Supplementary of High-fidelity Person-centric Subject-to-Image Synthesis

Yibin Wang[1, 3]*, **Weizhong Zhang**[2]*, **Jianwei Zheng**[3]†, **Cheng Jin**[1, 4]†

[1]School of Computer Science, Fudan University [2]School of Data Science, Fudan University

[3]College of Computer Science and Technology, Zhejiang University of Technology [4]Haina Lab

yibinwang1121@163.com, weizhongzhang@fudan.edu.cn, zjw@zjut.edu.cn, jc@fudan.edu.cn

A. More cases of problems

More cases of the challenges confronted by current SOTA methods are supplied in Fig. 1 and Fig. 2.

B. Algorithm

The computation pipeline of Saliency-adaptive Noise Fusion is illustrated in Algorithm 1.

Algorithm 1 SNF

Input: TDM ε_{θ_T}, SDM ε_{θ_S}, text prompt $\boldsymbol{c}$ and augmented text prompt $\boldsymbol{c}_{aug}$, the noise $\boldsymbol{x}_{T(1-\alpha)}$.

Output: The noise $\boldsymbol{x}_{T(1-\beta)}$

for each t from $T(1-\alpha)$ to $T(1-\beta)$ **do**

  $\boldsymbol{\varepsilon}_T = \varepsilon_{\theta_T}(\boldsymbol{x}_t|\boldsymbol{c})$

  $\boldsymbol{\varepsilon}_S = \varepsilon_{\theta_S}(\boldsymbol{x}_t|\boldsymbol{c}_{aug})$

  get $\boldsymbol{\Omega}^T$ and $\boldsymbol{\Omega}^S$ via Eq. (3) and Eq. (4)

  $\boldsymbol{M}$ = argmax(Softmax($\boldsymbol{\Omega}^T$), Softmax($\boldsymbol{\Omega}^S$))

  get predicted noises $\hat{\varepsilon_S}$ and $\hat{\varepsilon_T}$ via Eq. (2) and Eq. (1)

  $\hat{\varepsilon}$ = $\boldsymbol{M} \odot \hat{\varepsilon}_S$ + (1 - $\boldsymbol{M}$) $\odot$ $\hat{\varepsilon}_T$

  $\boldsymbol{x}_{t-1} \leftarrow \hat{\varepsilon}$

end for

return $\boldsymbol{x}_{T(1-\beta)}$

C. More implementation details

Baselines. We compare with recent state-of-the-art subject-to-image synthesis methods, which included optimization-based techniques like DreamBooth [3] and Custom-diffusion [1]. These models necessitate subject-specific fine-tuning for each subject. We utilize five images per subject for their fine-tuning in our work. We employed implementations from the diffuser library [4] for these methods. Additionally, we also compare with some tuning-free approaches, such as ELITE [5], Subject-diffusion [2], and Fastcomposer [6]. We utilized pre-trained models from the original authors for ELITE and Fastcomposer. However, since Subject-diffusion does not provide a pre-trained model or dataset to the public, we train it on the FFHQ-face [6] dataset, adhering to the original paper's settings as closely as possible. Subsequently, we selected its best model for our comparative analysis.

Training Configurations. During the training phase, we adopted a strategy following [6], where we freeze the text encoder and only train the U-Net, the MLP module, and the last two transformer blocks of the image encoder. For SDM, we trained only with text condition for 20% of the samples, a measure taken to preserve the model's capacity for text-only generation. Furthermore, we applied loss functions exclusively within the subject region for half of the training samples, a step taken to enhance the quality of generation in the subject area. Meanwhile, for TDM, we opted for training without any conditions in place for 20% of the instances, a choice made to facilitate classifier-free guidance sampling.

*Equal contribution.

†Corresponding author

Table 1. Additional quantitative comparison results. "N.A." indicates that the information is not available.

Methods	*Single-Subject*				*Multi-Subject*			
	FID ↓	IS ↑	CLIP-I ↑	DINO ↑	FID ↓	IS ↑	CLIP-I ↑	DINO ↑
ELITE	51.3	7.83	0.722	0.571	N.A.	N.A.	N.A.	N.A.
Dreambooth	41.6	7.98	0.763	0.648	N.A.	N.A.	N.A.	N.A.
Custom-Diffusion	35.7	8.44	0.785	0.662	N.A.	N.A.	N.A.	N.A.
Subject-Diffusion	31.4	8.92	0.778	0.727	36.7	7.44	0.718	0.583
Fastcomposer	29.8	9.16	0.795	0.719	32.1	8.17	0.721	0.602
Face-Diffuser	**21.2**	**11.42**	**0.832**	**0.753**	**25.9**	**10.33**	**0.754**	**0.633**

D. More qualitative comparison

Additional qualitative comparison results are presented in Fig. 3 and Fig. 4.

E. More quantitative comparison

Additional quantitative comparison results are presented in Tab. 1.

F. Ablation study

The functionality of three sampling stages. We conducte ablation experiments to assess the effectiveness of each stage by removing them individually. The results, as presented in

Table 2. The quantitative results for ablating each stage on both single- and multi-subject generation tasks. IP denotes identity reservation and PC denotes prompt consistency.

Methods	*Single-Subject*		*Multi-Subject*	
	IP ↑	PC ↑	IP ↑	PC ↑
w/o semantic scene construction	0.699	0.268	0.587	0.235
w/o subject-scene fusion	**0.710**	0.244	0.588	0.229
w/o subject enhancement	0.583	0.322	0.471	**0.322**
Face-Diffuser	0.708	**0.325**	**0.593**	0.319

Table 3. The quantitative results for replacing SNF with direct addition of predicted noises from SDM and TDM on both single- and multi-subject generation tasks. IP denotes identity reservation and PC denotes prompt consistency.

Methods	*Single-Subject*		*Multi-Subject*	
	IP ↑	PC ↑	IP ↑	PC ↑
addition	0.523	0.221	0.486	0.207
saliency-adaptive noise fusion	**0.708**	**0.325**	**0.593**	**0.319**

Tab. 2, highlight the significance of each stage. Removing the semantic scene construction stage notably affects prompt consistency, indicating its role in generating an initial layout for subsequent stages, thus ensuring overall semantic consistency in the generated images. The absence of the subject-scene fusion stage leads to a substantial drop in prompt consistency, emphasizing its importance in maintaining coherence between subjects and scenes, ultimately impacting image fidelity. Additionally, removing the subject enhancement stage resulted in a significant decrease in identity preservation performance, underscoring its role in enhancing the fidelity of generated persons.

The functionality of Saliency-adaptive Noise Fusion. To further underscore the effectiveness of our proposed Saliency-adaptive Noise Fusion (SNF), we conduct ablation experiments by replacing SNF with the direct addition of two predicted noises from SDM and TDM. The results, as presented in Table Tab. 3, clearly highlight the pivotal role of SNF in preserving the unique strengths of each model and achieving an effective collaboration between two generators. It is evident that direct addition leads to a significant degradation in both identity preservation and prompt consistency. This outcome is unsurprising, as direct addition disregards the specialized expertise of each model.

G. More cases of hyper-parameter analysis

Additional hyper-parameter analyses are presented in Fig. 5.

H. More visualized salience maps

Additional visualized salience maps are presented in Fig. 6.

I. Limitation

First, the persons generated by Face-diffuser closely match the reference images, which may inadvertently contribute to privacy and security concerns. It may cause the unauthorized use of face portraits, impacting the widespread adoption and ethical considerations. Additionally, our approach encounters challenges when it comes to editing attributes of given persons. Moving forward, we plan to engage in further research aimed at addressing these limitations and expanding the capabilities of our model.

J. Societal impact

The societal impact of subject-driven text-to-image generation technologies, such as Face-diffuser, is noteworthy. These advancements have far-reaching implications, fueling creativity in entertainment, virtual reality, and augmented reality industries. They enable more realistic content creation in video games and films, enhancing the overall user experience. However, as these technologies become more accessible, concerns about privacy, consent, and potential misuse have surfaced. Striking a balance between innovation and ethical considerations is crucial to harnessing the full potential of subject-driven text-to-image generation for the benefit of society.

References

[1] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *CVPR*, pages 1931–1941, 2023. 1

[2] Jian Ma, Junhao Liang, Chen Chen, and Haonan Lu. Subject-diffusion: Open domain personalized text-to-image generation without test-time fine-tuning. *arXiv preprint arXiv:2307.11410*, 2023. 1

[3] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *CVPR*, pages 22500–22510, 2023. 1

[4] Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, and Thomas Wolf. Diffusers: State-of-the-art diffusion models. `https://github.com/huggingface/diffusers`, 2022. 1

[5] Yuxiang Wei, Yabo Zhang, Zhilong Ji, Jinfeng Bai, Lei Zhang, and Wangmeng Zuo. Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. *arXiv preprint arXiv:2302.13848*, 2023. 1

[6] Guangxuan Xiao, Tianwei Yin, William T Freeman, Frédo Durand, and Song Han. Fastcomposer: Tuning-free multi-subject image generation with localized attention. *arXiv preprint arXiv:2305.10431*, 2023. 1

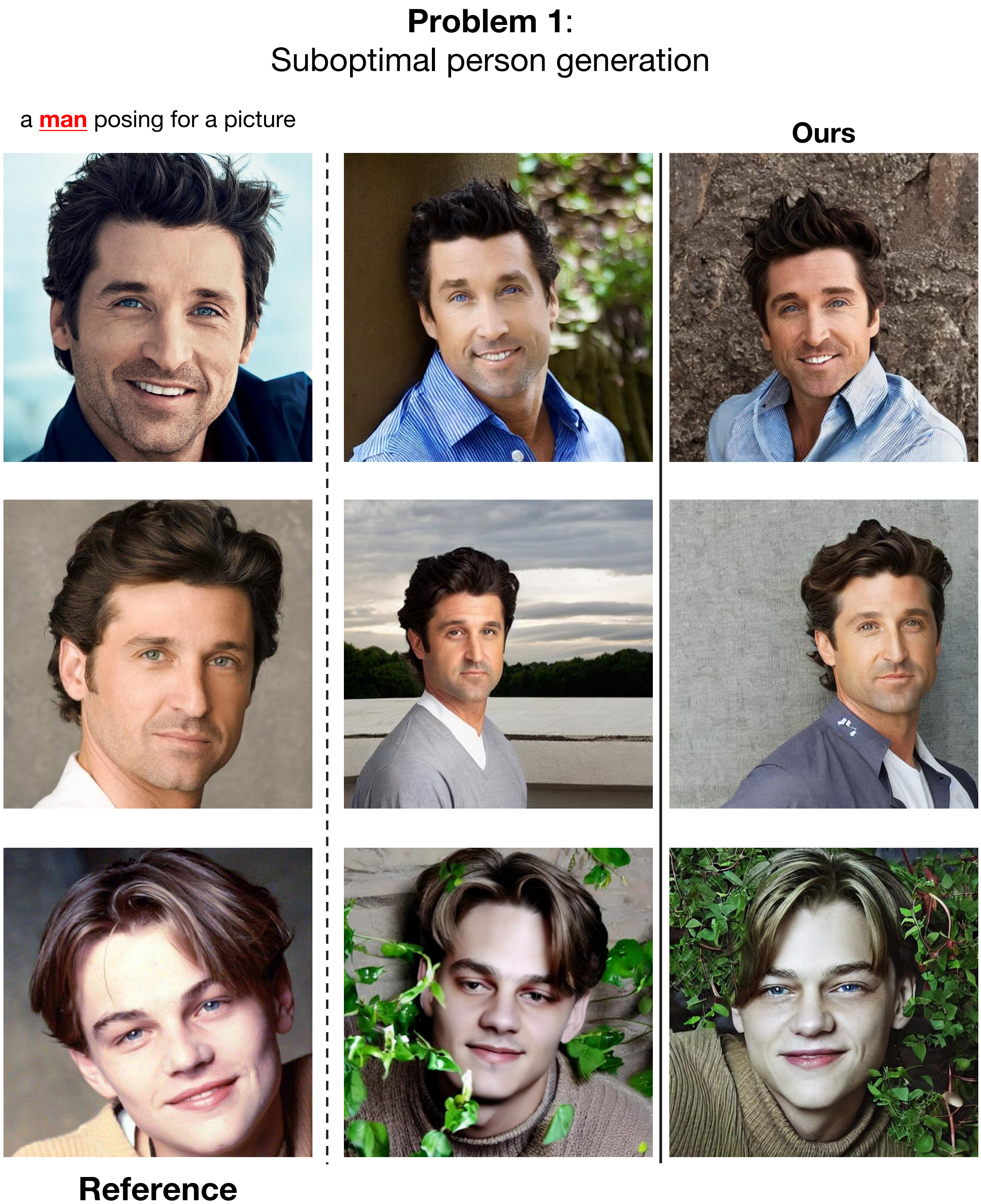

Figure 1. More problem cases of suboptimal person generation.

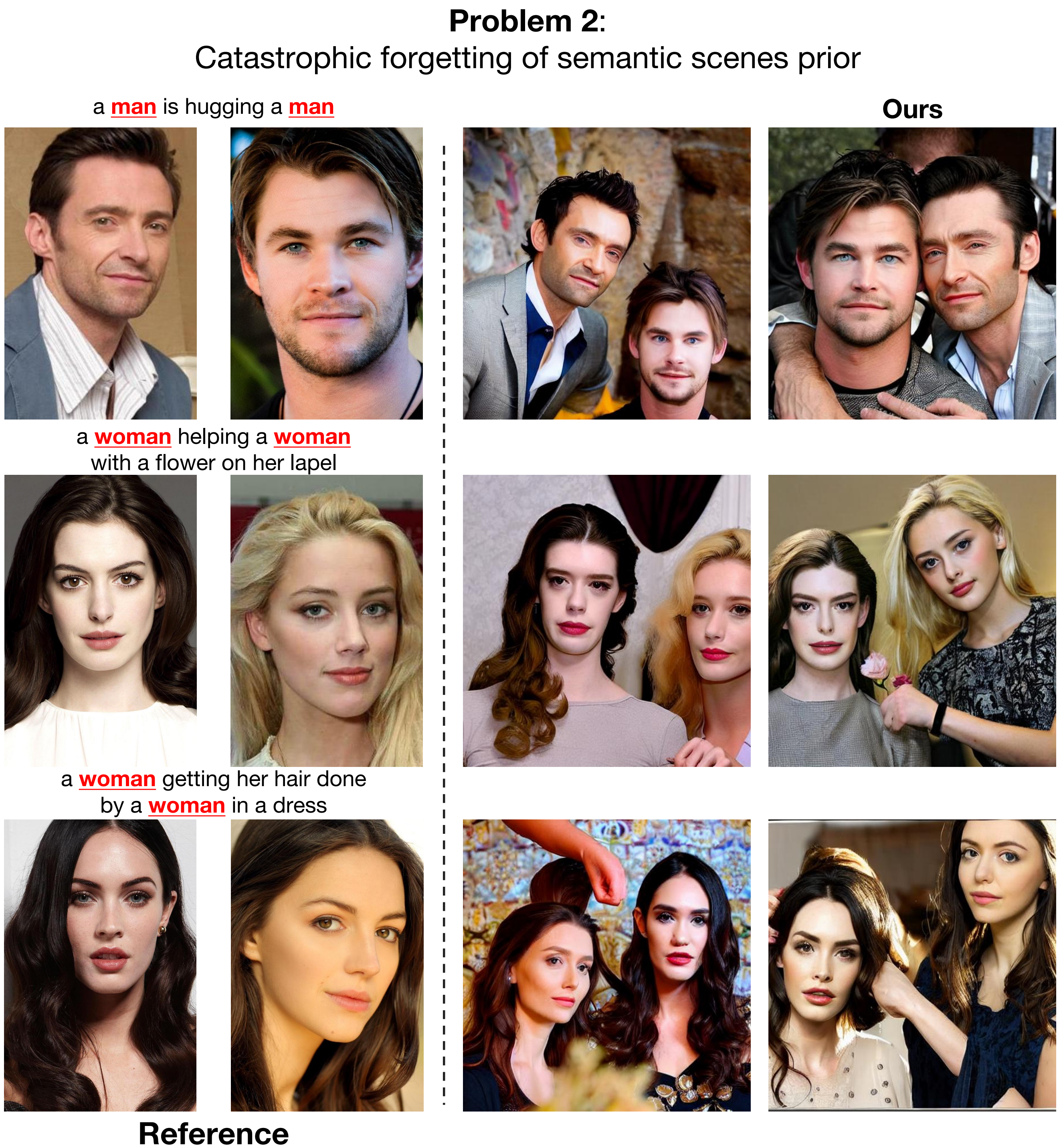

Figure 2. More problem cases of catastrophic forgetting of semantic scenes prior

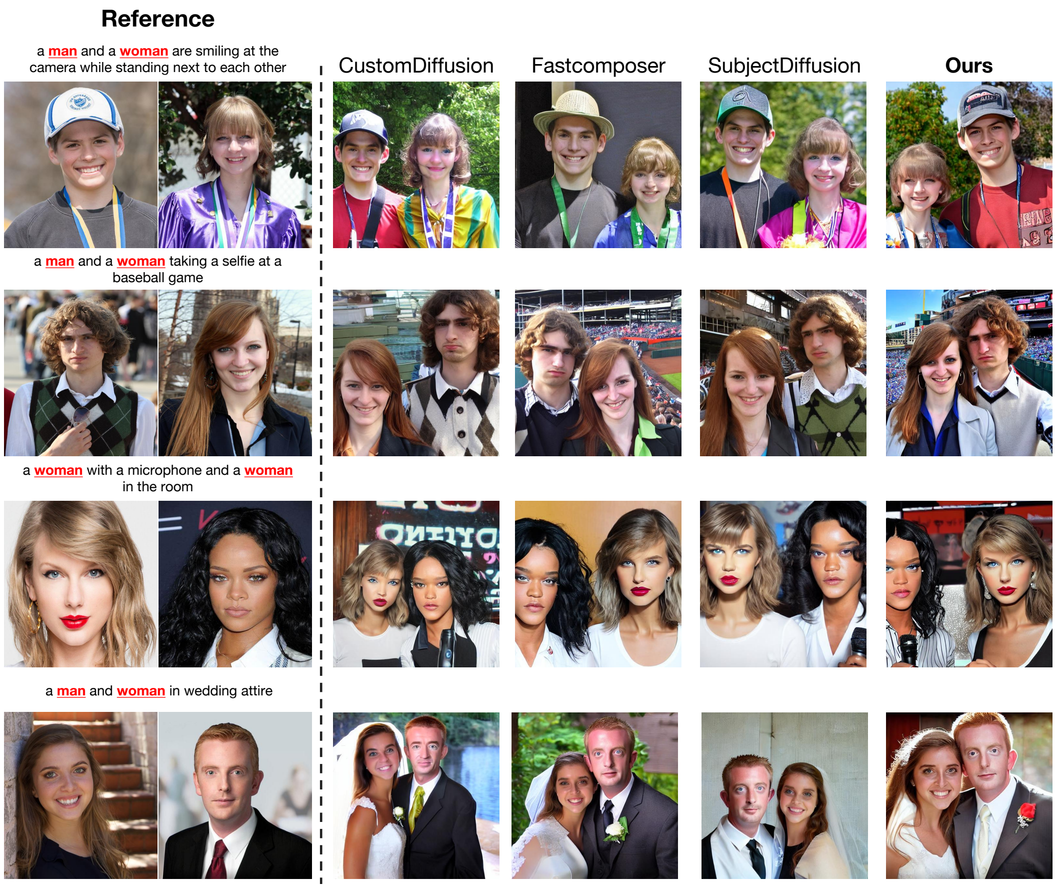

Figure 3. More qualitative comparative results against state-of-the-art methods on multi-subject generation.

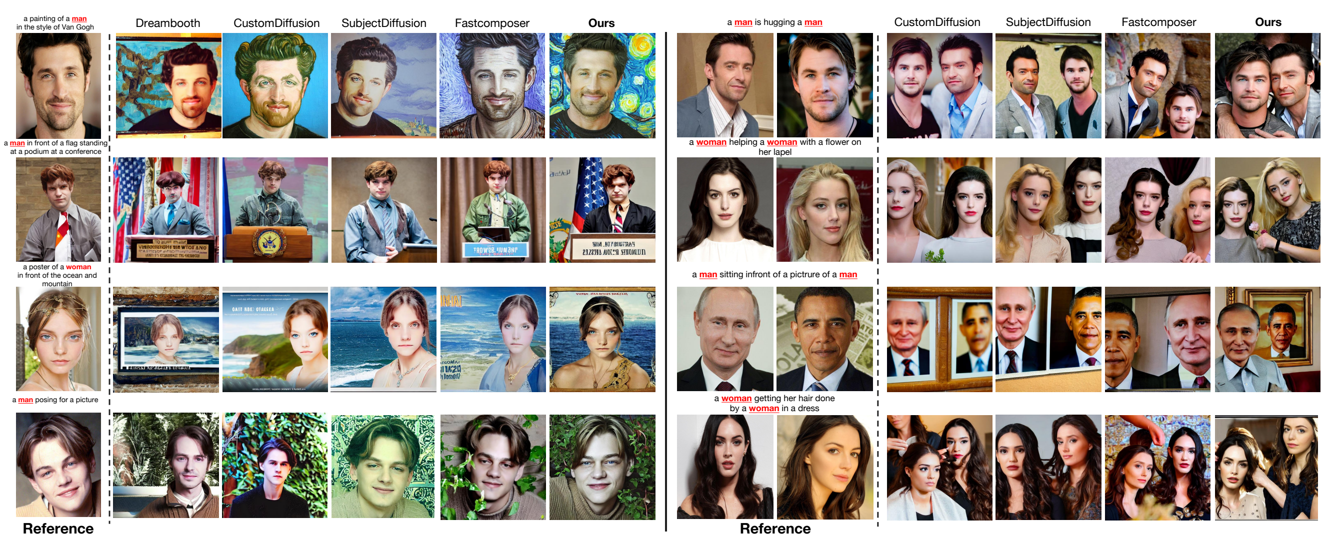

Figure 4. More qualitative comparative results.

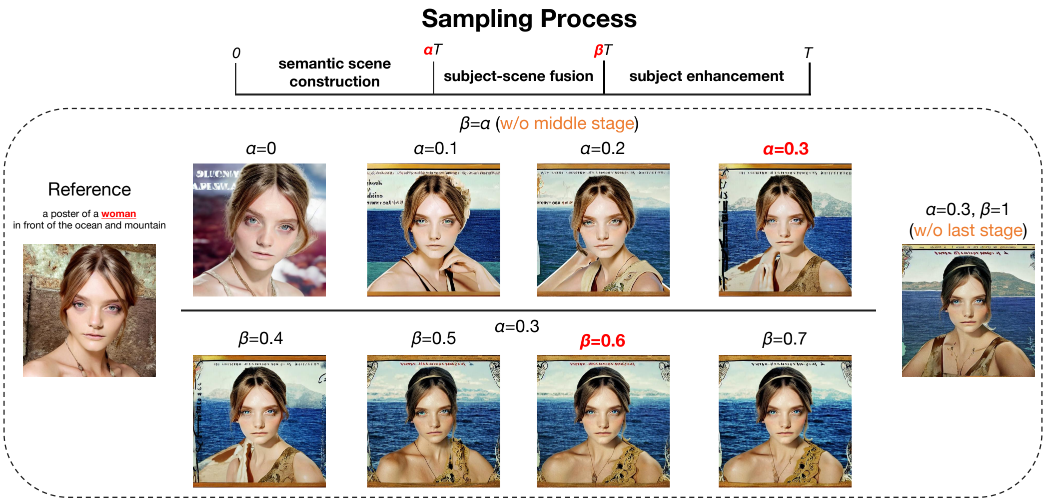

Figure 5. More hyper-parameter visualized analysis of α and β.

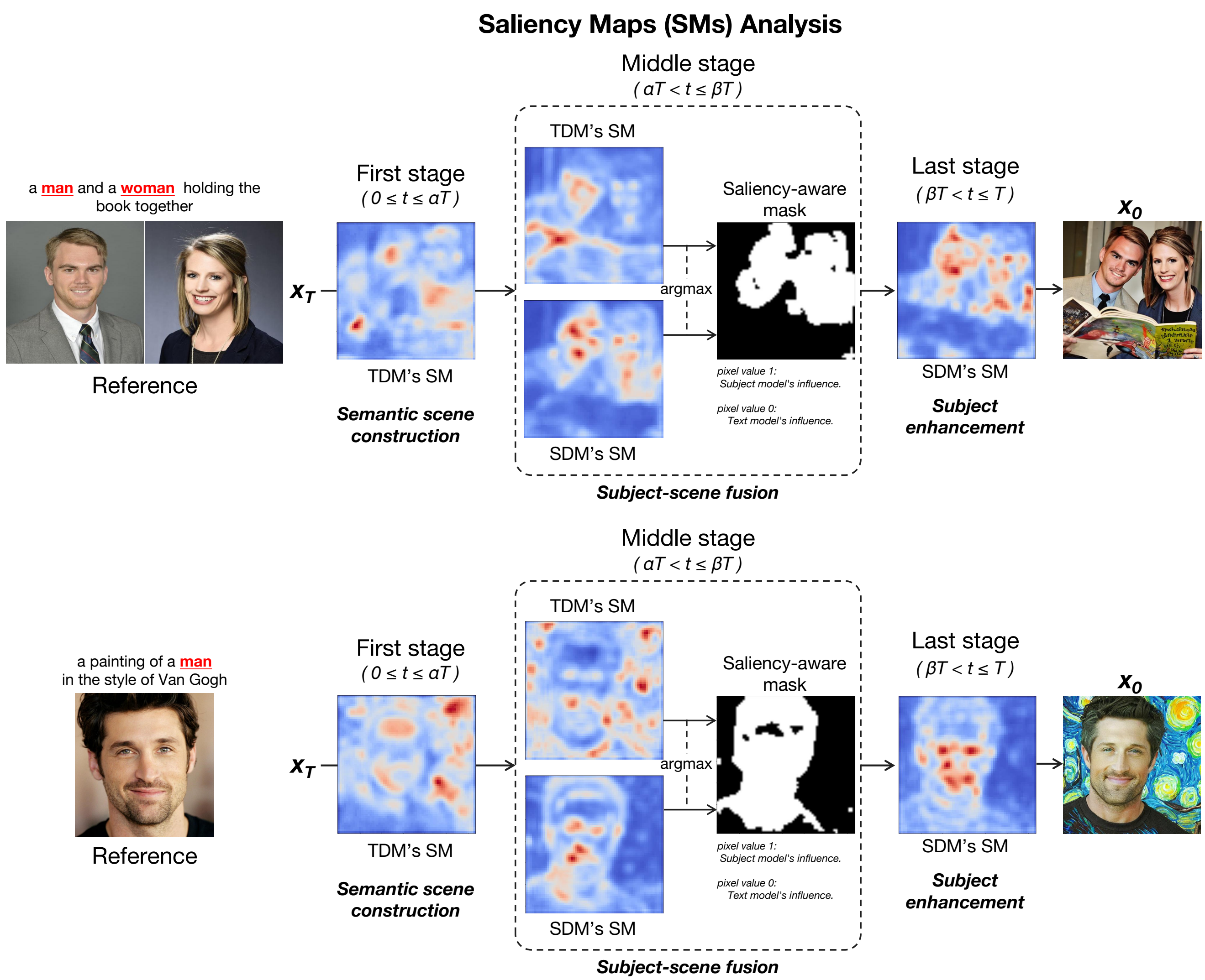

Figure 6. More cases of visualized salience maps of pre-trained models in each stage.