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
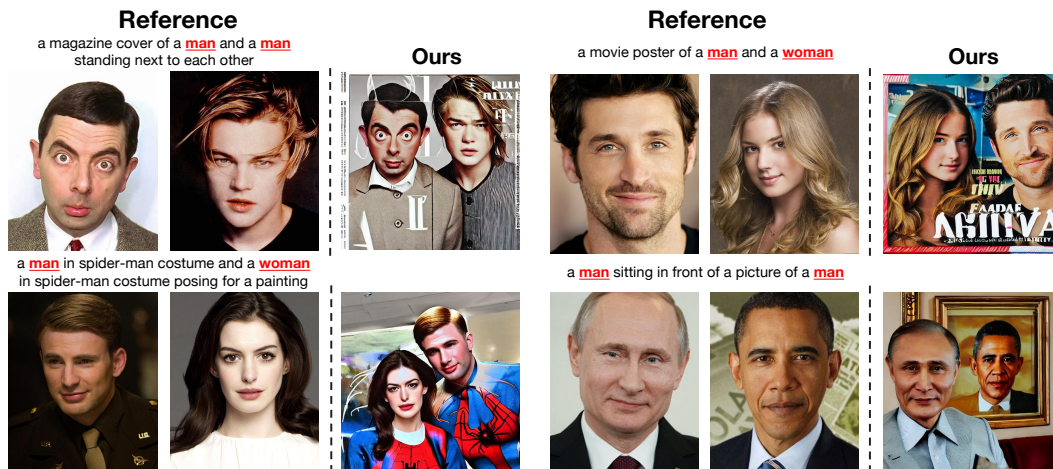
Figure 1. Displayed are the results generated using our Face-diffuser, showcasing its prowess across varied inputs. Each instance comprises two distinct inputs: a textual description, and reference images.

Abstract

Current subject-driven image generation methods encounter significant challenges in person-centric image generation. The reason is that they learn the semantic scene and person generation by fine-tuning a common pre-trained diffusion, which involves an irreconcilable training imbalance. Precisely, to generate realistic persons, they need to sufficiently tune the pre-trained model, which inevitably causes the model to forget the rich semantic scene prior and makes scene generation over-fit to the training data. Moreover, even with sufficient fine-tuning, these methods can still not generate high-fidelity persons since joint learning of the scene and person generation also lead to quality compromise. In this paper, we propose Face-diffuser, an effective collaborative generation pipeline to eliminate the above training imbalance and quality compromise. Specifically, we first develop two specialized pre-trained diffusion models, i.e., Text-driven Diffusion Model (TDM) and Subject-augmented Diffusion Model (SDM), for scene and person generation, respectively. The sampling process is divided into three sequential stages, i.e., semantic scene construction, subject-scene fusion, and subject enhancement. The first and last stages are performed by TDM and SDM respectively. The subject-scene fusion stage, that is the collaboration achieved through a novel and highly effective mechanism, Saliency-adaptive Noise Fusion (SNF). Specifically, it is based on our key observation that there exists a robust link between classifier-free guidance responses and the saliency of generated images. In each time step, SNF leverages the unique strengths of each model and allows for the spatial blending of predicted noises from both models automatically in a saliency-aware manner, all of which can be seamlessly integrated into the DDIM sampling process. Extensive experiments confirm the impressive effectiveness and robustness of the Face-diffuser in generating high-fidelity person images depicting multiple unseen persons with varying contexts. Code is available at https://github.com/CodeGoat24/Face-diffuser.

---

[*]Equal contribution.
[†]Corresponding author.

Problem 1: Suboptimal person generation

a **man** posing for a picture

Ours

Reference

Problem 2: Catastrophic forgetting of semantic scenes prior

a **man** and a **man** holding a piece of paper together

Ours

Reference

Figure 2. Current methods jointly learn the generation of semantic scenes and persons, which leads to a compromise in the quality of person generation (left), and the irreconcilable training imbalance issue leads to catastrophic forgetting of semantic scenes prior (right).

1. Introduction

Subject-driven text-to-image generation [3, 4, 7, 13, 21, 27] can synthesize personalized images depicting particular subjects defined by users with a few sample images. The basic idea in optimization-based approaches [1, 10, 12, 16, 20, 25, 28] is to fine-tune a pre-trained model (e.g., stable diffusion [19]) on a group of provided reference images, typically 3-5 images, for each subject. Another technique roadmap [5, 6, 17, 32, 33] is to retrain a text-to-image generation base model adapted from a pre-trained model with specially-designed structures or retrain the pre-trained model with specific training strategies on a large-scale personalized image dataset. These methods are more practical in real applications since no subject-specific fine-tunings are required in the test time.

However, we notice that there has been a notable dearth of research on person-centric image generation, and existing subject-driven text-to-image generation models [10, 12, 16, 17, 20, 30, 33] are inadequate for this task. To be precise, current state-of-the-art subject-driven text-to-image generation models, e.g., Fastcomposer [33] and Subject-diffusion [17], jointly learn the semantic scene and person generation capability based on some large-scale pre-trained language image model, which has been trained on extensive multimodal datasets like LAION 5B [24]. But they seem to struggle to effectively harness the inherent prior knowledge of semantic scenes encapsulated within these models. We argue that the reason is these models would over-fit to the textual descriptions and forget the rich semantic scene prior after prolonged training, see Fig. 2 (right) as an example. To validate this hypothesis, we perform experiments to assess their performance in terms of identity preservation and prompt consistency. The results, as illustrated in Fig. 3 (left), show that in the late stage of training, the prompt consistency scores for both models keep increasing for the training data, while exhibiting a noticeable decrease on the test data, providing strong empirical support for our hypothesis. Obviously, the most straightforward solution appears to be reducing the training duration. Unfortunately, it is proved to be invalid by the irreconcilable training imbalance issue demonstrated in Fig. 3. Precisely, Fig. 3 (right) indicates that to generate realistic persons, one needs to sufficiently

tune the pre-trained model, while from Fig. 3 (left), we observe that this inevitably causes the model to forget the rich semantic scene prior and makes scene generation over-fit to the training data. Moreover, even with sufficient fine-tuning, these methods are still unable to generate high-fidelity persons. An example is presented in Fig. 2 (left). The reason could be that jointly learning the generation of semantic scenes and persons may lead to a compromise in the quality of person generation.

To sum up, current methods suffer from catastrophic forgetting of semantic scenes prior due to the irreconcilable training imbalance and the suboptimal person generation due to the quality compromise for joint learning.

In this paper, we propose Face-diffuser, an effective collaborative generation pipeline for person and semantic scene synthesis. To break the training imbalance and quality compromise, we first independently fine-tune two specialized pre-trained diffusion models named Text-driven Diffusion Model (TDM) and Subject-augmented Diffusion Model (SDM) based on stable diffusion [19] for scene and person generation, respectively. Recent studies [26, 36] demonstrate that the generation of an image progresses from the overall scene to intricate details. Following this pipeline, Face-diffuser divides the sampling process into three consecutive stages: semantic scene construction by TDM, subject-scene fusion by collaboration between TDM and SDM, and subject enhancement by SDM. Nevertheless, since SDM and TDM are independent, developing an effective collaboration mechanism for them in the subject-scene fusion stage is the crux of achieving high-fidelity person and diverse semantic scene generation, which is the main contribution of this paper. The details are presented as follows.

It can be expected that to generate high-quality samples, the person and semantic scene generators should collaborate seamlessly in an evolving scheme in test time, that is they should be responsible for different areas in images at different time steps and in generating different images. To achieve this, we propose an effective fine-grained collaborative mechanism named Saliency-adaptive Noise Fusion (SNF) based on classifier-free guidance (CFG), which can be seamlessly integrated into the DDIM sampling process. SNF is motivated by our key observation that the CFG response, i.e., the noise distinction between a given condition and a null con-
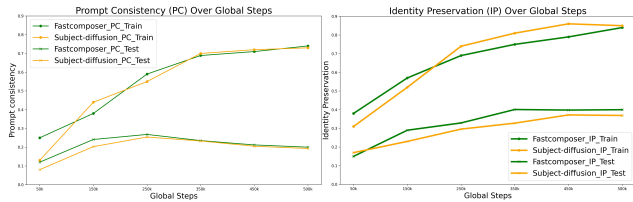
Figure 3. The experimental results showcasing the irreconcilable training imbalance between semantic scene and person generation of Fastcomposer [33] and Subject-diffusion [17]. We partitioned the FFHQ-wild [33] dataset into training and test sets following a 6:1 ratio and assessed their performance in terms of identity preservation and prompt consistency during continuous training.

dition of each generator can effectively evaluate the impact of the condition on each pixel. Similar finding has also been substantiated in the image editing study [35]. Therefore, in each step, we generate a saliency-adaptive mask derived from two models' responses of CFG to automatically allocate areas for them to synthesize. Precisely, for each pixel in the image, the responsibility for its synthesis is allocated to the model with the greater response to it in the current step.

Finally, we would like to highlight that in the test time, to capture and preserve the intricate details of persons given in reference images, we let the CFG response within SDM be the noise distinction between with and without reference person images, thereby neglecting the influence of text condition. This deliberate setting undoubtedly guides SDM to focus the saliency only on person-related areas, resulting in more high-fidelity person generation. Our main contributions can be summarized as follows.

- Our proposed Face-diffuser breaks the training imbalance and quality compromise problems in existing subject-to-image generation methods.
- We develop two independent models for scene and person generation, and a highly effective collaboration mechanism, Saliency-adaptive Noise Fusion, to utilize each model's strengths for higher-quality image synthesis.
- Extensive experiments validate the remarkable effectiveness and robustness of Face-diffuser in generating high-fidelity images portraying multiple unseen persons engaged in diverse contexts.

2. Related Work

2.1. Subject-driven Image Generation

Subject-driven image generation models aim to synthesize personalized images depicting particular subjects defined by users with a few sample images. DreamBooth [20], textual-inversion [10], and custom-diffusion [16] employ optimization-based techniques to incorporate subjects into diffusion models. This is achieved through either fine-tuning the model weights [16, 20] or by transforming subject im-

ages into a text token encoding the subject's identity [10]. However, these models suffer from inefficiency due to the extensive number of fine-tuning steps they demand. In addressing this concern, the tuning-encoder [18] approach mitigates the need for a large number of fine-tuning steps. It achieves this by initially generating an inverted set of latent codes using a pre-trained encoder and subsequently refining these codes through several fine-tuning iterations to better preserve subject identities. However, it's worth noting that all these tuning-based methods [10, 11, 16, 20] necessitate resource-intensive backpropagation. This requirement can pose challenges as it demands hardware capable of fine-tuning the model. Such a demand is neither feasible on edge devices, such as smartphones, nor scalable for cloud-based applications. To this end, several concurrent studies have explored tuning-free methods. For instance, X&Fuse [15] achieves this by concatenating the reference image with noisy latent variables for image conditioning. ELITE [32] and InstantBooth [25], on the other hand, employ global and local mapping networks to project reference images into word embeddings and inject reference image patch features into cross-attention layers to enhance local details. Despite their impressive results in the context of single-object customization, it's important to note that their architectural designs limit their applicability to scenarios involving multiple subjects. This limitation arises due to their reliance on global interactions between the generated image and the reference input image. In comparison, our Face-Diffuser amortizes the computationally expensive subject tuning during the training phase. This design enables instantaneous personalization for multiple subjects using straightforward feedforward methods during the testing process.

2.2. Multi-Subject-driven Image Generation

Custom-diffusion [16] offers the capability of multi-concept composition through joint fine-tuning of the diffusion model for multiple concepts. However, it primarily handles concepts that have clear semantic distinctions, such as animals and their related accessories or backgrounds. On the other hand, SpaText [2] and Collage Diffusion [22] enable multi-object composition by introducing a layout into the image generation process. These layouts are determined by user-provided segmentation masks and are then transformed into high-resolution images using a diffusion model. Nevertheless, these techniques either compose generic objects with customization [16] or rely on the resource-intensive textual-inversion process to encode instance-specific details [22]. Besides, they require users to provide segmentation maps.

Fastcomposer offers a solution for generating personalized, multi-subject images in an inference-only manner. It automatically derives plausible layouts from text prompts, eliminating the need for user-provided segmentation maps. Building upon these advancements, Subject-diffusion [17]
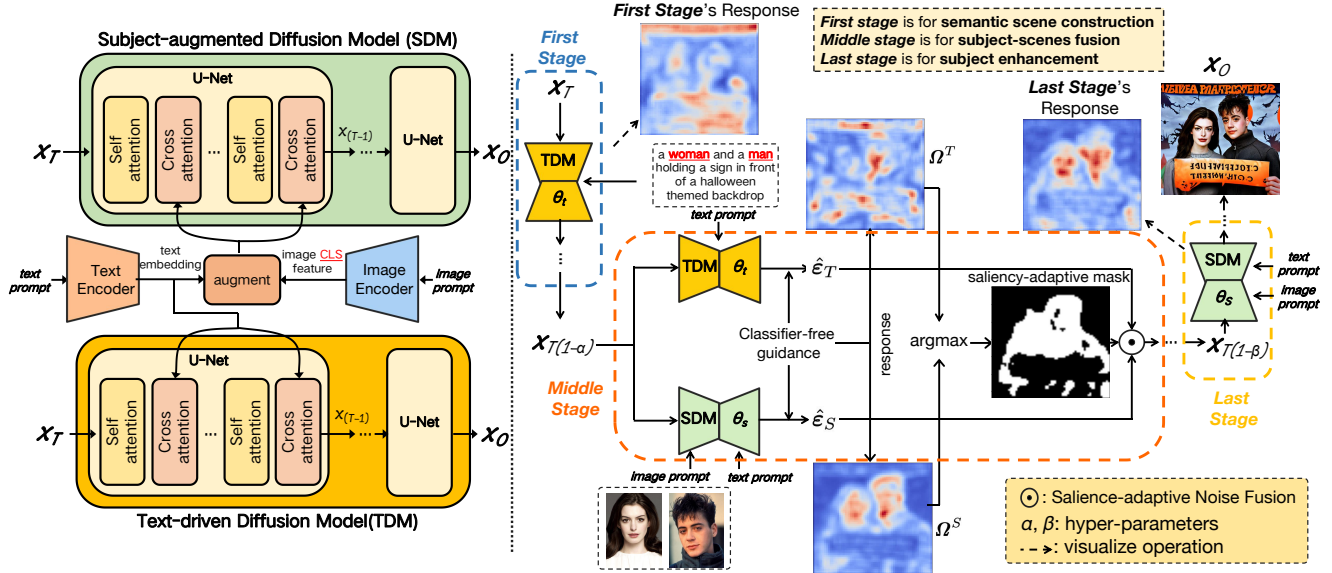
Figure 4. An overview of the Face-diffuser framework. On the left, we display the architectures of two pre-trained models, derived from Stable Diffusion [19], while omitting the autoencoder for simplicity. On the right, we outline our sampling process, which consists of three well-designed stages.

introduces a method that integrates text and image semantics, incorporating coarse location and fine-grained reference image control to enhance subject fidelity and generalization. However, they face notable challenges in person-centric image generation. Specifically, their models jointly learn the semantic scene and person generation, which leads to training imbalance and quality compromises. In contrast, our Face-diffuser excels at generating high-fidelity characters in diverse semantic scenes without the need for fine-tuning, all in an inference-only manner.

3. Face-diffuser

In this section, we will first elaborate on the overview of our architecture as illustrated in Fig. 4. After that, we will introduce two specialized models in our pipeline, i.e., Text-driven Diffusion Model (TDM) and Subject-augmented Diffusion Model (SDM). Last, we will delve into the three-stage sampling process, with a focus on the details of our proposed collaboration synthesis mechanism, Saliency-adaptive Noise Fusion (SNF).

3.1. Overview

In the training stage, as shown in Fig. 4 (left), to eliminate the training imbalance, we *independently* fine-tune the pre-trained Stable-Diffusion [19] as our TDM and another similar model with an additional image prompt as our SDM. These two base models will be responsible for the semantic scene and person generation, respectively.

In the test time (Fig. 4 (right)), unlike previous studies [16,

20], Face-diffuser eliminates the need for subject-specific fine-tuning. To be precise, our sampling process comprised three sequential stages with total T denoising steps. In the first stage, we employ TDM to construct the preliminary semantic scene for αT steps.

Then, TDM and SDM collaborate to infuse the person into the scene in the following βT steps based on our effective collaboration mechanism SNF. At each step, SNF utilizes the responses from classifier-free guidance of both models to produce a saliency-adaptive mask. This mask, with the same size as the target image, serves as a saliency indicator, containing values of 0 and 1 that denote the regions generated by TDM and SDM, respectively. They are responsible for different areas at different time steps for flexible evolving collaboration.

In the last stage, SDM is further utilized to refine the quality of generated persons.

3.2. Semantic Scene and Person Generators

3.2.1 Text-driven Diffusion Model (TDM)

Stable diffusion (SD) is employed as our TDM. For semantic scene generation, given the semantic scene prompt c and the input image x, the VAE first encodes the x into a latent representation z, perturbed by Gaussian noise ε to get z_t at t step during diffusion. Then the text encoder ψ maps semantic scene prompts c to conditional embeddings $\psi(c)$ which would be integrated into the denoiser ε_θ, U-Net through cross-attention [8, 9, 29, 31]. The training objective is to

minimize the loss function as follows:

$$L_{noise} = \mathbb{E}_{z,c,\varepsilon \sim N(0,1),t} \parallel \varepsilon - \varepsilon_\theta(z_t, t, \psi(c)]) \parallel_2^2$$

During inference, a random noise z_T is sampled from a normal distribution $N(0, 1)$, and this noise is iteratively denoised by the U-Net to produce the initial latent representation z_0. Subsequently, the VAE decoder maps these latent codes back to pixel space to generate the final image.

3.2.2 Subject-augmented Diffusion Model (SDM)

The SDM model tailored for subject generation is also based on the SD model but includes an additional reference image condition r. Inspired by previous works like [17, 33], we adopt a tuning-free approach by enhancing text prompts with visual features extracted from reference images. When given a text prompt and a list of reference images, we begin by encoding the text prompt and reference subjects into embeddings using pre-trained CLIP text and image encoders, respectively. Following this, we replace the user-specific word embeddings with these visual features and input the resulting augmented embeddings into a multilayer perceptron (MLP). This process yields the final conditioning embeddings, denoted as c_{aug}. The loss function of SDM closely resembles the one in Eq. (1), with the substitution of $\psi(c)$ by $\psi(c)_{aug}$.

$$L_{noise} = \mathbb{E}_{z,c,\varepsilon \sim N(0,1),t} \parallel \varepsilon - \varepsilon_\theta(z_t, t, \psi(c)_{aug}]) \parallel_2^2$$

3.2.3 Condition Effectiveness Enhancement

For TDM, to reinforce the effectiveness of the conditions in scene generation, Classifier-free Guidance (CFG) is employed in each step to extrapolate the predicted noise along the direction specified by certain conditions. Specifically, at step t, CFG takes the form of

$$\hat{\varepsilon}_T = \varepsilon_\theta(z_t|\varnothing) + s(\underbrace{\varepsilon_\theta(z_t|c) - \varepsilon_\theta(z_t|\varnothing)}_{R_T}), \quad (1)$$

where $\varnothing$ signifies a null condition. The hyperparameter $s > 0$ denotes the guidance weight, and the reinforcing effect becomes stronger when s increases.

For SDM, we notice that it is conditioned on both the text and reference images. To eliminate the impact of text condition c and emphasize the noise distinction ablating the reference person images only, we extend CFG as follows:

$$\hat{\varepsilon}_S = \varepsilon_\theta(x_t|\varnothing) + s(\underbrace{\varepsilon_\theta(x_t|c, r) - \varepsilon_\theta(x_t|c)}_{R_S}). \quad (2)$$

This careful design undoubtedly strengthens its ability to generate higher-fidelity characters, as this leads the model to

become more adept at capturing and preserving the subtle details of the reference images.

We call the predicted noise differences R_T and R_S in Eqn.(1) and (2) the responses of the semantic scene and reference image conditions. They play a fundamental role in developing our collaborative mechanism in Face-Diffuser, which is represented in the next subsection.

3.3. Collaboratively Synthesis

Face-Diffuser employs the following 3 stages to generate each image:

Stage I: Given initial noise x_T, we first employ TDM to construct the scene for αT steps and output $x_{T(1-\alpha)}$.

Stage II: After construction of the preliminary scene, we take previous stage's output as input and leverage TDM and SDM to collaboratively infuse the person into the scene for $(\beta - \alpha)T$ steps, outputting $x_{T(1-\beta)}$.

Stage III: Taking $x_{T(1-\beta)}$ as input, SDM is further utilized to enhance the details generation of persons in the remaining steps.

In the following, we will delve into the collaboration details between TDM and SDM in the middle stage.

3.3.1 Saliency-adptive Noise Fusion

Note that the responses R_T and R_S actually evaluate the impact of the semantic scene and reference images on each pixel of the predicted noises, the regions with large values mean the conditions have significant impacts on these pixels, which naturally defines the responsibility of TDM and STM in this step.

Formally, we first define the following two salience maps based on R_T and R_S:

$$\Omega^T = \text{Smooth}(\text{Abs}(R_T)), \quad (3)$$

$$\Omega^S = \text{Smooth}(\text{Abs}(R_S)), \quad (4)$$

where the operator Abs($\cdot$) calculates the absolute values of the input variables, while the Smooth($\cdot$) function is applied to reduce high-frequency noise, effectively eliminating local outliers and enhancing the coherence of adjacent regions. The empirical validations of Ω^T and Ω^S, i.e., their visualizations are presented in Sec. 4.3.

Given Ω^T and Ω^S, we proceed to develop the saliency-adaptive fusion mask through a comparison between these two salience maps:

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

The softmax operation here is crucial, as the values of Ω^T and Ω^S could have different magnitudes, it ensures that the sum of each salience map remains constant thus let them comparable. The mask M is adopted to define the collaboration mechanism, i.e, the pixels with $M_{ij} = 0$ and $M_{ij} = 1$
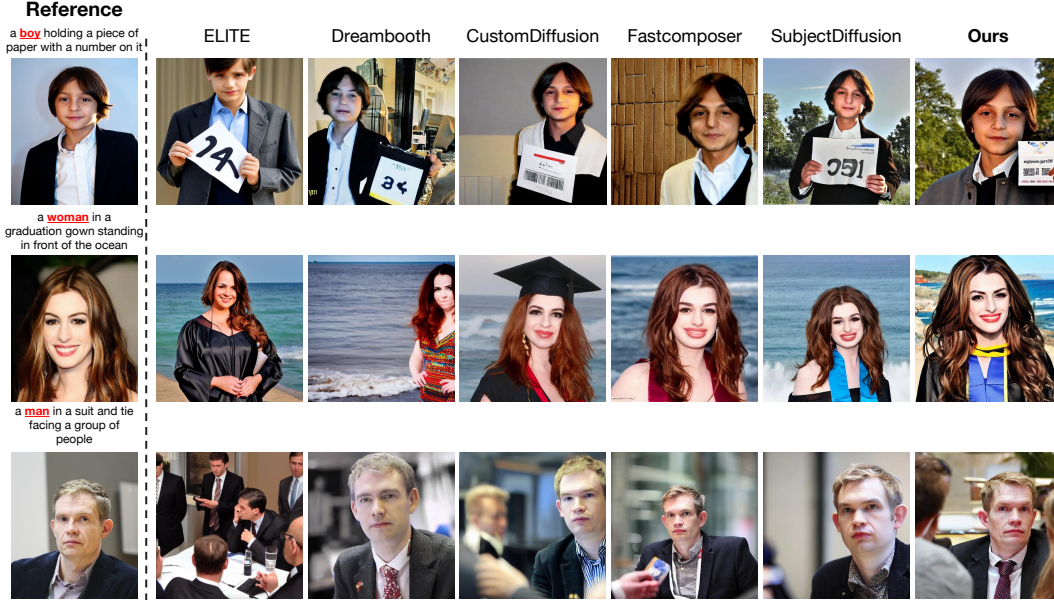
Figure 5. Qualitative comparative results against state-of-the-art methods on single-subject generation.

Figure 6. Qualitative comparative results against state-of-the-art methods on multi-subject generation.

are allocated to TDM and SDM in generation, respectively. Finally, the fused noise can be obtained through the following process:

$$\hat{\boldsymbol{\varepsilon}} = \boldsymbol{M} \odot \hat{\boldsymbol{\varepsilon}}_S + (1 - \boldsymbol{M}) \odot \hat{\boldsymbol{\varepsilon}}_T, \qquad (5)$$

where $\odot$ denotes Hadamard Product, and we omit t for simplicity. It is essential to note that in each sampling step, both models take the blended z_t as input, facilitating the automatic semantic alignment of the two models' noise space.

4. Experiment

4.1. Implementation details

Datasets. Two datasets are used in our experiments. One is FFHQ-face [33], which is constructed based on the FFHQ-

wild dataset [14] and comprised of a total of 70,000 samples, 60,000 for training, and 10,000 for testing. Each sample contains a caption and one or more persons. The other dataset is the Single-benchmark dataset employed in the recent work [17, 33]. It comprises 15 subjects, each with 30 text prompts.

Training Configurations. We train our models SDM and TDM based on the pre-trained Stable-Diffusion (SD) v1-5 model [19] with the FFHQ-face dataset. For SDM's image encoder, we utilized OpenAI's *clip-vit-large-patch14* vision model, which acts as the companion model to the text encoder in SDv1-5. We conducted training on the SDM for a total of 450k steps and on the TDM for 250K steps, utilizing 4 NVIDIA A100 GPUs. We set a constant learning rate of 1e-5 and a batch size of 8.

Evaluation. Quantitively, we evaluate the single-subject generation on Single-benchmark and the multi-subject generation on the test set of FFHQ-face. To further assess the robustness and superiority of our method, we use a diverse set of unseen persons and semantic scenes for qualitative comparison. We assess image generation quality using two metrics: identity preservation (IP) and prompt consistency (PC). IP score is obtained by face detection in both reference and generated images using MTCNN [34] and then pair-wise identity similarity is calculated using FaceNet [23]. For multi-subject evaluation, we identified and matched faces in generated images with reference subjects, measuring overall identity preservation based on the minimum similarity value. PC is assessed through the average CLIP-L/14 image-text similarity, following the approach in textual-inversion [10].
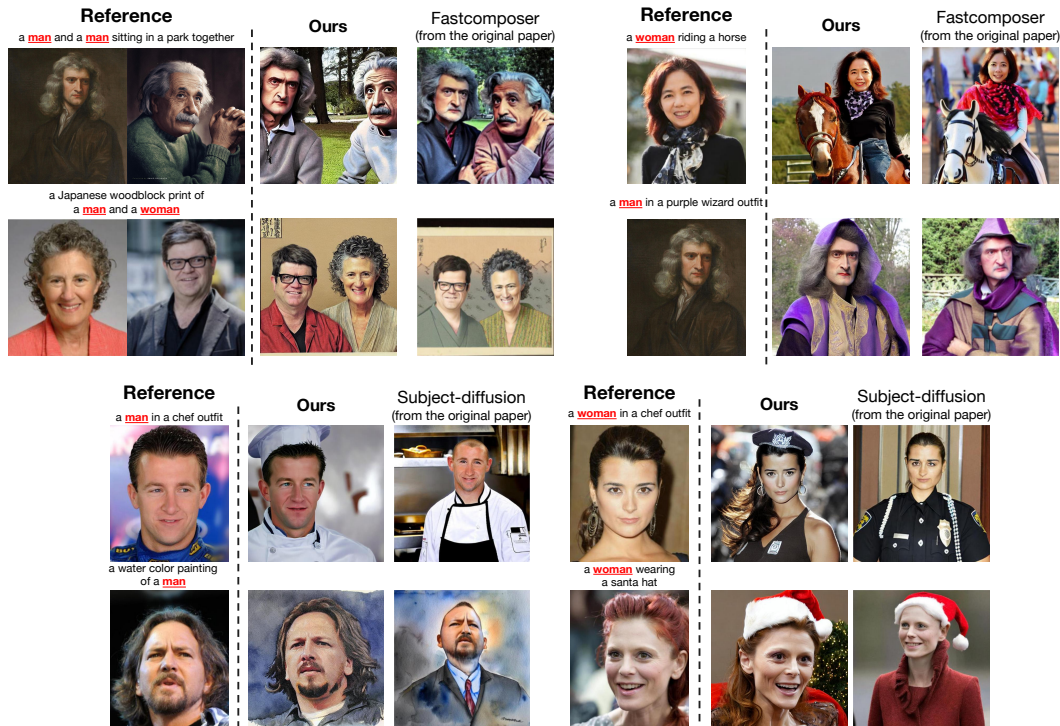
Figure 7. Qualitative comparative results against Fastcomposer and Subject-diffusion using the samples provided in their original papers. The generated results of Fastcomposer and Subject-diffusion are all from their original papers.
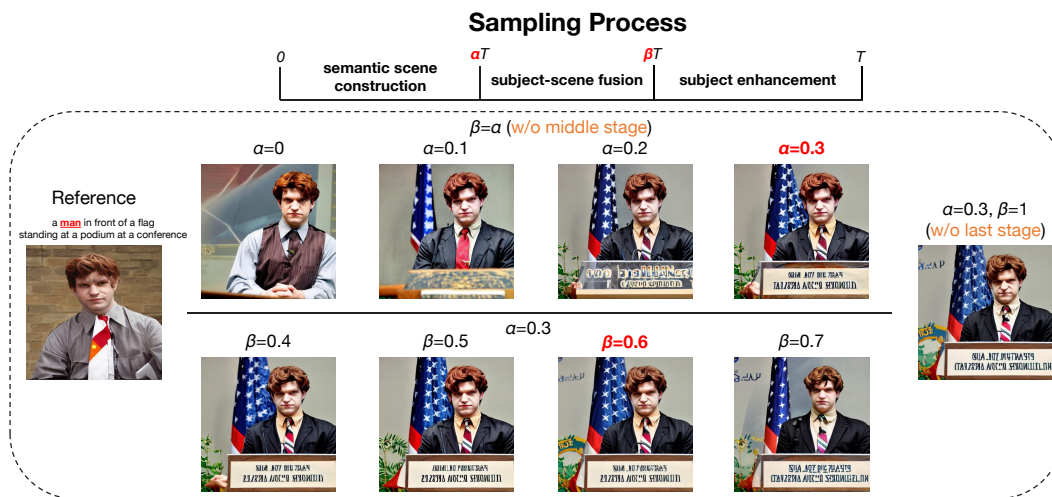

Figure 8. Visualized analysis of hyper-parameters α and β. We initially set β equal to α to explore optimal values for α. Once α is determined, we then vary β to determine its value. We also intentionally set $\beta = 1$ to assess the effectiveness of the last stage.

4.2. Results

In Tab. 1, we present a quantitative analysis comparing the performance of Face-Diffuser against baseline methods. Our method exhibits a significant advantage in both single-subject and multi-subject image generation. Notably, when compared to the previous state-of-the-art model Fastcomposer on multi-subject generation, Face-diffuser outperforms it by 0.132 and 0.084 in terms of identity preservation and prompt consistency, respectively. While Subject-diffusion and Fastcomposer excel in identity preservation compared to other methods, except our Face-diffuser, they tend to perform less satisfactorily in terms of prompt consistency, potentially due to their overfitting to the text prompts.

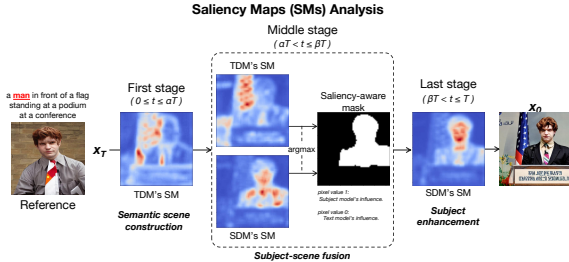Additionally, we provide some qualitative comparisons.

Figure 9. The visualization of average salience maps of our model in each sampling stage.

Table 1. Quantitative results on single- and multi-subject generation. IP denotes identity reservation and PC denotes prompt consistency. "N.A." indicates that the information is not available.

Methods	Single-Subject		Multi-Subject	
	IP ↑	PC ↑	IP ↑	PC ↑
ELITE (zero-shot)	0.228	0.146	N.A.	N.A.
Dreambooth (fine-tune)	0.273	0.239	N.A.	N.A.
Custom-Diffusion (fine-tune)	0.434	0.233	N.A.	N.A.
Subject-Diffusion(zero-shot)	0.605	0.228	0.432	0.205
Fastcomposer (zero-shot)	0.514	0.243	0.461	0.235
Face-Diffuser (zero-shot)	**0.708**	**0.325**	**0.593**	**0.319**

As illustrated in Fig. 5 and Fig. 6, our method excels in synthesizing more consistent persons with given reference images and semantic scenes compared to other baselines. For instance, in Fig. 5 (1st row), Fastcomposer and CustomDiffusion fail to generate the boy holding a piece of paper. Similarly, in Fig. 5 (3rd row), all compared methods except ELITE and CustomDiffusion, struggle to generate the scene of a man facing a group of people, but the two methods also produce unsatisfactory persons. In Fig. 6, it is evident that all compared methods fall short of producing satisfactory scenes.

Furthermore, we conducted additional comparisons with the two most recent methods, Fastcomposer and Subject-diffusion, using the samples provided in their original papers for further qualitative evaluation. The visualization results can be observed in Fig. 7, providing additional evidence of our method's superiority in high-fidelity image generation.

4.3. Hyperparameter and Effectiveness Analysis

Choice of α and β. We explore the optimal values for α and β in our work since they significantly impact the overall performance of our method. Fig. 8 shows the impact of varying the ratio of α and β.

Initially, we prioritize investigating the value of α and set $\beta = \alpha$ (ablating the middle stage) to explore the optimal timesteps for semantic scene construction (Fig. 8 (first row)). We observe that when $\alpha = \beta = 0$, it leads to subject over-fitting [33], which is expected due to the absence of TDM. As we increase the ratio, the generated semantic scene improves. When $\alpha = 0.3$, the scene is effectively constructed. However, we encounter issues related to coherence between the character and the scene, highlighting the importance of the middle stage.

Subsequently, we set $\alpha = 0.3$ to determine the value of β, which represents the sampling steps assigned to the semantic-scene fusion stage (Fig. 8 (second row)). As we increase the ratio, the coherence issue between the person and the scene improves, along with the overall quality of person generation. Optimal fusion between the person and the scene is achieved at $\beta = 0.6$, leading us to determine $\beta = 0.6$.

To further assess the significance of the subject enhance-

ment stage, we set $\alpha = 0.3$ and $\beta = 1$ (removing the last stage). By observing the results, the fidelity of character generation deteriorates. This underscores the substantial role of the subject enhancement stage in enhancing fidelity.

Visualization of Salience-adaptive Mask. We further analyze the effectiveness of our collaborative sampling process by visualizing the average salience map of each model, as shown in Fig. 9. These visualizations reveal that TDM mainly focuses on the semantic scene construction, while SDM places its emphasis on subject generation. This alignment with our design philosophy reinforces the distinct roles and responsibilities assigned to each model within the Face-diffuser framework.

5. Conclusion

Current subject-driven person-centric image generation models jointly learn the semantic scene and person generation by fine-tuning a common pre-trained diffusion model, which leads to training imbalance and quality compromises. In this paper, we propose Face-diffuser, an effective collaborative generation pipeline that develops two independent diffusion models for semantic scenes and person generation. The sampling process is divided into three sequential stages, i.e., semantic scene construction, subject-scene fusion, and subject enhancement. The subject-scene fusion stage, that is the collaboration achieved through our novel and highly effective mechanism, Saliency-adaptive Noise Fusion, which spatial blending of predicted noises from both models automatically in a saliency-aware manner in each step. Extensive experiments demonstrate Face-diffuser's effectiveness in generating high-quality images depicting multiple unseen persons in various scenarios.

6. Acknowledgments

References

[1] Omri Avrahami, Kfir Aberman, Ohad Fried, Daniel Cohen-Or, and Dani Lischinski. Break-a-scene: Extracting multiple concepts from a single image. *arXiv preprint arXiv:2305.16311*, 2023. 2

[2] Omri Avrahami, Thomas Hayes, Oran Gafni, Sonal Gupta, Yaniv Taigman, Devi Parikh, Dani Lischinski, Ohad Fried, and Xi Yin. Spatext: Spatio-textual representation for controllable image generation. In *CVPR*, pages 18370–18380, 2023. 3

[3] Daniel Bolya and Judy Hoffman. Token merging for fast stable diffusion. In *CVPR*, pages 4598–4602, 2023. 2

[4] Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T Freeman, Michael Rubinstein, et al. Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704*, 2023. 2

[5] Wenhu Chen, Hexiang Hu, Chitwan Saharia, and William W Cohen. Re-imagen: Retrieval-augmented text-to-image generator. *arXiv preprint arXiv:2209.14491*, 2022. 2

[6] Wenhu Chen, Hexiang Hu, Yandong Li, Nataniel Rui, Xuhui Jia, Ming-Wei Chang, and William W Cohen. Subject-driven text-to-image generation via apprenticeship learning. *arXiv preprint arXiv:2304.00186*, 2023. 2

[7] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, et al. Cogview: Mastering text-to-image generation via transformers. In *NeruIPS*, pages 19822–19835, 2021. 2

[8] Yuchao Feng, Jiawei Jiang, Honghui Xu, and Jianwei Zheng. Change detection on remote sensing images using dual-branch multilevel intertemporal network. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–15, 2023. 4

[9] Yuchao Feng, Yanyan Shao, Honghui Xu, Jinshan Xu, and Jianwei Zheng. A lightweight collective-attention network for change detection. In *ACM MM*, page 8195–8203, 2023. 4

[10] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *ICLR*, 2023. 2, 3, 6

[11] Rinon Gal, Moab Arar, Yuval Atzmon, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. Designing an encoder for fast personalization of text-to-image models. In *Siggraph*, 2023. 3

[12] Shaozhe Hao, Kai Han, Shihao Zhao, and Kwan-Yee K Wong. Vico: Detail-preserving visual condition for personalized text-to-image generation. *arXiv preprint arXiv:2306.00971*, 2023. 2

[13] Minguk Kang, Jun-Yan Zhu, Richard Zhang, Jaesik Park, Eli Shechtman, Sylvain Paris, and Taesung Park. Scaling up gans for text-to-image synthesis. In *CVPR*, pages 10124–10134, 2023. 2

[14] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, pages 4401–4410, 2019. 6

[15] Yuval Kirstain, Omer Levy, and Adam Polyak. X&fuse: Fusing visual information in text-to-image generation. *arXiv preprint arXiv:2303.01000*, 2023. 3

[16] 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. 2, 3, 4

[17] 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. 2, 3, 5, 6

[18] Daniel Roich, Ron Mokady, Amit H Bermano, and Daniel Cohen-Or. Pivotal tuning for latent-based editing of real images. *ACM Transactions on Graphics (TOG)*, 42(1):1–13, 2022. 3

[19] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 2, 4, 6

[20] 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. 2, 3, 4

[21] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. In *NeurIPS*, pages 36479–36494, 2022. 2

[22] Vishnu Sarukkai, Linden Li, Arden Ma, Christopher Ré, and Kayvon Fatahalian. Collage diffusion. *arXiv preprint arXiv:2303.00262*, 2023. 3

[23] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, pages 815–823, 2015. 6

[24] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. In *NeurIPS*, pages 25278–25294, 2022. 2

[25] Jing Shi, Wei Xiong, Zhe Lin, and Hyun Joon Jung. Instantbooth: Personalized text-to-image generation without test-time finetuning. *arXiv preprint arXiv:2304.03411*, 2023. 2, 3

[26] Chenyang Si, Ziqi Huang, Yuming Jiang, and Ziwei Liu. Freeu: Free lunch in diffusion u-net. *arXiv preprint arXiv:2309.11497*, 2023. 2

[27] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, pages 2256–2265, 2015. 2

[28] Yoad Tewel, Rinon Gal, Gal Chechik, and Yuval Atzmon. Key-locked rank one editing for text-to-image personalization. In *ACM SIGGRAPH*, pages 1–11, 2023. 2

[29] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 4

[30] Yibin Wang, Yuchao Feng, Jie Wu, Honghui Xu, and Jianwei Zheng. Ca-gan: Object placement via coalescing attention

based generative adversarial network. In *ICME*, pages 2375–2380, 2023. 2

[31] Yibin Wang, Haixia Long, Qianwei Zhou, Tao Bo, and Jianwei Zheng. Plsnet: Position-aware gcn-based autism spectrum disorder diagnosis via fc learning and rois sifting. *Computers in Biology and Medicine*, page 107184, 2023. 4

[32] 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. 2, 3

[33] 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. 2, 3, 5, 6, 8

[34] 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. 6

[35] Jing Zhao, Heliang Zheng, Chaoyue Wang, Long Lan, and Wenjing Yang. Magicfusion: Boosting text-to-image generation performance by fusing diffusion models. *arXiv preprint arXiv:2303.13126*, 2023. 3

[36] Guangcong Zheng, Xianpan Zhou, Xuewei Li, Zhongang Qi, Ying Shan, and Xi Li. Layoutdiffusion: Controllable diffusion model for layout-to-image generation. In *CVPR*, pages 22490–22499, 2023. 2