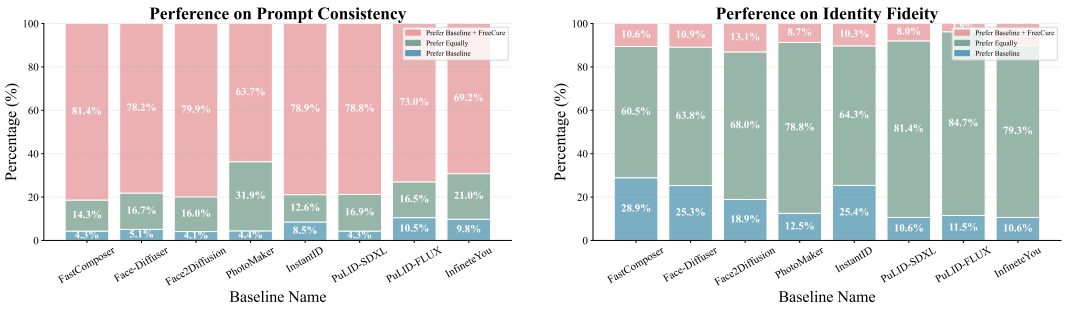
Table 3: **Quantitative ablation analysis for FASA and APG's independent effect.** After calculating the metrics for each method, we compute their mean values based on the corresponding foundation model type. For metrics of each baseline, please refer to Appendix.A.4.4.

Foundation Model	FASA	APG	PC(%) ↑	IF(%) ↑	PC × IF (hMean) ↑
SDv1.5	✓	✗	21.83	46.39	29.69
	✗	✓	20.58	46.81	28.59
SDXL	✓	✗	24.55	56.97	34.31
	✗	✓	23.97	56.87	33.73
FLUX	✓	✗	24.50	75.66	37.01
	✗	✓	24.09	75.85	36.57

Hyperparameter Analysis. Our supplemental ablation analysis indicates that the hyperparameters ω and γ consistently fall within specific ranges across different baselines: $\omega \in [1.8, 2.4], \gamma \in [0.5, 0.6]$. Baselines based on Stable Diffusion v1.5 [49, 64, 62] require larger ω values (up to 2.4), while other methods [16, 18, 27, 59] perform well around $\omega = 2.0$. Since results remain stable within these ranges, we adopt a unified setting ($\omega = 2.0, \gamma = 0.5$) throughout our experiments.

5.4 User Study

To further validate FreeCure's superiority, we conducted an online user study with 30 participants. The study was designed as follows: for each baseline method, we randomly selected 10 samples. Each sample included a reference image, a text prompt, the baseline's output, and the result refined by FreeCure. Participants were asked to evaluate the prompt consistency and identity fidelity of each output. As shown in Fig.10a and Fig.10b, the results demonstrate a clear preference for FreeCure on prompt consistency and equal preference on identity fidelity, indicating that FreeCure's potential negative impact on identity preservation is minimal.



(a) User preference on prompt consistency

(b) User preference on identity fidelity

Figure 10: **User study of FreeCure**. The preference ratio indicate that FreeCure can improve prompt consistency without undermining identity fidelity of different personalization models.

6 Limitation and Conclusion

Limitation. FreeCure is influenced by inherent bias and maximum achievable capabilities of personalization models, which is commonly observed in plug-and-play adaptation frameworks. Furthermore, it struggles with certain transparent objects, such as glass bottles. In some cases, we also observe slight attribute entanglement between FD and PD processes. Furthermore, FreeCure's application to personalization models based on auto-regressive architectures could be further explored.

Conclusion. Facing the challenge that personalization models employing identity embeddings frequently struggle to preserve prompt consistency, we introduce FreeCure, a training-free framework that leverages the high prompt consistency inherent in foundation models to refine the output of personalization models, leading to a remarkable improvement in their prompt consistency and minimal disruption to their original identity fidelity Our experiments validate the effectiveness of FreeCure on popular baselines from different foundation models, particularly in scenarios with complex prompts that encompass multiple attributes.

Acknowledgments

The research was supported in part by Theme-based Research Scheme (T45-205/21-N) from Hong Kong RGC, Generative AI Research, Development Centre from InnoHK, in part by the National Natural Science Foundation of China (Grant No. 62372480), in part by HKUST-MetaX Joint Lab Fund (No. METAX24EG01-D). We also thank Yubai Wei, Moheng Li, and Meixin Zhou for their assistance with the user study.

References

[1] `https://github.com/black-forest-labs/flux`, 2024.

[2] Yuval Alaluf, Daniel Garibi, Or Patashnik, Hadar Averbuch-Elor, and Daniel Cohen-Or. Cross-image attention for zero-shot appearance transfer. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–12, 2024.

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

[4] Moab Arar, Rinon Gal, Yuval Atzmon, Gal Chechik, Daniel Cohen-Or, Ariel Shamir, and Amit H. Bermano. Domain-agnostic tuning-encoder for fast personalization of text-to-image models. In *SIGGRAPH Asia 2023 Conference Papers*, pages 1–10, 2023.

[5] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18392–18402, 2023.

[6] Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22560–22570, 2023.

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

[8] Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM transactions on Graphics (TOG)*, 42(4):1–10, 2023.

[9] Zhuowei Chen, Shancheng Fang, Wei Liu, Qian He, Mengqi Huang, and Zhendong Mao. Dreamidentity: enhanced editability for efficient face-identity preserved image generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 1281–1289, 2024.

[10] Siying Cui, Jia Guo, Xiang An, Jiankang Deng, Yongle Zhao, Xinyu Wei, and Ziyong Feng. Idadapter: Learning mixed features for tuning-free personalization of text-to-image models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 950–959, 2024.

[11] Ganggui Ding, Canyu Zhao, Wen Wang, Zhen Yang, Zide Liu, Hao Chen, and Chunhua Shen. Freecustom: Tuning-free customized image generation for multi-concept composition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9089–9098, 2024.

[12] Ziyi Dong, Pengxu Wei, and Liang Lin. Dreamartist: Towards controllable one-shot text-to-image generation via positive-negative prompt-tuning. *arXiv preprint arXiv:2211.11337*, 2022.

[13] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

[14] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *The Eleventh International Conference on Learning Representations*, 2023.

[15] Rinon Gal, Or Lichter, Elad Richardson, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. Lcm-lookahead for encoder-based text-to-image personalization. In *European Conference on Computer Vision*, pages 322–340. Springer, 2024.

[16] Jing Gu, Yilin Wang, Nanxuan Zhao, Tsu-Jui Fu, Wei Xiong, Qing Liu, Zhifei Zhang, He Zhang, Jianming Zhang, HyunJoon Jung, et al. Photoswap: Personalized subject swapping in images. *Advances in Neural Information Processing Systems*, 36, 2024.

[17] Yuchao Gu, Xintao Wang, Jay Zhangjie Wu, Yujun Shi, Yunpeng Chen, Zihan Fan, Wuyou Xiao, Rui Zhao, Shuning Chang, Weijia Wu, et al. Mix-of-show: Decentralized low-rank adaptation for multi-concept customization of diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024.

[18] Zinan Guo, Yanze Wu, Chen Zhuowei, Lang chen, Peng Zhang, and Qian HE. Pulid: Pure and lightning id customization via contrastive alignment. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 36777–36804. Curran Associates, Inc., 2024.

[19] Ligong Han, Yinxiao Li, Han Zhang, Peyman Milanfar, Dimitris Metaxas, and Feng Yang. Svdiff: Compact parameter space for diffusion fine-tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7323–7334, 2023.

[20] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-or. Prompt-to-prompt image editing with cross-attention control. In *The Eleventh International Conference on Learning Representations*, 2023.

[21] Amir Hertz, Andrey Voynov, Shlomi Fruchter, and Daniel Cohen-Or. Style aligned image generation via shared attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4775–4785, 2024.

[22] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A reference-free evaluation metric for image captioning. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[23] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[24] 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. *arXiv preprint arXiv:2106.09685*, 2021.

[25] Miao Hua, Jiawei Liu, Fei Ding, Wei Liu, Jie Wu, and Qian He. Dreamtuner: Single image is enough for subject-driven generation. *arXiv preprint arXiv:2312.13691*, 2023.

[26] Yi Huang, Jiancheng Huang, Yifan Liu, Mingfu Yan, Jiaxi Lv, Jianzhuang Liu, Wei Xiong, He Zhang, Shifeng Chen, and Liangliang Cao. Diffusion model-based image editing: A survey. *arXiv preprint arXiv:2402.17525*, 2024.

[27] Liming Jiang, Qing Yan, Yumin Jia, Zichuan Liu, Hao Kang, and Xin Lu. Infiniteyou: Flexible photo recrafting while preserving your identity. *arXiv preprint arXiv:2503.16418*, 2025.

[28] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023.

[29] Lingjie Kong, Kai Wu, Xiaobin Hu, Wenhui Han, Jinlong Peng, Chengming Xu, Donghao Luo, Jiangning Zhang, Chengjie Wang, and Yanwei Fu. Anymaker: Zero-shot general object customization via decoupled dual-level id injection. *arXiv preprint arXiv:2406.11643*, 2024.

[30] 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 *European Conference on Computer Vision*, 2024.

[31] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1931–1941, June 2023.

[32] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[33] Xiaoming Li, Xinyu Hou, and Chen Change Loy. When stylegan meets stable diffusion: a w+ adapter for personalized image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2187–2196, 2024.

[34] Zhen Li, Mingdeng Cao, Xintao Wang, Zhongang Qi, Ming-Ming Cheng, and Ying Shan. Photomaker: Customizing realistic human photos via stacked id embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8640–8650, 2024.

[35] Bingyan Liu, Chengyu Wang, Tingfeng Cao, Kui Jia, and Jun Huang. Towards understanding cross and self-attention in stable diffusion for text-guided image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7817–7826, 2024.

[36] Jian Ma, Junhao Liang, Chen Chen, and Haonan Lu. Subject-diffusion: Open domain personalized text-to-image generation without test-time fine-tuning. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–12, 2024.

[37] Yuhang Ma, Wenting Xu, Jiji Tang, Qinfeng Jin, Rongsheng Zhang, Zeng Zhao, Changjie Fan, and Zhipeng Hu. Character-adapter: Prompt-guided region control for high-fidelity character customization. *arXiv preprint arXiv:2406.16537*, 2024.

[38] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 4296–4304, 2024.

[39] Jisu Nam, Heesu Kim, DongJae Lee, Siyoon Jin, Seungryong Kim, and Seunggyu Chang. Dreammatcher: Appearance matching self-attention for semantically-consistent text-to-image personalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8100–8110, 2024.

[40] Lianyu Pang, Jian Yin, Haoran Xie, Qiping Wang, Qing Li, and Xudong Mao. Cross initialization for face personalization of text-to-image models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8393–8403, 2024.

[41] Or Patashnik, Rinon Gal, Daniil Ostashev, Sergey Tulyakov, Kfir Aberman, and Daniel Cohen-Or. Nested attention: Semantic-aware attention values for concept personalization. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pages 1–12, 2025.

[42] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023.

[43] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27080–27090, 2024.

[44] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

[45] 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*, pages 10684–10695, 2022.

[46] 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.

[47] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6527–6536, 2024.

[48] 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*, pages 815–823, 2015.

[49] Kaede Shiohara and Toshihiko Yamasaki. Face2diffusion for fast and editable face personalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6850–6859, 2024.

[50] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[51] Zhenxiong Tan, Songhua Liu, Xingyi Yang, Qiaochu Xue, and Xinchao Wang. Ominicontrol: Minimal and universal control for diffusion transformer. *arXiv preprint arXiv:2411.15098*, 2024.

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

[53] Yoad Tewel, Rinon Gal, Dvir Samuel, Yuval Atzmon, Lior Wolf, and Gal Chechik. Add-it: Training-free object insertion in images with pretrained diffusion models. *arXiv preprint arXiv:2411.07232*, 2024.

[54] Yoad Tewel, Omri Kaduri, Rinon Gal, Yoni Kasten, Lior Wolf, Gal Chechik, and Yuval Atzmon. Training-free consistent text-to-image generation. *ACM Transactions on Graphics (TOG)*, 43(4):1–18, 2024.

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

[56] 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.

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

[58] Kuan-Chieh Wang, Daniil Ostashev, Yuwei Fang, Sergey Tulyakov, and Kfir Aberman. Moa: Mixture-of-attention for subject-context disentanglement in personalized image generation. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–12, 2024.

[59] 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.

[60] Ye Wang, Xuping Xie, Lanjun Wang, Zili Yi, and Rui Ma. Dp-adapter: Dual-pathway adapter for boosting fidelity and text consistency in customizable human image generation. *arXiv preprint arXiv:2502.13999*, 2025.

[61] Yibin Wang, Weizhong Zhang, and Cheng Jin. Magicface: Training-free universal-style human image customized synthesis. *arXiv preprint arXiv:2408.07433*, 2024.

[62] Yibin Wang, Weizhong Zhang, Jianwei Zheng, and Cheng Jin. High-fidelity person-centric subject-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7675–7684, 2024.

[63] Yuxiang Wei, Yabo Zhang, Zhilong Ji, Jinfeng Bai, Lei Zhang, and Wangmeng Zuo. Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15943–15953, 2023.

[64] 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.

[65] 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.

[66] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 325–341, 2018.

[67] Ge Yuan, Xiaodong Cun, Yong Zhang, Maomao Li, Chenyang Qi, Xintao Wang, Ying Shan, and Huicheng Zheng. Inserting anybody in diffusion models via celeb basis. *Advances in Neural Information Processing Systems*, 36, 2024.

[68] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters*, 23(10):1499–1503, 2016.

[69] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3836–3847, 2023.

[70] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

[71] Shilong Zhang, Lianghua Huang, Xi Chen, Yifei Zhang, Zhi-Fan Wu, Yutong Feng, Wei Wang, Yujun Shen, Yu Liu, and Ping Luo. Flashface: Human image personalization with high-fidelity identity preservation. *arXiv preprint arXiv:2403.17008*, 2024.

[72] 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.

[73] Yuxin Zhang, Weiming Dong, Fan Tang, Nisha Huang, Haibin Huang, Chongyang Ma, Tong-Yee Lee, Oliver Deussen, and Changsheng Xu. Prospect: Prompt spectrum for attribute-aware personalization of diffusion models. *ACM Transactions on Graphics (TOG)*, 42(6):1–14, 2023.

[74] Jing Zhao, Heliang Zheng, Chaoyue Wang, Long Lan, and Wenjing Yang. Magicfusion: Boosting text-to-image generation performance by fusing diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22592–22602, October 2023.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: All claims made in the abstract and introduction are supported by results of the experiments. See Sec.5.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: FreeCure can be influenced by original models' inherent biases and maximum achievable capabilities.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.

- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: This paper does not include theoretical results.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We report the experiment details and we will release related codes. See Sec.5.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.

- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Our dataset consists of public part and self-collected part. We will release them together with codes. See Sec.5.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: Our method does not include any training process, and we have specified the hyperparameter of our framework. See Sec.5.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: All evaluation samples are randomly initialized and follows the identical evaluation process.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: We run experiments on one H800 GPU (80GB).

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The paper confirms with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: There is no societal impact of the work performed.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.

- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: This paper cites the original papers, codes, and datasets which are used.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: This paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: The LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of this paper.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

A Appendix

The Appendix consists of three sections:

- Sec.A.1 includes an ethics statement that addresses the privacy of our evaluation datasets and our methods for mitigating potential risks.
- Sec.A.2 provides implementation details, including details of setting up the dual denoising paradigm for every baseline. Details of segmentation models, prompt datasets as well as evaluation metrics are also included.
- Sec.A.3 provides additional results of FreeCure in a boarder range of references, including non-celebrities. It also includes complementary FD outputs corresponding to the results of the main submission (Fig.5, 6), and broader attribute enhancement achieved through the integration of advanced models, such as Segment-Anything Model [28].
- Sec.A.4 provides a comprehensive analysis of FreeCure, including analysis for initial noise conditions, detailed visualization of FASA modules, more detailed ablation studies, runtime analysis and a template of the user study.

A.1 Ethics Statements

We confirm that our human face datasets, featuring both celebrities and non-celebrities, are sourced from public datasets or Google Images, limited to CC-licensed content. We have made extensive efforts to balance critical attributes like gender, race, and age to mitigate significant bias. Furthermore, we acknowledge that FreeCure could be misused for malicious purposes, such as creating impersonating identities. To mitigate these risks, we are committed to ethical governance and a controlled release strategy. For instance, we release the method only to accredited researchers under a license prohibiting misuse, with mandatory ethics training.

A.2 More Implementation Details

A.2.1 Triggering Hidden Foundation Knowledge in Personalization Baselines

The core architecture of FreeCure is based on a dual denoising paradigm of the FD and PD process, which triggers hidden foundation knowledge in personalization models. This paradigm deactivates the identity embeddings which are integrated into the cross-attention layers. In Table 4 we illustrate how we deactivate identity embedding for different baselines mentioned in the main paper since they have various identity embedding fusion strategies (e.g., LoRA, adapters) and prompt templates.

Table 4: **Strategies for implementing dual denoising paradigms on different baselines.** "black curly hair" is used here as an example and can be replaced with other facial attributes.

Method	PD inputs	FD inputs
FastComposer	a <P> <|image|> with black curly hair	a <P> image with black curly hair
Face-Diffuser	a <P> <|image|> with black curly hair	a <P> image with black curly hair
Face2Diffusion	a f l with black curly hair	a <P> image with black curly hair
InstantID	a <P> with black curly hair	a <P> image with black curly hair
PhotoMaker	a <P> img with black curly hair	a <P> image with black curly hair
PuLID	a <P> with black curly hair	a <P> image with black curly hair
InfiniteYou	a <P> with black curly hair	a <P> image with black curly hair

FastComposer [64], Face-Diffuser [62] and PhotoMaker [34]share the similar prompting template, wherein their identity embeddings are integrated into the "<P>" word, which is initialized with class-specific words (*e.g., man, woman, boy, girl*). To establish the FD process, we maintain the original plain text embedding for "<P>" without incorporating any identity information, effectively equivalent to applying a null identity embedding. Face2Diffusion [49] integrated the identity embedding into the trigger word "f". Therefore, we replace the identity embedding with a zero-valued tensor before its

Table 5: **Information for facial attribute and label relationship of BiSeNet.**

label	attribute	label	attribute	label	attribute
0	background	6	glasses	12	lower lip
1	facial skin	7	right ear	13	neck
2	right eyebrow	8	left ear	14	necklace
3	left eyebrow	9	nose	15	cloth
4	left eye	10	earrings	16	hair
5	right eye	11	upper lip	17	hat

Table 6: **Prompts for evaluation categorized by the number of included attributes (range from 1 to 3).** <S> will be replaced with placeholder tokens such as man, woman, boy, etc.

Attribute	Prompt
1	a <S> with black curly hair a <S> with blonde curly hair a <S> with red long straight hair a <S> with very angry looking a <S> with frowning worriedly a <S> laughing happily a <S> wearing a mask a <S> wearing sunglasses
2	a <S> with white curly hair, frowning worriedly a <S> with black curly hair, laughing happily a <S> with blonde curly hair and blue eyes a <S> with blue eyes, laughing happily a <S> wearing sunglasses, laughing happily a <S> with black hair, wearing silver earrings a <S> with blonde hair and blue eyes a <S> with sunglasses, frowning worriedly
3	a <S> with red curly hair, wearing pearl earrings, unhappy looking a <S> with blue eyes and blonde curly hair, smiling a <S> with white curly hair, wearing sunglasses, laughing happily a <S> with black curly hair, wearing silver earrings, frowning worriedly

merging process. For baselines such as InstantID [59], PuLID [18], and InfiniteYou [27] that employ residual cross-attention adapters for identity fusion, we initialize their identity embeddings as zero tensors with matching dimensionality while preserving the original textual embeddings unchanged.

A.2.2 Attribute Segmentation Models

As mentioned in Sec.5, we use two mainstream segmentation models: BiSeNet and Segment-Anything models for attribute-aware mask generation. BiSeNet [2] is a popular framework for face parsing that is capable of generating semantic masks corresponding to facial attributes. Table 5 shows BiSeNet's prediction label and corresponding facial attributes, demonstrating that it can address the majority of facial attributes, thereby verifying the robustness of FreeCure. Additionally, FreeCure is designed to be seamlessly integrated with various segmentation models, not limited to BiSeNet. We have showcased results that utilize Segment-Anything as the face parsing model in Sec.A.3.4.

A.2.3 Facial Prompts for Evaluation

Table 6 introduces the prompts for evaluating enhancement performance for facial attributes. Generally, our prompts include multiple facial attributes, ensuring that previous baselines' weaknesses in prompt consistency can be fully uncovered and highlighting the enhanced performance of FreeCure.

[2]https://github.com/CoinCheung/BiSeNet

A.2.4 Details of Metrics

Prompt Consistency (PC, also known as CLIP-T). We leverage the official implementation of the vision transformer[3] provided by OpenAI.

Identity Fidelity (IF). We use the official implementation of MTCNN + FaceNet pipeline[4] to conduct the processes of face detection and feature extraction from facial regions. Additionally, we compute the cosine similarity between two face embeddings to evaluate their similarity.

Face Diversity (Face Div.). We utilize the official implementation of LPIPS[5] to quantify the perceptual distance between two facial images.

A.3 More Results

A.3.1 More Results on Celebrity & Non-Celebrity References

In Fig.11, 12 and 13, we provide additional experimental results on more reference images to further validate the consistent performance of FreeCure. This robust performance shows the potential of FreeCure to enhance various personalization models in practical applications, as the ability to handle non-celebrity personalization is a critical requirement in real-world scenarios.

[3] https://github.com/openai/CLIP

[4] https://github.com/timesler/facenet-pytorch

[5] https://github.com/richzhang/PerceptualSimilarity

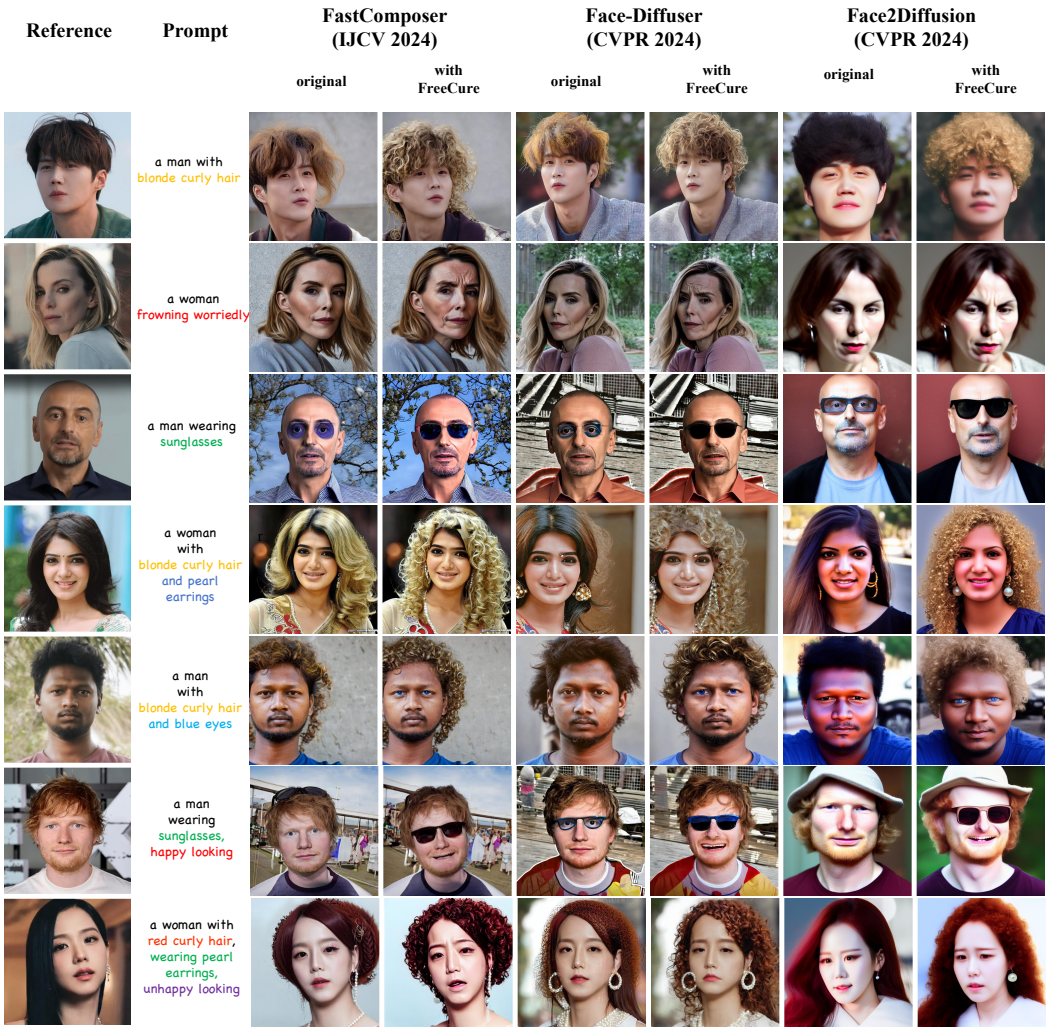


Figure 11: **More results for baselines implemented with SDv1.5, including FastComposer, Face-Diffuser and Face2Diffusion.** We provide personalization results including non-celebrity identities.

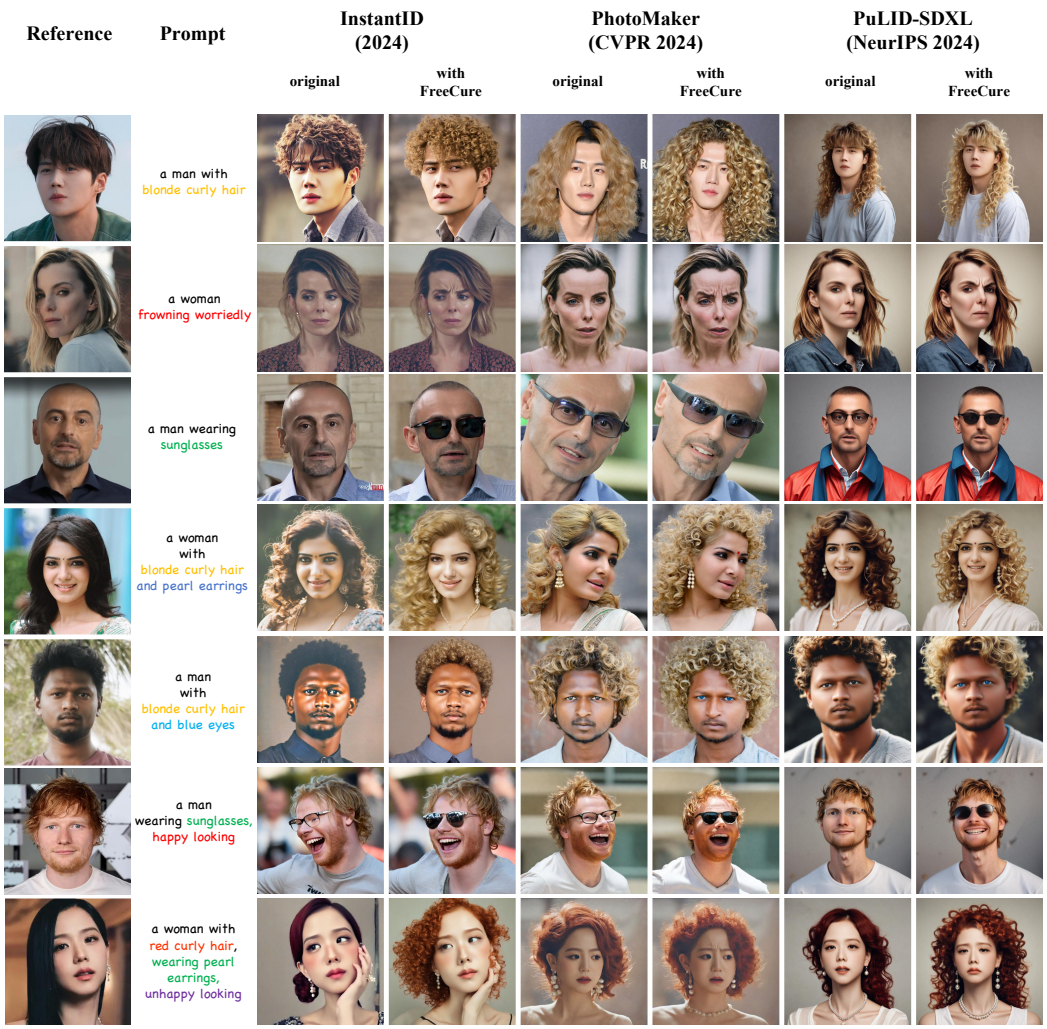


Figure 12: **More results for baselines implemented with SDXL, including InstantID, PhotoMaker and PuLID-SDXL.** We provide personalization results including non-celebrity identities.

Figure 13: **More results for baselines implemented with FLUX, including PuLID-FLUX and InfiniteYou.** We provide personalization results including non-celebrity identities.

A.3.2 Supplementary Comparison of Results for Main Paper

To better demonstrate the efficacy of FreeCure, Fig.14, 15 present a comparative analysis of the results generated by the FD, PD, and enhanced outputs by FreeCure linked to Fig.5. The results indicate that FreeCure faithfully transfers target prompt-aligned attributes from the FD into the PD processes while preserving the original identity information.

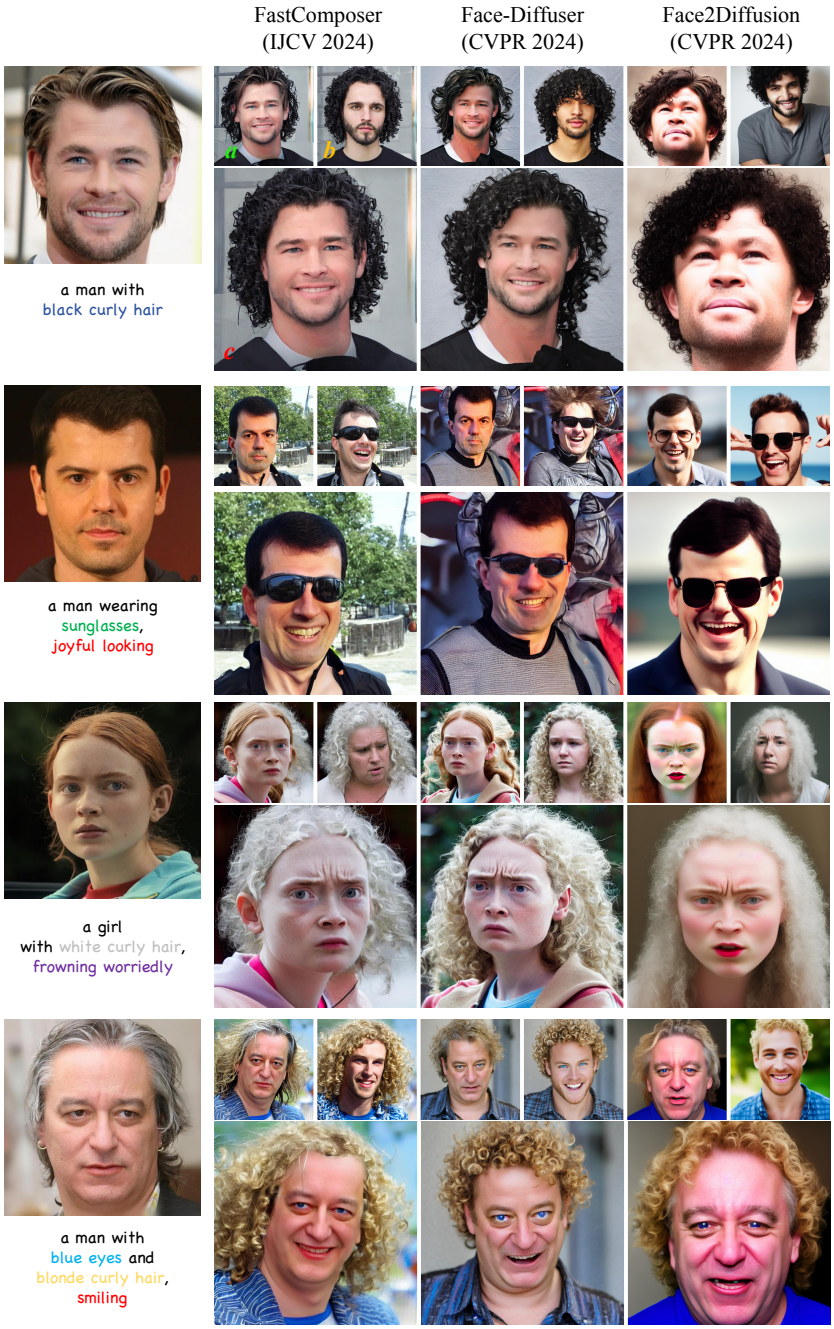


Figure 14: **Corresponding comparative FD & PD outputs for results in the main submission. (FastComposer, Face-Diffuser and FaceDiffusion)** (a) refers to original personalization outputs, (b) refers to foundation outputs, and (c) refers to outputs enhanced with proposed FreeCure.

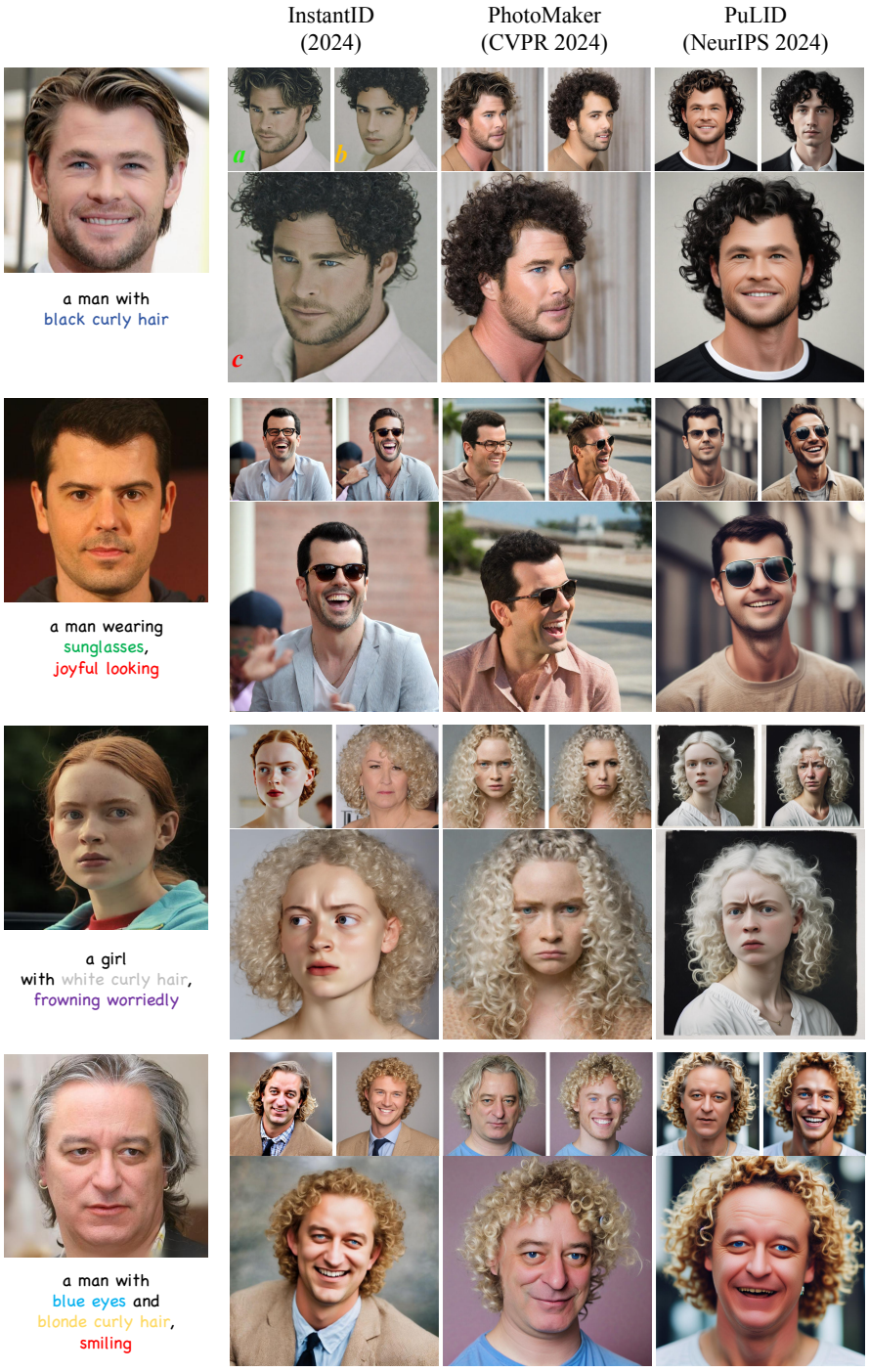


Figure 15: **Corresponding comparative FD & PD outputs for results in the main paper. (InsantID, PhotoMaker and PuLID)** (a) refers to original personalization outputs, (b) refers to foundation outputs, and (c) refers to outputs enhanced with proposed FreeCure.

A.3.3 More Detailed Quantitative Results for Prompt Consistency with Multiple Attributes.

As supplementary results for Table.2, we provide source prompt consistency results of each baseline considering prompts with various attributes in Table.7. It shows that as the prompt becomes increasingly complicated, all baselines are facing more decreasing in prompt consistency, which can limit their application in real scenarios. The results also support the fact that FreeCure's improvement in PC becomes more significant as the complexity of the prompt increases.

Table 7: Quantitative comparison of prompt consistency with different number of attributes.

Method	PC (1 Attr.) ↑	PC (2 Attr.) ↑	PC (3 Attr.) ↑
FastComposer	18.83	18.05	16.92
FastComposer + **FreeCure**	**21.42** (+13.75%)	**21.29** (+17.95%)	**19.69** (+16.37%)
Face-Diffuser	21.68	20.98	18.02
Face-Diffuser + **FreeCure**	**22.89** (+5.58%)	**22.72** (+8.29%)	**21.17** (+17.48%)
Face2Diffusion	22.54	21.98	20.54
Face2Diffusion + **FreeCure**	**23.81** (+5.63%)	**23.02** (+4.73%)	**22.63** (+10.18%)
InstantID	22.43	21.81	20.95
InstantID + **FreeCure**	**23.98** (+6.91%)	**23.52** (+7.84%)	**23.11** (+10.31%)
PhotoMaker	23.51	23.01	22.57
PhotoMaker + **FreeCure**	**25.04** (+6.51%)	**24.91** (+8.26%)	**24.64** (+11.19%)
PuLID (SDXL)	25.56	25.12	24.43
PuLID (SDXL) + **FreeCure**	**26.30** (+3.34%)	**25.96** (+5.32%)	**25.73** (+3.55%)
PuLID (FLUX)	23.29	22.08	21.32
PuLID (FLUX) + **FreeCure**	**25.52** (+9.57%)	**24.49** (+10.91%)	**23.87** (+11.96%)
InfiniteYou	25.01	23.19	22.43
InfiniteYou + **FreeCure**	**25.98** (+3.88%)	**24.92** (+7.46%)	**24.45** (+9.01%)

A.3.4 Integration with Segment-Anything for Attribute Extraction

Fig.16 showcases FreeCure's consistent performance when integrated with more advanced segmentation models such as Segment-Anything. For instance, the attribute **mask** cannot be handled by BiSeNet since its label is absent. However, by replacing BiSeNet with Segment-Anything and specifying the target prompt as "mask," valid semantic masks can be accurately generated, thereby guaranteeing effective attribute enhancement through FASA. Given that Segment-Anything extracts attributes based on flexible prompts, its seamless integration significantly enhances the versatility of FreeCure, enabling it to address a broader range of scenarios.

Figure 16: **Results for FreeCure's integration with Segment-Anything.** FreeCure can seamlessly integrate with more advanced segmentation models such as Segment-Anything for extraction of a boarder range of attributes. Target prompt: "a person wearing a mask".

A.3.5 Analysis on non-facial attributes enhancement

We have conducted supplementary experiments about attributes beyond facial elements. Following the prompt set of PuLID, we evaluated prompts involving common objects, clothing, and other typical non-facial attributes, using Segment Anything for mask generation. The experimental results across several baselines are shown in Table.8:

Table 8: Quantitative comparison of on non-facial attributes.

	PC(%) ↑	IF(%) ↑	Face Div. (%) ↑
Face2Diffusion	22.87	41.08	41.34
Face2Diffusion + FreeCure	23.34 (+2.06%)	40.75 (-0.80%)	41.91 (+1.38%)
InstantID	22.43	65.55	47.73
InstantID + FreeCure	23.51 (+4.81%)	65.34 (-0.32%)	48.06 (+0.69%)
PhotoMaker	24.67	52.14	46.02
PhotoMaker + FreeCure	25.37 (+2.84%)	51.77 (-0.71%)	46.58 (+1.21%)
PuLID (SDXL)	27.43	59.41	41.97
PuLID (SDXL) + FreeCure	28.13 (+2.55%)	59.03 (-0.64%)	42.05 (+0.19%)

Experimental results indicate that the facial personalization model exhibits a weaker copy-paste effect in non-facial regions, leading to less pronounced improvements in prompt consistency compared to the facial attributes specifically addressed in our submission. However, we observed that FreeCure still demonstrates significant restoration effects for attributes such as headphones and hats, which are in close proximity to the face (see Fig.17). Moreover, since the restoration occurs outside the facial region, FreeCure has a more negligible negative impact on identity fidelity.

Figure 17: **Results on non-facial attributes.** FreeCure can still improve prompt consistency on non-facial attributes, especially on objects whose location is close to face.

A.4 More Analysis

A.4.1 Non-interference Enhancement of Multiple Facial Attributes

We validate that the subsequent enhancement processes of FreeCure do not impact attributes that have already been enhanced. Fig.18 visualizes the intermediate results at FreeCure's different stages. When APG is used in the latter stages, attributes previously enhanced through FASA remain unaffected. This illustrates that FreeCure's improvement across multiple attributes is highly robust and consistent.

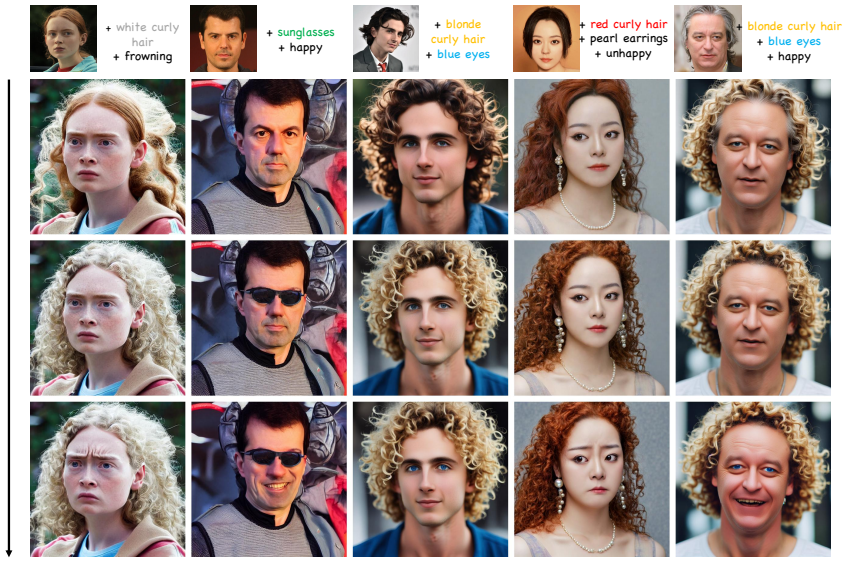


Figure 18: **Demonstration of FreeCure's non-interference manner**. **top row**: personalization models' outputs; **middle row**: intermediate enhanced results with FASA; **bottom row**: final results after APG enhancement.

A.4.2 More Robustness Justification under Inaccurate Attribute Masks

Although FreeCure rely on attribute segmentation process based on pretrained semantic segmentation models. Its performance does not sensitive to the accuracy of these models. We design two-way validation experiments that introduce inaccuracy to attribute masks. First, we apply a dilation operation on masks, which might include area which does not belong to the target attribute area. Second, we apply an erosion operation, which might cause loss of some useful information about attributes. To create significant distortion of masks, we set the receptive field of both dilation and erosion to 25 pixels. We conduct experiments on several baselines, and the Table.9 below reports the results for respective prompt consistency (PC) and identity fidelity (IF).

Table 9: PC and IF Performance for Inaccurate Masks

	original IF	original PC	dialated IF	dialated PC	eroded IF	eroded PC
InstantID	62.01	23.62	62.23	23.45	62.23	23.21
PhotoMaker	50.15	24.91	50.01	24.87	50.44	24.34
PuLID-SDXL	56.95	26.05	56.74	25.51	57.01	25.36
PuLID-FLUX	72.61	24.78	72.39	24.52	72.97	24.32

The results demonstrate that identity fidelity (IF) remains largely consistent across both dilated and eroded mask conditions, with slight improvements observed in eroded masks. This occurs because mask erosion preserves certain attributes similar to the reference image. As for prompt consistency, both dilation and erosion cause minimal degradation, though erosion exhibits marginally greater

impact. We consider that this stems from cases where fine-grained attribute restoration (e.g., iris color, earrings) may fail when the mask completely degenerates due to erosion.

Generally, the performance degradation of FreeCure due to inaccurate masks is limited. Moreover, in practical applications, specialized segmentation models (e.g., BiSeNet, Segment Anything) exhibit high robustness, making failure cases caused by mask inaccuracies even rarer than observed in these deliberately designed experiments. This further validates the inherent robustness of the FreeCure framework.

A.4.3 More Comparison Results w/ and w/o Identical Initial Noise

We conduct extensive experiments to evaluate the impact of initial noise conditions. Fig.19 demonstrates that under identical initial noise conditions, FreeCure consistently enhances target attributes across different input reference images. This finding highlights the robustness and practical applicability of FreeCure. Fig.20 illustrates that FreeCure can successfully inject attribute information from the FD process to the PD process even if their final generated faces have various locations, angles, and sizes. This finding particularly demonstrates the robustness of the FASA module.

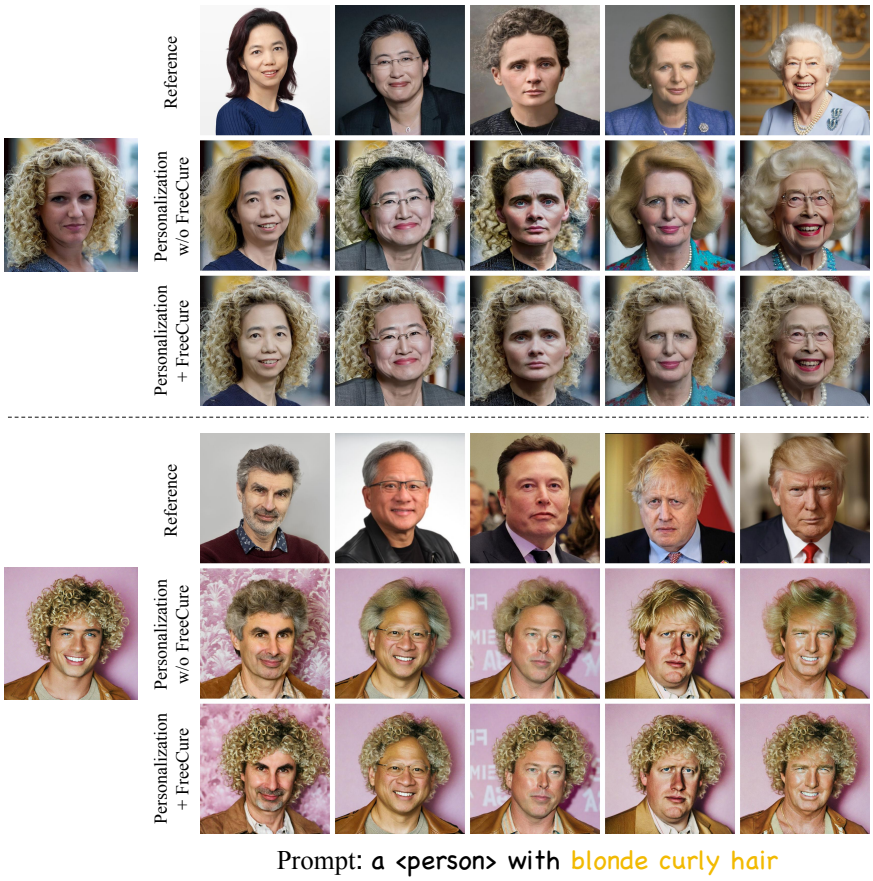


Figure 19: **More studies for attribute enhancement start with identical initial noise.** FreeCure exhibits robust performance with the single foundation output's guidance on personalized outputs from various references (generated based on Face-Diffuser and PhotoMaker).

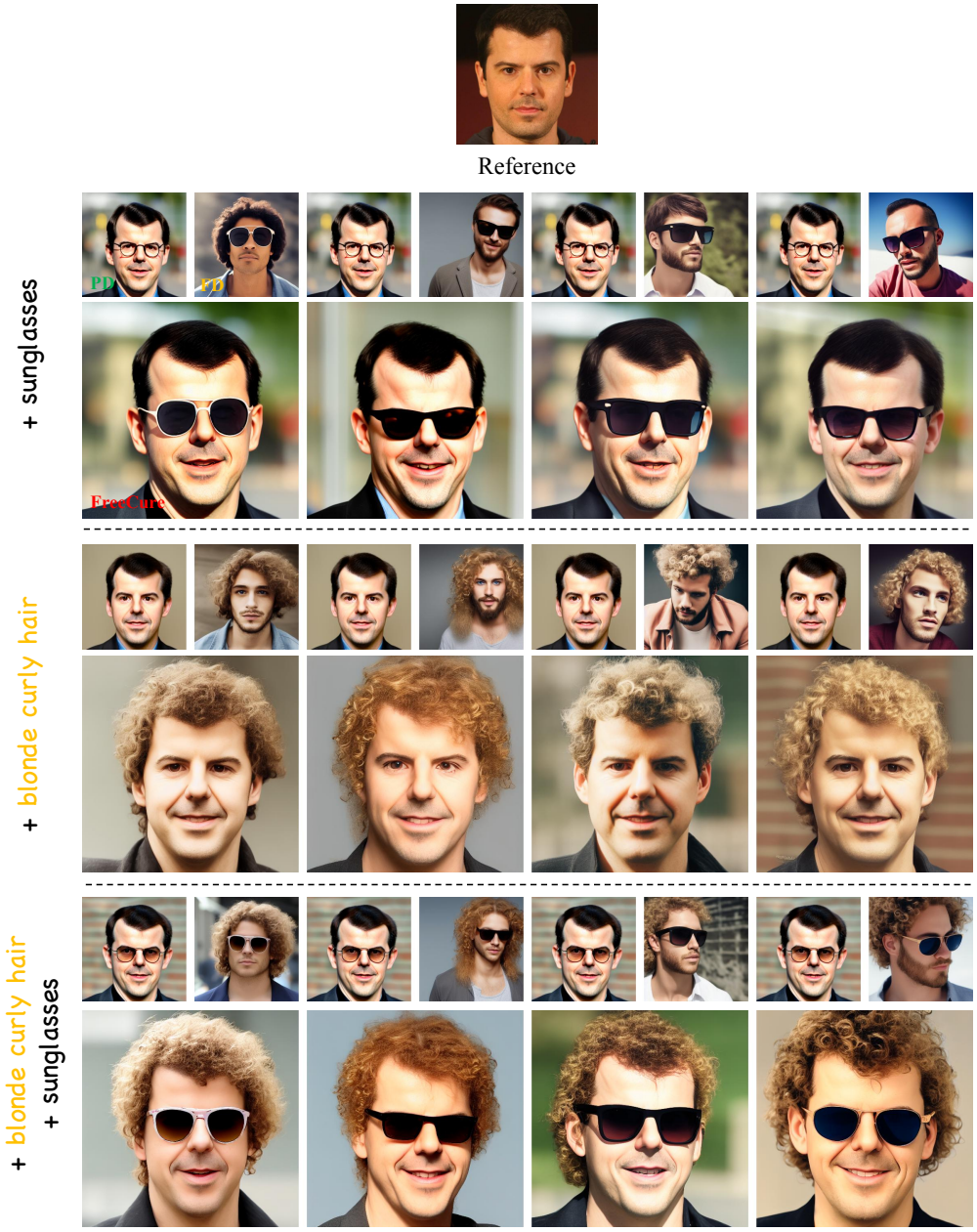


Figure 20: **More validation on FreeCure's robust performance with different initial noises for PD and FD processes**. Even if PD and FD's outputting faces have totally different locations, angles, and sizes, FreeCure can still exhibit stable enhancement performance.

A.4.4 More Detailed Quantitative Results for Ablation Studies

Table.10 provides a supplementary details for ablation studies mentioned in Table.3. The results indicate that both FASA and APG consistently improve prompt consistency across all original personalization models. Notably, FASA yields more substantial improvements compared to APG, aligning with our findings in Sec.5.3.

Table 10: **Quantitative ablation analysis for FASA and APG's independent effect on overall performance.** The table provides detailed ablation studies on each baselines.

Method	FASA	APG	PC(%) ↑	IF(%) ↑	PC × IF (hMean) ↑
FastComposer	✓	✗	20.91	41.94	27.91
	✗	✓	19.03	42.84	26.35
Face-Diffuser	✓	✗	21.86	57.92	31.74
	✗	✓	20.74	58.11	30.60
Face2Diffusion	✓	✗	22.73	39.30	28.80
	✗	✓	21.98	39.47	28.24
InstantID	✓	✗	23.02	62.45	33.64
	✗	✓	22.76	62.51	33.40
PhotoMaker	✓	✗	24.76	50.62	33.25
	✗	✓	23.61	50.86	32.25
PuLID (SDXL)	✓	✗	25.86	57.85	35.74
	✗	✓	25.54	57.24	35.33
PuLID (FLUX)	✓	✗	24.12	72.98	36.26
	✗	✓	23.86	73.03	35.97
InfiniteYou	✓	✗	24.87	78.34	37.75
	✗	✓	24.32	78.67	37.15

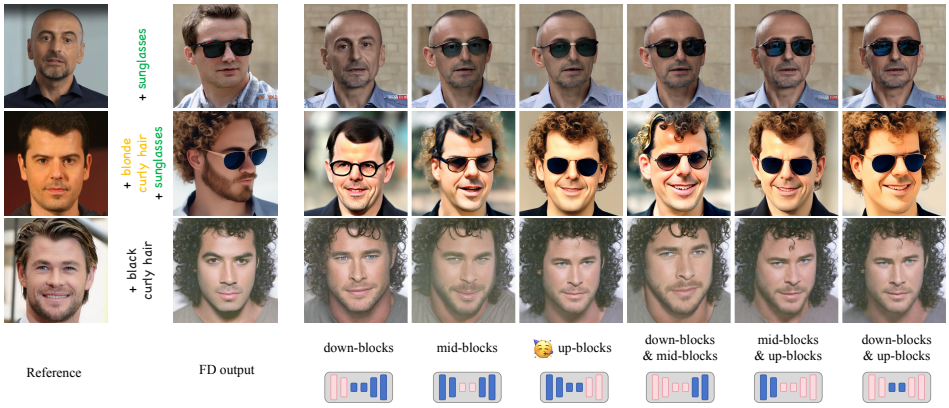


Figure 21: **Ablation study for applying FASA at different positions of UNet**. We find out that applying FASA modules to upsampling blocks alone will be sufficient enough to exhibit promising enhancement for target attributes.

A.4.5 Ablations of Applying FASA at Different Positions in the Denoising Model

We conduct detailed ablation studies to determine the optimal placement of the FASA module within the denoising model, as illustrated in Fig.21. When FASA is applied exclusively to the downsampling and middle blocks, its effectiveness is limited. In contrast, applying FASA to the upsampling blocks yields the most significant performance improvements. Furthermore, adding FASA to the middle and downsampling blocks provides negligible additional benefits when the upsampling blocks are already

utilized, as the upsampling blocks alone are sufficient to accurately transfer attribute information from the FD to the PD process. Based on these findings, we conclude that applying FASA to the denoising model's upsampling blocks represents the optimal configuration.

A.4.6 More Visualization of FASA

We provide a more detailed visualization of FASA in Fig.22, which substantiates the claim of Sec. 5.2 that FASA enhances facial attributes in a fine-grained manner. For instance, FASA effectively captures and faithfully transfers information of attributes localized in small areas (e.g. eyes, pearl earrings). Furthermore, in regions unrelated to the target attributes, FASA maintains a strong alignment with the original PD attention maps, demonstrating that it preserves the core functionality of the personalization models.

Figure 22: **More visualization of FASA map on attribute-relevant and attribute-irrelevant areas.** Red points refer to the area of target attributes and green points refer to the area which is not related to target attributes.

A.4.7 More Studies on Identity Embedding

Fig.23 illustrates the impact of identity embedding interpolation when integrated into the cross-attention layers of different blocks, as mentioned in Sec.3. When cross-attention maps of the FD process are injected into the downsampling blocks or middle blocks, the changes in output are minimal, even if all identity embedding's cross-attention maps are replaced ($\alpha = 1$). It is only when the interpolation is applied to the upsampling blocks that significant degradation of identity information is evident, while other facial attributes are effectively restored. In summary, the identity embedding exerts its most significant influence within the upsampling blocks of the denoising model. However, a small value of α can cause significant identity information loss, supporting the argument in Sec.3, that the well-trained cross-attention layers for identity information extraction is susceptible.

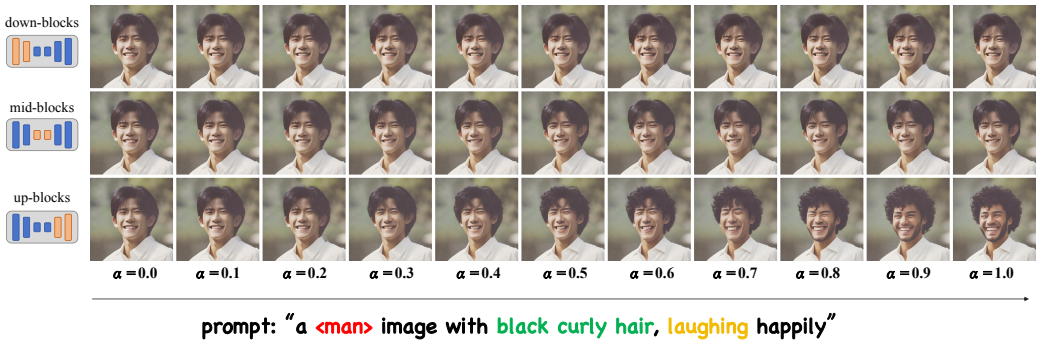


Figure 23: **Identity embedding interpolation's influence on cross-attention layers of different blocks in the denoising model.** It is applied on: downsampling blocks (top), bottleneck blocks (middle), and upsampling blocks (bottom).

We have conducted a detailed analysis using finer-grained α values and their corresponding quantitative outcomes. Below are the results:

Table 11: Fined-grained Quantitative Metrics for Examples of Cross-attention Interpolation in Sec.3

	α	0	0.2	0.4	0.6	0.8	1
top	Identity Fidelity	50.23	24.53	10.34	6.42	3.43	2.1
	Prompt Consistency	24.98	25.5	26.39	28.54	30.43	33.01
bottom	Identity Fidelity	52.34	49.52	29.12	15.75	6.32	3.2
	Prompt Consistency	26.76	30.34	31.09	32.98	34.23	35.02

The results demonstrate that identity fidelity declines sharply even when only a small portion of the original cross-attention map is altered, suggesting that identifying *a well-optimized sweet spot* through cross-attention manipulation alone is highly challenging. In summary, difficulties in cross-attention manipulation provide us with a more solid motivation in the exploration of self-attention modules.

A.4.8 Inference Time Analysis

We conduct experiments on a single H800 GPU with 80 GB VRAM to measure each baseline's inference time before and after applying FreeCure. Table.12 below reports the results ("*" means 30-step denoising due to official guidance and 50-step denoising is applied for the rest baselines)

It is true that introducing extra calculation can lead to longer runtime, we would like to emphasize that most training-free methods introduce extra processes, including inversion operation, denoising processes, and extra attention calculation. Plus, comparing to the fact that most encoder-based personalization methods require a long range time of data collection, preprocessing and tuning (this may consume a large array of GPUs and thousands of GPU-hours), our training-free method provides an innovative perspective for prompt consistency improvement by the knowledge in themselves, without the need for designing new model architecture and dataset curation.

Table 12: Runtime analysis

	Baseline (seconds)	Baseline + FreeCure (seconds)			
		Stage 1	Stage 2 with FASA	APG	Total
FastComposer	1.79	1.96	2.19	1.83	5.98
Face-Diffuser	2.13	2.63	2.33	1.92	6.88
Face2Diffusion*	1.13	1.25	1.47	1.23	3.95
PhotoMaker	6.52	7.04	10.03	6.1	23.17
InstantID*	7.12	7.63	10.65	6.23	24.51
PuLID-SDXL	7.43	8.02	10.96	7.28	26.26
PuLID-FLUX*	12.84	14.89	16.86	12.95	44.7
InfineteYou*	8.34	10.23	13.53	9.03	32.79

A.4.9 User Study Template

Fig.24 show an example of our proposed user study on the performance of FreeCure.

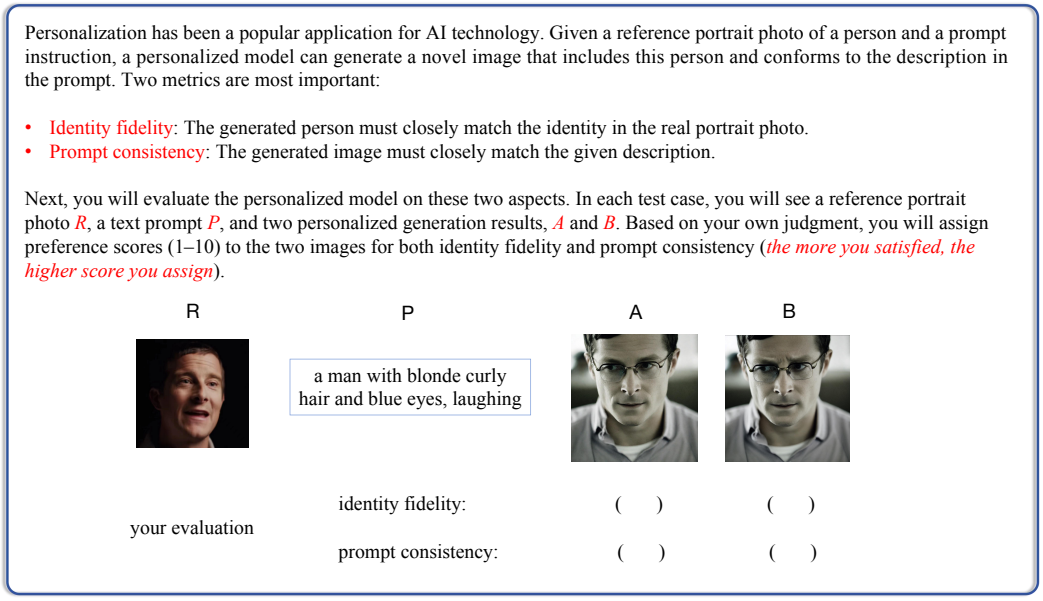


Figure 24: **An example of user study.** We use scoring strategy to collect user assessments about the samples with/without FreeCure.