Face Aging via Diffusion-based Editing

Xiangyi Chen[1,2]
xiangyi.chen@telecom-paris.fr
Stéphane Lathuilière[2]
stephane.lathuiliere@telecom-paris.fr

[1] Shanghai Jiao Tong University
Shanghai, China

[2] LTCI, Télécom Paris
Institut Polytechnique de Paris
Palaiseau, France

Abstract

In this paper, we address the problem of *face aging*—generating past or future facial images by incorporating age-related changes to the given face. Previous aging methods rely solely on human facial image datasets and are thus constrained by their inherent scale and bias. This restricts their application to a limited generatable age range and the inability to handle large age gaps. We propose FADING, a novel approach to address **F**ace **A**ging via **DI**ffusion-based editi**NG**. We go beyond existing methods by leveraging the rich prior of large-scale language-image diffusion models. First, we specialize a pre-trained diffusion model for the task of face age editing by using an age-aware fine-tuning scheme. Next, we invert the input image to latent noise and obtain optimized null text embeddings. Finally, we perform text-guided local age editing via attention control. The quantitative and qualitative analyses demonstrate that our method outperforms existing approaches with respect to aging accuracy, attribute preservation, and aging quality.

1 Introduction

Have you ever looked in the mirror and wondered what you might look like in a few decades? Digital face aging techniques make it possible now. This exciting field aims to create realistic transformations of a person's face, simulating the effects of aging or de-aging. These techniques have critical applications in various fields including entertainment, forensics, and healthcare. A number of face-aging methods have been introduced. Most recent approaches [1, 7, 10, 11, 23, 28, 42] are based on deep generative models, such as generative adversarial networks (GANs) [8] and have shown promising results. But to our knowledge, all existing learning-based methods rely solely on datasets of human facial images (*e.g*. FFHQ [17] or CelebA [16, 22]), and are thus constrained by the inherent scale and bias of these datasets. For example, most methods have a limited transformation range (mostly less than 70 years old) and may fail when faced with large age gaps, occlusions, as well as extreme head poses due to the limited number of these rare cases in the dataset.

Meanwhile, the recently proposed diffusion models [13] exhibit comparable or even superior generation quality compared to GANs. In light of this, we propose to extend a diffusion-based large-scale language image model to tackle the specific task of face aging. Our motivation is that these models have learned, through language supervision, a rich image prior on a vast diversity of images, including faces, and have extensive semantic knowledge on diverse concepts (such as *"woman"/"man"*, *"glasses"*, etc) that could be potentially exploited for age editing. While some recent research [3, 4, 6, 12, 19, 24] has explored the potential of leveraging diffusion models for image editing tasks, they are limited to general-purpose editing methods. In contrast, no studies have demonstrated how these approaches can be adapted to tailor highly specific tasks such as face aging.

To this end, we propose FADING : Face Aging via DIffusion-based editiNG. The proposed method consists of two stages: specialization and editing. Specialization is a training stage where we re-target a pre-trained diffusion-based language-image model for face aging. In this stage, we employ an age-aware fine-tuning scheme that achieves better disentanglement of the age from age-irrelevant features (*e.g*. gender). For the editing stage, we first employ a well-chosen inversion technique to invert the input image into latent noise. Subsequently, we leverage a pair of text prompts containing both initial and target age information to perform text-based localized age editing, via attention control. Our contribution can be summarized as follows: (i) FADING is the first method to extend large-scale diffusion models for face aging; (ii) we successfully leverage the attention mechanism for accurate age manipulation and disentanglement; (iii) we qualitatively and quantitatively demonstrate the superiority of FADING over state-of-the-art methods through extensive experiments. [1]

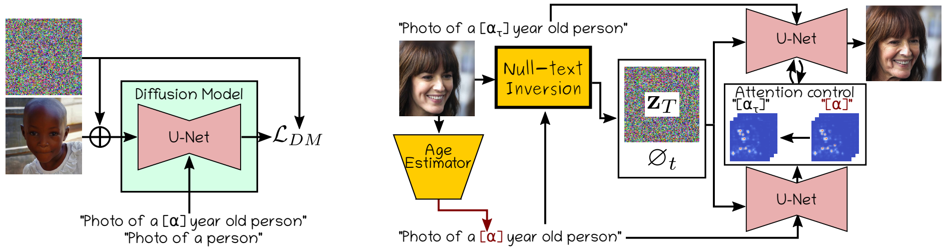
2 Related Work

Face-Aging Most of the recent methods rely on the well-known Generative Adversarial Networks (GANs) [8]. On the one hand, *condition-based* methods follow the conditional GAN framework [25]. This means they include age as an extra condition into the GAN framework to guide age-aware synthesis [2, 14, 21, 40, 43]. The age estimator can be embedded into the generator and trained simultaneously with it [21]. Alternatively, recurrent neural networks are used in [38, 39] to iteratively synthesize aging effects. Pre-trained face recognizers are employed to preserve age-irrelevant features (*i.e*. identity) [2, 14, 40, 41].

On the other hand, other methods [10, 11, 15, 23, 42] resort to *latent space manipulation* [9, 34]. An age modulation network is designed to fuse age labels with the latent vectors in HRFAE [42], or to output age-aware transformation to apply to the decoder in RAGAN [23]. SAM [1] relies on the latent space of a pre-trained GAN and employs an age regressor to explicitly guide the encoder in generating age-aware latent codes. Huang *et al.* [15] learn a unified embedding of age and identity. Some works also adopt a style-based architecture [17, 18]. LATS [28] follows StyleGAN2 [18] to perform modulated convolutions to inject learned age code into the decoder. CUSP [7] disentangles style and content representations and uses a decoder to combine the two representations with a style-based strategy. We highlight that one drawback of these methods is the significant discrepancy in identity that arises when real images are inverted into the GAN's latent space [37]. Consequently, the reconstruction of the initial image may be inaccurate, which can lead to suboptimal results.

Image editing with Diffusion Models (DMs) Large-scale diffusion models have raised the bar for text-to-image synthesis [29, 30, 32]. Naturally, works have attempted to adapt text-guided diffusion models to image editing. SDEdit [24] is among the first to propose diffusion-based image editing. It adds noise to the input image and then performs a text-guided denoising process from a predefined step. However, SDEdit lacks specific control

[1] Code available at https://github.com/MunchkinChen/FADING.



(a) Specialization to aging via fine-tuning of a pre-trained diffusion model.

(b) Age editing: given an input image, the diffusion process is inverted. The image is then edited replacing the estimated age with the target age.

Figure 1: FADING addresses **f**ace **a**ging via **di**ffusion-based editi**ng**: In the specialization stage, a pre-trained diffusion model is fine-tuned for the aging task. Editing is achieved via age estimation, image inversion, and attention control.

over edited region. With the help of a mask provided by the user, [3, 4, 27] better address this problem and enable more meaningful local editing. After each denoising step, the mask is applied to the latent image while also adding the noisy version of the original image. DiffEdit [6] gets rid of the need for a user-provided mask by automatically generating one that highlights regions to be edited based on the text description. Prompt-to-prompt [12] proposes a text-only editing technique based on a pair of *"before-after"* text descriptions. Null-text inversion [26] enables real image editing with prompt-to-prompt thanks to its accurate inversion of real images. Concurrently, Imagic [19] enables text-guided real image editing by fine-tuning the diffusion model to capture the input image's appearance. However, it is important to note that all these methods are general-purpose editing techniques. As such, our work aims to showcase the potential for adapting these broad approaches for use in more specific tasks, such as face aging.

3 FADING: Face Aging via DIffusion-based editiNG

The objective of this work is to transform an input image $\mathbf{x}$ to make the person in the image appear to be of a specific target age α_τ. For this, we employ a dataset of N face images $\mathbf{x}^{(n)} \in \mathbb{R}^{H\times W\times 3}$, $n = 1,...,N$ with their corresponding age labels $\alpha^{(n)} \in \{1,..K\}$, where K is the maximum age in our training dataset. The age labels $\alpha^{(n)}$ can be obtained either via manual labeling or using a pre-trained age classifier.

The proposed approach relies on a specialization and an edition stage illustrated in Figure 1. In the first stage, a pre-trained diffusion model is re-targeted for the task of face age editing. This training procedure is detailed in Sec. 3.1. To better disentangle age information from other age-irrelevant features, our specialization procedure employs an age-aware fine-tuning scheme. Then, our inference consists of two steps: inversion and editing. In the inversion step, we inverse the diffusion process using a recent optimization-based inversion [26] as detailed in Sec. 3.2. In the editing step, we use a new prompt that contains the target age to guide a localized age editing with attention control (see Sec. 3.3). We also provide a solution to improve the prompts used for editing to achieve higher image quality.

3.1 Specialization to Face Aging

FADING leverages a pre-trained text-to-image Diffusion Model (DM) [13]. While the proposed method could be applied to any text-to-image DM, in our experiments, we employ a variant of DM named Latent Diffusion Model (LDMs) [30]. LDMs operate in the latent space of an auto-encoder to achieve lower computation complexity. As traditional DMs, LDMs are composed of a forward and a backward pass.

In the forward process, the input image $\mathbf{x}_0$ is projected to the auto-encoder latent space, $\mathbf{z}_0 = \mathcal{E}(\mathbf{x}_0)$. Then, random Gaussian noises are added to the original latent embedding $\mathbf{z}_0$ in a stepwise manner to create a sequence of noisy samples $(\mathbf{z}_1 ..., \mathbf{z}_T)$. Learning an LDM consists in training a neural network ε_θ to estimate the corresponding noise from a given sample $\mathbf{z}_t$. In the reverse process, on the other hand, new data points are generated by sampling from a normal distribution and gradually denoising the sample using ε_θ. The generated image $\hat{\mathbf{x}}_0$ is obtained by feeding the estimated latent tensor $\hat{\mathbf{z}}_0$ to the decoder. To enable generation conditioned on a text prompt $\mathcal{P}$, a sequence of token embeddings is extracted from $\mathcal{P}$ and given to ε_θ via cross-attention layers, where keys and values are estimated from the token embedding. In the case of unconditional generation, the token embeddings are replaced by fixed embeddings referred to as *null-text embedding* and denoted by $\varnothing_t$.

Age editing with a pre-trained DM can be performed without any training stage [24, 26], but this produces unsatisfactory results since they are generally not specialized for human faces. Also, coarse conditioning prompts, such as "*man in his thirties*", can capture age-related semantics but we observe that they often fail to capture more specific textual descriptions of age as numbers, such as "*32-year-old man*". To address these issues, we propose a specialization stage that re-purposes a pre-trained DM toward the aging task. For every face image $\mathbf{x}$ with its corresponding age α, fine-tuning is performed using an image-prompt pair, with the following prompt: $\mathcal{P}_\alpha$="*photo of a* $[\alpha]$ *year old person*", where α is the age of the person written as numerals. We have observed better performance when adding another age-agnostic prompt $\mathcal{P}$="*photo of a person*" at every iteration. We refer to this fine-tuning scheme as the *double-prompt* scheme. One assumption to justify this observation is that it can allow better disentangling of age information from other age-irrelevant features (*i.e.* identity and context features). Regarding the training loss, we employ the reconstruction objective of DMs which, in our case, can be written as follows:

$$\mathcal{L}_{DM} = \mathbb{E}_{\mathbf{z}_0 \sim \mathcal{E}(x), \alpha, \varepsilon, \varepsilon', t}[||\varepsilon - \varepsilon_\theta(\mathbf{z}_t, t, \mathcal{P})||_2^2 + ||\varepsilon' - \varepsilon_\theta(\mathbf{z}'_t, t, \mathcal{P}_\alpha)||_2^2], \tag{1}$$

where ε and ε' are random Gaussian noises, and $\mathbf{z}_t$ and $\mathbf{z}'_t$ are the respective noisy latent codes obtained from $\mathbf{z}_0$. To preserve the rich image prior learned by the DM, we restrict the number of fine-tuning steps to a small value, typically around 150 steps.

3.2 Age Editing: Image Inversion

After the specialization stage, our DM can generate face images either unconditionally or conditionally on a target age α with prompts $\mathcal{P}$ and $\mathcal{P}_\alpha$ respectively. To enable real image editing, we need to inverse the diffusion process of the input image. In this task, we leverage an inversion algorithm, known as *null-text inversion* [26], which consists in modifying the unconditional textual embedding that is used for classifier-free guidance such that it leads to accurate reconstruction. To be specific, we use the specialized model to invert the input image $\mathbf{x}$ to the noise space through DDIM inversion [35]. We obtain a diffusion trajectory $\{z_t^{inv}\}, t = 1 \dots T$ from Gaussian noise to the input image. Unfortunately, previous studies [35] show that classifier-free guidance amplifies the accumulated error of DDIM inversion, resulting in poor reconstruction of $\mathbf{x}$. *Null-text inversion* optimizes the null-text embedding $\varnothing_t$ used in classifier guidance at every step t such that, assuming a conditioning prompt $\mathcal{P}_{inv}$ corresponding to the input image, the forward process leads to an accurate reconstruction of $\mathbf{x}$. The unconditionally inverted sequence of noisy latents $\{z_t^{inv}\}_{t=1}^{T}$ serves as our pivot trajectory for optimization: the unconditional null embeddings over all time-steps $\{\varnothing_t\}_{t=1}^{T}$ are sequentially optimized such that the noise estimator network ε_θ predicts latent codes close to z_{t-1}^{inv} at every step t. More precisely, for every step t in the order of the diffusion process $t = T \rightarrow t = 1$, the following minimization problems are sequentially considered:

$$\min_{\varnothing_t} \| z_{t-1}^{inv} - z_{t-1}(\bar{z}_t, t, \mathcal{P}_{inv}; \varnothing_t) \|_2^2 \tag{2}$$

where $\bar{z}_t$ is the noisy latent code obtained by solving the optimization problem of the previous step, and z_{t-1} is the latent code at step $t-1$ estimated using $\bar{z}_t$. To enable age editing, we need to provide a prompt corresponding to the content of the input image. In this task, we propose to employ a pre-trained age estimator. Assuming an input image $\mathbf{x}$, we obtain its estimated age α and employ as prompt $\mathcal{P}_{inv} = \mathcal{P}_\alpha =$"*photo of a* $[\alpha]$ *year old person*".

3.3 Age Editing: Localized Age Editing with Attention Control

We now explain how we edit an image $\mathbf{x}$ to make the person in the image appear to be of a target age α_τ. To achieve this, we take inspiration from recent literature [12] and act on the cross-attention maps used for text-conditioning, forcing the model to modify only age-related areas via attention map injection. After inversion, we know the latent noise $\mathbf{z}_T$ and the optimized unconditional embeddings $\{\varnothing_t\}_{t=1}^{T}$ leads to an accurate reconstruction of $\mathbf{x}$ when conditioned on prompt $\mathcal{P}_\alpha$. In every cross-attention layer of ε_θ, we compute the reference cross-attention maps generated during the diffusion process $\{M_t^\alpha = \text{Softmax}(Q_t^{\mathbf{Z}} K_t^\alpha)\}_{t=1}^{T}$, where $Q_t^{\mathbf{Z}}$ are queries computed from $\mathbf{z}_t$ and K_t^α keys computed from the prompt $\mathcal{P}_\alpha$. As shown in [12, 36], these attention maps contain rich semantic relations between the spatial layout of the image and each word in $\mathcal{P}_\alpha$. In our case, the attention maps corresponding to the token $[\alpha]$ indicate which pixels are related to the age of the person.

Next, we replace the initial estimated age α in the inversion prompt $\mathcal{P}_\alpha$ with a target age α_τ and obtain a new target prompt $\mathcal{P}_\tau$="*photo of a* $[\alpha_\tau]$ *year old person*". We then use $\mathcal{P}_\tau$ to guide the generation: during the new sampling process, we inject the cross-attention maps $\{M_t^\alpha\}_{t=1}^{T}$, but keep the cross-attention values from the new prompt $\mathcal{P}_\tau$. In this way, the generated image is conditioned on the target age information provided by the target prompt $\mathcal{P}_\tau$ through the cross-attention values, while preserving the original spatial structure. Specifically, as only age-related words are modified in the new prompt, only pixels that attend to age-related tokens receive the greatest attention. Note that, we follow [12] and perform a soft attention constraint by swapping only the first t_M steps, as the attention maps play an important role mostly in the early stages.

Enhancing prompts FADING can achieve satisfying aging performance with the very generic prompts given above. Nevertheless, the results can be further improved by using more specific prompts in the inversion and editing stages. While this can be achieved with manual prompt engineering, we propose a simple and automatic way to improve our initial prompts $\mathcal{P}_\alpha$ and $\mathcal{P}_\tau$. First, we can leverage pre-trained gender classifiers to predict the gender of the person in the input image. Then, the word *"person"* in both $\mathcal{P}_\alpha$ and $\mathcal{P}_\tau$ can be replaced by either *"woman"* or *"man"*. Second, our experiments show that in the case of young ages, either in $\mathcal{P}_\alpha$ and $\mathcal{P}_\tau$, the use of words such as *"person"*, *"woman"* or *"man"* do not perform well. Therefore, if the target age α_τ or the age α estimated by our classifier is below 15, the words *"woman/man"* are replaced by *"girl/boy"* in $\mathcal{P}_\tau$ or $\mathcal{P}_\alpha$.

4 Experiments

Implementation details We employ Stable Diffusion [30] pre-trained on the LAION-400M dataset [33]. 150 training images are sampled from FFHQ-Aging[28] to finetune the pre-trained model for 150 steps, with a batch size of 2. We used the central age of the true label age group as α in the finetuning prompt $\mathcal{P}_\alpha$. We employed Adam optimizer with a learning rate of 5×10^{-6} and $\beta_1 = 0.9$, $\beta_2 = 0.999$. During attention control, we set the cross-attention replacing ratio t_M/T to 0.8. All experiments are conducted on a single A100 GPU. It takes 1 minute for finetuning, 1 minute for inversion, and 5 seconds for age editing.

Evaluation protocol We utilized two widely-used high-resolution **datasets** as in [7]. *FFHQ-Aging* [28] is an extension of the NVIDIA FFHQ [17] dataset containing 70k 1024×1024 resolution images. Images are manually labeled into 10 age groups ranging from 0-2 to 70+ years old. *CelebA-HQ* [16] consists of 30k images. This dataset is used only for evaluation, not for training. Age labels are obtained using the DEX classifier [31] as used in previous studies [42, 43]. Images are downsampled to 512×512 resolution for our experiments. Regarding the **metrics**, we evaluate aging methods from three perspectives: aging accuracy, age-irrelevant attribute reservation, and aging quality. Following [7], we employ: *Mean Absolute Error (MAE)*: the prediction of an age estimator is compared with the target age. *Gender, Smile, and Face expression preservation*: we report the percentage to which the original attribute is preserved. *Blurriness*: indicates face blur condition. *Kernel-Inception Distance* [5] assesses the discrepancy between generated and real images for similar ages. We report the KID between original and generated images within the same age groups. For evaluation, Face++[2] is used for aging accuracy, attribute preservation, and blurriness evaluation.

4.1 Comparison with State-of-the-Art

We conduct comparisons with state-of-the-art aging approaches, including HRFAE [42], LATS [28], and CUSP [7]. We are unable to include Re-aging GAN [23], another recent aging method, in our comparison due to the unavailability of its source code. Moreover, the lack of detailed information regarding its evaluation protocol prevents us from conducting a fair and reliable comparison following its evaluation protocol. We start the comparison on the CelebA-HQ[16] dataset. In this case, we follow the evaluation protocol used in [42] and sample 1000 test images with *"young"* labels and translate them to the target age of 60.

Qualitative comparison The comparative study on CelebA-HQ is shown in Figure 2. Note that these images are extracted from [7], and consequently have not been cherry-picked. We observe that FaderNet [20] introduces little modifications, PAG-GAN [41] and

[2] Face++ Face detection API: https://www.faceplusplus.com/

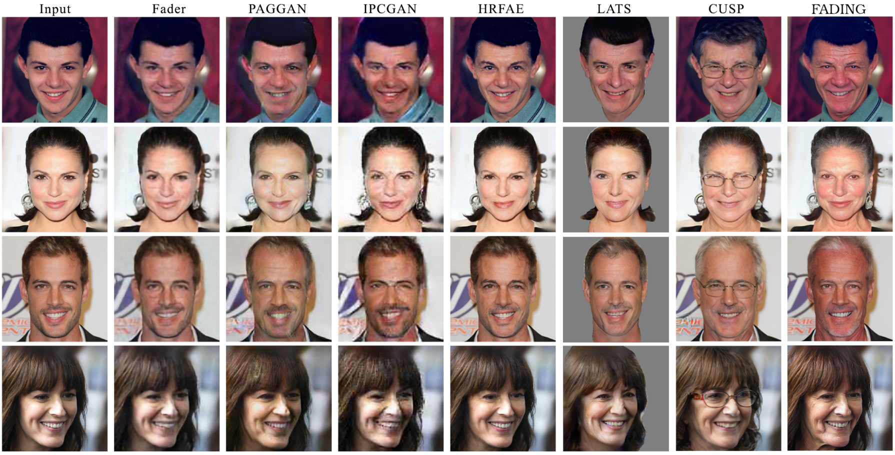


Figure 2: Qualitative comparison with state-of-the-art methods on CelebA-HQ. Images for the other approaches are extracted from [7].

IPC-GAN [40] produce pronounced artifacts or degradation. HRFAE [42] generates plausible aged faces with minor artifacts but is mostly limited to skin texture changes, such as adding wrinkles. LATS [28], CUSP [7], and our approach introduce high-level semantic changes, such as significant receding of the hairline (see third row). But LATS operates only in the foreground; it does not deal with backgrounds or clothing and requires a previous masking procedure. On the other hand, CUSP always introduces glasses with aging. This is likely due to the high correlation between age and glasses in their training set. Our method does not introduce these undesired additional accessories, produces fewer artifacts on backgrounds, and possesses more visual fidelity to the input image.

We now expand the comparison with the best-performing competitor, namely CUSP [7], on FFHQ-Aging [28]. We translate input images to all age groups and report per-age-group results, for a more comprehensive analysis with a complete sense of continuous transformation throughout the lifespan. Figure 3 shows qualitative results. We have the following key observations. (1) In general, our approach introduces fewer artifacts, generates realistic textural and semantic modification, and achieves better visual fidelity across all age groups. (2) We achieve significant improvement for extreme target ages (infant and elderly, see columns for (4-6) and (70+)). (3) Our model handles better rare cases, such as accessories or occlusions. CUSP fails when the source person wears facial accessories. Typically, for the person on the right who wears sunglasses, CUSP falsely translates sunglasses to distorted facial components. In contrast, our method preserves accessories accurately while correctly addressing structural changes elsewhere. These results confirm our initial hypothesis that utilizing a specialized DM pre-trained on a large-scale dataset increases robustness compared to methods exclusively trained on facial datasets, which are susceptible to data bias.

Interestingly, we observe a slight variation in skin tone when addressing age change with FADING. It is important to note that a similar shift in skin tone is also observed for the training-free baseline (vanilla implementation of prompt-to-prompt editing using pretrained Stable Diffusion, referred to as Training-free in Table 3), as shown in Figure 4b (see more results in supplementary material). This suggests that the entanglement between age and skin tone is inherent to the pre-trained Stable Diffusion model and is not a result of our

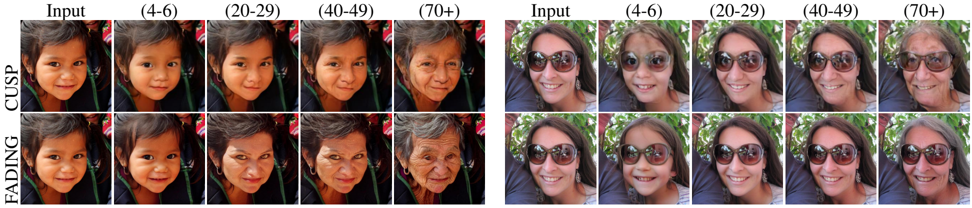


Figure 3: Qualitative comparison with state-of-the-art methods on FFHQ-Aging. For CUSP, we translate each image to the corresponding age group. For FADING, we translate to the central age of each group. For the oldest age group (70+), we translate to 80 years old.

Table 1: Quantitative comparison on CelebA-HQ on the young-to-60 task. Except for FADING, the scores are extracted from [7].

Method	Predicted Age	Blur	Gender	Smiling	Neutral	Happy
Real images	68.23 ± 6.54	2.40	-	-	-	-
FaderNet [20]	44.34 ± 11.40	9.15	97.60	95.20	90.60	92.40
PAGGAN [41]	49.07 ± 11.22	3.68	95.10	93.10	90.20	91.70
IPCGAN [40]	49.72 ± 10.95	9.73	96.70	93.60	89.50	91.10
HRFAE [42]	54.77 ± 8.40	**2.15**	97.10	**96.30**	**91.30**	**92.70**
HRFAE-224 [42]	51.87 ± 9.59	5.49	97.30	95.50	88.30	92.50
LATS [28]	55.33 ± 9.33	4.77	96.55	92.70	83.77	88.64
CUSP [7]	**67.76** ± **5.38**	2.53	93.20	88.70	79.80	84.60
FADING (Ours)	66.49 ± 6.46	2.35	**98.40**	90.20	84.50	86.80

specialization stage.

Quantitative comparison Table 1 presents quantitative results on CelebA-HQ[16] dataset. Note that an 8.23-year discrepancy is reported between the DEX classifier utilized for inference and the Face++ classifier utilized for evaluation[7]. FADING is on par with CUSP for aging accuracy. We achieve the highest gender preservation, proving our capability to retain age-irrelevant features. However, we report lower scores for other attributes. As is discussed in the qualitative analysis, this is because previous methods primarily generate texture-level modification, which preserves high-level attributes. In contrast, FADING yields more profound but realistic semantic changes, thus slightly compromising preservation metrics.

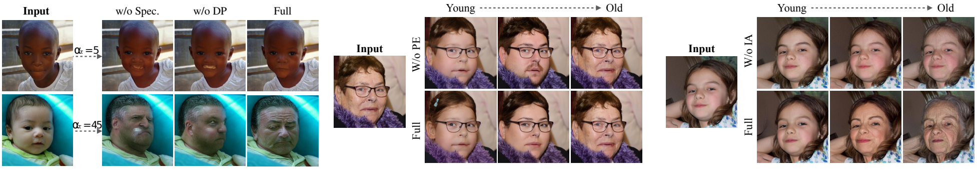
Table 2 presents quantitative results on FFHQ-Aging[28] dataset. Lower MAE suggests that we have a better aging accuracy. FADING also reports better gender preservation for most age groups. Note that, for middle-aged group from 30-50, an almost perfect preservation rate is achieved. Our qualitative analysis is supported by the quantitative KID analysis, with one order of magnitude lower than CUSP for nearly all age groups. Again this demonstrates that FADING achieves higher aging performance.

4.2 Ablation studies

Specialization (Spec.) and Double-Prompt (DP) scheme To assess the influence of the design of the specialization step, we consider a variant where we skip the specialization step and directly use a pre-trained Stable Diffusion instead. This baseline can be seen as a vanilla implementation of prompt-to-prompt editing [12] with null-text inversion [26] in the case of aging. The second variant includes the specialization step but omits the double-prompt scheme. The results shown in Figure 4a and Table 3 demonstrate the effectiveness

Table 2: Quantitative comparison between CUSP and FADING on FFHQ-Aging.

Metric	Method	0-2	3-6	7-9	10-14	15-19	20-29	30-39	40-49	50-69	70+	Mean
MAE	CUSP	9.41	16.28	20.24	18.16	11.88	10.36	12.70	**11.08**	**8.13**	8.05	12.63
	FADING	**5.70**	**11.72**	**13.66**	**11.22**	**6.86**	**6.23**	**9.60**	12.04	8.39	**6.20**	**9.16**
Gender(%)	CUSP	71.5	**73.5**	**74.5**	**78.0**	73.5	80.5	85.5	81.5	82.0	76.0	77.7
	FADING	**72.0**	72.0	67.5	68.0	**88.0**	**96.0**	**98.0**	**97.0**	**95.0**	**87.5**	**84.1**
KID($\times$100)	CUSP	4.19	3.22	3.14	3.18	3.60	3.63	3.98	4.69	4.07	4.57	3.83
	FADING	**1.41**	**0.11**	**0.45**	**0.25**	**0.52**	**0.16**	**1.00**	**0.59**	**1.50**	**0.61**	**0.66**



(a) Specialization step

(b) Enhanced Prompts

(c) Initial Age

Figure 4: Qualitative ablation studies on several aspects of FADING: impact of the specialization step (*Spec.*), the use of Double Prompt (*DP*) , the Enhanced Prompts (*EP*) and the use of the Initial Age (*IA*).

of our specialization step in generating more realistic images. Our qualitative analysis indicates that the images edited with a non-specialized model exhibit noticeable aberrations, especially around the mouth area and facial contours. The quantitative metrics also support the observation that our method achieves higher aging quality (lowest blurriness and KID). Furthermore, the training-free editing approach reports the highest aging error and a low attribute preservation rate. Regarding our double-prompt scheme, Figure 4a shows that it improves the structural alignment with the original image. Quantitatively, as shown in Table 3, the slight increase in age-MAE brought by *DP* is vastly complemented by the large gains in attribute preservation metrics. This improvement suggests that the *DP* indeed enhances the disentanglement of age from age irrelevant features by keeping them better retained. Besides, the age-MAE metric may be a less strong indicator of disentanglement capability, given that differences of 0.38 year in facial appearance are often imperceptible in real photos.

Enhanced Prompts (EP) and Initial Age (IA) We now analyze the edition stage considering two other variants: one without our enhanced prompts and another which does not use the initial age of the source image and instead uses $(\mathcal{P}, \mathcal{P}_\tau)$ as editing prompts. The positive impacts of enhanced prompts and the use of the estimated initial age are demonstrated in Table 4 where we observe consistent gains in all metrics. Qualitatively, *EP* plays an important role in preserving age irrelevant attributes: we observe significant improvements in gender consistency in Figure 4b. Surprisingly, the use of gender information in our enhanced prompts also helps to improve aging accuracy. We hypothesize that this is because more detailed prompts (we assume that "woman" contains more information than "person") lead to more specialized attention maps for each semantic component, resulting in more accurate targeting of age-related pixels. The impact of *IA* is illustrated in Figure 4c. Without information on the initial age, the appearance of the person barely changes, except for slight variations in hair color. This indicates that the use of initial age (*IA*) in guiding prompts prevents the model from reproducing the original image without effectively addressing the

Table 3: Ablation study on the Specialization stage

Method	Spec.	DP	MAE	Gender	Smiling	Happy	Neutral	Blur	KID($\times 100$)
Training-free ∼[26]	✗	-	9.295	82.40	82.95	78.35	78.80	2.226	0.668
Single prompt	✓	✗	**8.781**	81.95	85.05	81.55	81.05	2.275	0.707
Full	✓	✓	9.162	**84.10**	**86.60**	**81.95**	**81.75**	**2.030**	**0.660**

Table 4: Ablation study on the Editing stage

Method	MAE	Gender	KID($\times 100$)
w/o EP	9.830	79.90	0.668
w/o IA	13.703	80.05	1.164
Full	**9.162**	**84.10**	**0.660**

age change.

5 Conclusion

In this paper, a novel method for face age editing based on diffusion models was presented. The proposed model leverages the rich image and semantic prior of large-scale text-image models, via a training stage that specializes the diffusion model for aging tasks. Qualitative and quantitative analyses on two different datasets demonstrated that our method produces natural-looking re-aged faces across a wider range of age groups with higher re-aging accuracy, better aging quality, and greater robustness compared to state-of-the-art methods. The effectiveness of each component of our method was also validated through extensive experiments. In future works, we plan to extend our enhanced prompts strategy to preserve other age-agnostic attributes by leveraging corresponding pre-trained attribute classifiers. For example, we could include *"wearing glasses"* in the editing prompts when glasses are detected.

Acknowledgments

This paper has been supported by the French National Research Agency (ANR- 20-CE23-0027).

References

[1] Yuval Alaluf, Or Patashnik, and Daniel Cohen-Or. Only a matter of style: Age transformation using a style-based regression model. *ACM Transactions on Graphics (TOG)*, 2021.

[2] Grigory Antipov, Moez Baccouche, and Jean-Luc Dugelay. Face aging with conditional generative adversarial networks. *IEEE International Conference on Image Processing (ICIP)*, 2017.

[3] Omri Avrahami, Ohad Fried, and Dani Lischinski. Blended latent diffusion. *arXiv:2206.02779*, 2022.

[4] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[5] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. *International Conference on Learning Representations (ICLR)*, 2018.

[6] Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Diffedit: Diffusion-based semantic image editing with mask guidance. *International Conference on Learning Representations (ICLR)*, 2022.

[7] Guillermo Gomez-Trenado, Stéphane Lathuilière, Pablo Mesejo, and Óscar Cordón. Custom structure preservation in face aging. *Computer Vision–European Conference on Computer Vision (ECCV)*, 2022.

[8] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 2020.

[9] Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. Ganspace: Discovering interpretable gan controls. *Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

[10] Sen He, Wentong Liao, Michael Ying Yang, Yi-Zhe Song, Bodo Rosenhahn, and Tao Xiang. Disentangled lifespan face synthesis. *International Conference on Computer Vision (ICCV)*, 2021.

[11] Zhenliang He, Meina Kan, Shiguang Shan, and Xilin Chen. S2gan: Share aging factors across ages and share aging trends among individuals. *International Conference on Computer Vision (ICCV)*, 2019.

[12] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv:2208.01626*, 2022.

[13] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

[14] Gee-Sern Hsu, Rui-Cang Xie, and Zhi-Ting Chen. Wasserstein divergence gan with cross-age identity expert and attribute retainer for facial age transformation. *IEEE Access*, 2021.

[15] Zhizhong Huang, Junping Zhang, and Hongming Shan. When age-invariant face recognition meets face age synthesis: A multi-task learning framework. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[16] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *International Conference on Learning Representations (ICLR)*, 2018.

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

[18] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[19] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. *arXiv:2210.09276*, 2022.

[20] Guillaume Lample, Neil Zeghidour, Nicolas Usunier, Antoine Bordes, Ludovic Denoyer, and Marc'Aurelio Ranzato. Fader networks: Manipulating images by sliding attributes. *Conference on Neural Information Processing Systems (NeurIPS)*, 2017.

[21] Zeqi Li, Ruowei Jiang, and Parham Aarabi. Continuous face aging via self-estimated residual age embedding. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[22] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. *International Conference on Computer Vision (ICCV)*, 2015.

[23] Farkhod Makhmudkhujaev, Sungeun Hong, and In Kyu Park. Re-aging gan: toward personalized face age transformation. *International Conference on Computer Vision (ICCV)*, 2021.

[24] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. *International Conference on Learning Representations (ICLR)*, 2021.

[25] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv:1411.1784*, 2014.

[26] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. *arXiv:2211.09794*, 2022.

[27] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob Mcgrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *International Conference on Machine Learning (ICML)*, 2022.

[28] Roy Or-El, Soumyadip Sengupta, Ohad Fried, Eli Shechtman, and Ira Kemelmacher-Shlizerman. Lifespan age transformation synthesis. *Computer Vision–European Conference on Computer Vision (ECCV)*, 2020.

[29] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv:2204.06125*, 2022.

[30] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[31] Rasmus Rothe, Radu Timofte, and Luc Van Gool. Dex: Deep expectation of apparent age from a single image. *International Conference on Computer Vision (ICCV) workshops*, 2015.

[32] 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. *Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

[33] Christoph Schuhmann, Robert Kaczmarczyk, Aran Komatsuzaki, Aarush Katta, Richard Vencu, Romain Beaumont, Jenia Jitsev, Theo Coombes, and Clayton Mullis. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *NeurIPS Workshop Datacentric AI*, 2021.

[34] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of gans for semantic face editing. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[35] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *International Conference on Learning Representations (ICLR)*, 2020.

[36] Raphael Tang, Akshat Pandey, Zhiying Jiang, Gefei Yang, Karun Kumar, Jimmy Lin, and Ferhan Ture. What the daam: Interpreting stable diffusion using cross attention. *arXiv:2210.04885*, 2022.

[37] Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. Designing an encoder for stylegan image manipulation. *ACM Transactions on Graphics (TOG)*, 2021.

[38] Wei Wang, Zhen Cui, Yan Yan, Jiashi Feng, Shuicheng Yan, Xiangbo Shu, and Nicu Sebe. Recurrent face aging. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[39] Wei Wang, Yan Yan, Zhen Cui, Jiashi Feng, Shuicheng Yan, and Nicu Sebe. Recurrent face aging with hierarchical autoregressive memory. *IEEE transactions on pattern analysis and machine intelligence*, 2018.

[40] Zongwei Wang, Xu Tang, Weixin Luo, and Shenghua Gao. Face aging with identity-preserved conditional generative adversarial networks. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[41] Hongyu Yang, Di Huang, Yunhong Wang, and Anil K Jain. Learning continuous face age progression: A pyramid of gans. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[42] Xu Yao, Gilles Puy, Alasdair Newson, Yann Gousseau, and Pierre Hellier. High resolution face age editing. *International Conference on Pattern Recognition (ICPR)*, 2021.

[43] Zhifei Zhang, Yang Song, and Hairong Qi. Age progression/regression by conditional adversarial autoencoder. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.