# Supplementary Materials for Pluralistic Aging Diffusion Autoencoder

Peipei Li[1], Rui Wang[1], Huaibo Huang[2], Ran He[2], Zhaofeng He[1⋆]
[1]Beijing University of Posts and Telecommunications
[2]CRIPAC&MAIS, Institute of Automation, Chinese Academy of Sciences
{lipeipei, wr_bupt, zhaofenghe}@bupt.edu.cn, huaibo.huang@cripac.ia.ac.cn, rhe@nlpr.ia.ac.cn

In this supplementary material, we first introduce the theory validation in Sec. 1. Then, we show the training and inference algorithms in Sec. 2. Additional qualitative and quantitative comparison results are shown in Sec. 3. In this section, we propose a straightforward technique for achieving gender transformation utilizing the proposed PAEs. Meanwhile, we conduct disentanglement experiments in terms of different timesteps and compare PADA with previous method on Morph and CACD2000. Also, we measure the diversity boundary of PADA. Finally, we show more pluralistic face aging results in Sec. 4, including reference-guided face aging, text-guided face aging, diverse face aging, and intermediate generation results of diffusion decoder.

## 1. Theory Validation

**Theorem 1.** In the normalized CLIP latent space, according to the Law of Cosines, the Euclidean distance $D(e^{age}, e^{txt})$ between probabilistic aging embedding $e^{age}$ and text-based age representation $e^{txt}$ is optimally equivalent to the cosine similarity.

**Proof.** In practice, we normalize all the features in the CLIP latent space by $L_2$ norm. Hence, according to the Law of Cosines, the equivalent form of $D(e^{age}, e^{txt})$ can be rewritten as:

$$\begin{aligned}&D(e^{age}, e^{txt}) = \left\|e^{age} - e^{txt}\right\|_2^2 \\ &= \left\|e^{age}\right\|_2^2 + \left\|e^{txt}\right\|_2^2 - 2\left\|e^{age}\right\|_2\left\|e^{txt}\right\|_2 cos(e^{age}, e^{txt}) \\ &= 2 - 2cos(e^{age}, e^{txt})\end{aligned}$$

Therefore, when calculating the loss $L_{tKL}$, the optimization objectives for the Euclidean distance $D(e^{age}, e^{txt})$ and cosine distance $-cos(e^{age}, e^{txt})$ are equivalent.

**Theorem 2.** For directly sampling PAE from text-based age prior, the Euclidean distance $D$ between the probabilistic aging embedding $e^{age}$ and corresponding aging text representation $e^{txt}$ for $\forall m^*$ satisfies $D(e^{txt}, e^{age}) \leq m^*$ with probability at least:

$$\begin{aligned}&Prob(D(e^{txt}, e^{age}) \leq m^*) \\ &= 1 - \int_{-1}^{1-\frac{m^*}{2}-\frac{m^*}{2\epsilon}} \frac{\Gamma(d/2+1/2)}{\sqrt{\pi}\Gamma(d/2)}(1-x^2)^{d/2-1}dx,\end{aligned}$$

where $\Gamma(\cdot)$ is Gamma function, i.e. $\Gamma(\cdot) = \int_0^\infty x^{t-1}e^{-x}dx$. $d$ is the dimension of input feature, $\epsilon$ is hyperparameter for sampling intensity, and $\eta$ is normalized sampling from normal Gaussian distribution.

**Proof.** In practice, we normalize the features in CLIP latent space by $L_2$ norm.

According to the Law of Cosines, we get:

$$\begin{aligned}D(e^{txt}, e^{age}) &= \left\|e^{txt} - e^{age}\right\|_2^2 \\ &= 2\left(1 - \frac{(e^{txt})^T e^{age}}{\left\|e^{txt}\right\|_2\left\|e^{age}\right\|_2}\right) \\ &= 2\left(1 - \frac{\left\|e^{txt}\right\|_2^2 + \epsilon\cdot(e^{txt})^T\eta}{\left\|e^{txt}+\epsilon\cdot\eta\right\|_2}\right) \\ &= 2\left(1 - \frac{1 + \epsilon\cdot(e^{txt})^T\eta}{\left\|e^{txt}+\epsilon\cdot\eta\right\|_2}\right) \\ &\leq 2\left(1 - \frac{1 + \epsilon\cdot(e^{txt})^T\eta}{\left\|e^{txt}\right\|_2+\epsilon\left\|\eta\right\|_2}\right) \\ &= 2\left(1 - \frac{1 + \epsilon\cdot(e^{txt})^T\eta}{1+\epsilon}\right)\end{aligned}$$

Therefore, we find an lower bound for our original probability:

$$\begin{aligned}&Prob(D(e^{txt}, e^{age}) \leq m^*) \\ &\geq Prob\left(2\left(1 - \frac{1 + \epsilon\cdot(e^{txt})^T\eta}{1+\epsilon}\right) \leq m^*\right) \\ &= 1 - Prob\left((e^{txt})^T\eta \leq 1 - \frac{m^*}{2} - \frac{m^*}{2\epsilon}\right)\end{aligned}$$

In [2], the Cumulative Distribution Function (CDF) of the inner product of two random vectors, i.e. $x = u^T v$ on a standard unit sphere is:

⋆Corresponding author

$$F(x) = \int_{-1}^{x} \frac{\Gamma(d/2+1/2)}{\sqrt{\pi}\Gamma(d/2)}(1-x^2)^{d/2-1}dx \quad (1)$$

Thus, we complete our proof:

$$\begin{aligned} &Prob\left(D(e^{txt}, e^{age}) \leq m^*\right) \\ &\geq 1 - Prob\left((e^{txt})^T\eta \leq 1 - \frac{m^*}{2} - \frac{m^*}{2\epsilon}\right) \\ &= 1 - \int_{-1}^{1-\frac{m^*}{2}-\frac{m^*}{2\epsilon}} \frac{\Gamma(d/2+1/2)}{\sqrt{\pi}\Gamma(d/2)}(1-x^2)^{d/2-1}dx \end{aligned}$$

## 2. Details on Methods

For detailed explanation, we show the Training pipeline and Inference pipeline of our PADA in Algorithm 1 and 2, respectively.

**Algorithm 1** Training stage of PADA: given a pre-trained conditional noise prediction network $\epsilon(x_t, t, z)$, a pre-trained semantic encoder $E_{sem}$, and a pre-trained CLIP image/text encoder $E_{img}/E_{txt}$

**Input:** source image $x_0^{src}$, reference image $x_0^{ref}$, reference text $t^{ref}$, diffusion step $T$
**Output:** $\theta^*$ (the parameters of CLIP-guided Age Encoder $E_{age}$)

**repeat**
  $t \sim Uniform(1, ..., T)$
  $x_t^{src} \sim \mathcal{N}(\sqrt{\overline{\alpha}_t}x_0^{src}, (1-\overline{\alpha}_t)\boldsymbol{I})$
  $x_t^{ref} \sim \mathcal{N}(\sqrt{\overline{\alpha}_t}x_0^{ref}, (1-\overline{\alpha}_t)\boldsymbol{I})$
  $z^{src}, z^{ref} \leftarrow E_{sem}(x_0^{src}), E_{sem}(x_0^{ref})$
  $\hat{x}_0^{src} \leftarrow \frac{x_t^{src}}{\sqrt{\alpha}_t} - \frac{\sqrt{1-\alpha_t}}{\sqrt{\alpha}_t}\epsilon(x_t^{src}, t, z^{src})$
  $\hat{x}_0^{ref} \leftarrow \frac{x_t^{ref}}{\sqrt{\alpha}_t} - \frac{\sqrt{1-\alpha_t}}{\sqrt{\alpha}_t}\epsilon(x_t^{ref}, t, z^{ref})$
  $r \leftarrow random(0, 1)$
  **if** $r \leq 0.5$ **then**
    $z^{age} \leftarrow E_{age}\left(E_{img}(x_0^{ref})\right)$
  **else if** $r \leq 0.8$ **then**
    $z^{age} \leftarrow E_{age}\left(E_{txt}(t^{ref})\right)$
  **else**
    $z^{age} \leftarrow E_{age}\left(E_{img}(x_0^{src})\right)$
  **end if**
  $z^{tar} \leftarrow z^{src} + z^{age}$
  $\hat{x}_0^{tar} \leftarrow \frac{x_t^{src}}{\sqrt{\alpha}_t} - \frac{\sqrt{1-\alpha_t}}{\sqrt{\alpha}_t}\epsilon(x_t^{src}, t, z^{tar})$
  Compute total loss $L(\hat{x}_0^{tar}, \hat{x}_0^{src}, \hat{x}_0^{ref}, t^{ref})$
  Take a gradient step on $\nabla_\theta L$
**until** coveraged

**Algorithm 2** Inference stage of PADA: given a pre-trained conditional noise prediction network $\epsilon(x_t, t, z)$, a semantic encoder $E_{sem}$, a pre-trained CLIP image/text encoder $E_{img}/E_{txt}$, and lerned CLIP-guided age encoder $E_{age}$

**Input:** source image $x_0^{src}$, reference image $x^{ref}$ or reference text $t^{ref}$, generation step $T$
$x_T^{tar} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$
$z^{src} \leftarrow E_{sem}(x_0^{src})$
**if** *image − guided* **then**
  $z^{age} \leftarrow E_{age}\left(E_{img}(x_0^{ref})\right)$
**else if** *text − guided* **then**
  $z^{age} \leftarrow E_{age}\left(E_{txt}(t^{ref})\right)$
**end if**
$z^{tar} \leftarrow z^{src} + z^{age}$
**for** $t = T, ..., 1$ **do**
  $\hat{x}_0^{tar} \leftarrow \frac{x_t^{tar}}{\sqrt{\alpha}_t} - \frac{\sqrt{1-\alpha_t}}{\sqrt{\alpha}_t}\epsilon(x_t^{tar}, t, z^{tar})$
  $x_{t-1}^{tar} \leftarrow \sqrt{\alpha_{t-1}}\hat{x}_0^{tar} + \sqrt{1-\alpha_{t-1}} \cdot \epsilon(x_t^{tar}, t, z^{tar})$
**end for**
**Output:** target aging result $x_0^{tar}$.

## 3. Qualitative and Quantitative Comparisons

We compare the continuous face aging capabilities of our PADA with DLFS [5], SAM [1], and CUSP [4] on CelebA-HQ test set in Fig. 1. Obviously, both the aging accuracy and age-irrelevant information preservation of our method are superior to these methods. Meanwhile, in Fig. 2 and Fig. 3, we show more comparison results with the three state-of-the-art methods on FFHQ-AT test set.

**Gender Adjustment.** As our PAE is proposed in CLIP latent space and incorporates gender information during aging training, we are able to perform gender adjustment using the formula $e^{rec} = e^{age} \pm \Delta e^{gend}$, where $\Delta e^{gend} = e^m - e^w$ and $e^m$ and $e^w$ correspond to the embeddings of 'man's face' and 'woman's face', respectively. Fig. 4 displays the results obtained after applying gender adjustment. More results can be found in Fig. 5 and Fig. 6.

Table 1. Quantitative analysis of diversity boundaries.

| Variance | +Low-level | +Low-level+High-level ($\epsilon$) | | | |
|---|---|---|---|---|---|
| | | 0.01 | 0.1 | 0.25 | 0.5 |
| LPIPS (↑) | 0.189 | 0.193 | 0.194 | 0.199 | 0.203 |
| ID (↑) | 0.668 | 0.649 | 0.633 | 0.617 | 0.593 |


Figure 1. Continuous face aging by interpolation in latent space. Best viewed zoomed-in.

**Diversity Boundary.** Following PICNet[6], we evaluate our diversity with LPIPS. The average score is calculated between 1k pairs generated with and without variations. In Table 1, as the sampling intensity $\epsilon$ of high-level variations increases, the diversity score increases, while the ID score slightly decreases. These indicate the promising performance of our PADA for generating diverse results while preserving identity.

**Disentanglement in PADA.** As shown in Fig 7, the early denoising steps (T=25 to T=10) prioritize shape, while the later steps (T=10 to T=0) prioritize texture. For example, if we replace C1 with C2 in later denoising steps, the generated texture corresponds to C2, while the generated shape corresponds to C1. This verifies the effectiveness of PADA for face aging.

**Comparison with StyleAging [3] on Morph and CACD2000.** We compare PADA with StyleAging [3]. Since there is the domain bias between Morph and FFHQ, so we first finetune the pretrained DiffAE on Morph dataset with 2 epochs. Compared with StyleAging [3], our method achieves better generation quality and aging fidelity. The results are shown in Fig 8.

**Effectiveness of CLIP Space.** To validate the effectiveness of CLIP feature space, we replace the CLIP image encoder with a pre-trained age estimator and adopt PAE in its latent space(called PADA_AGE). As shown in Fig. 9, it can generate diverse aging results, indicating the effectiveness of our PAE. However, PADA_AGE has limited flexibility, as it cannot directly

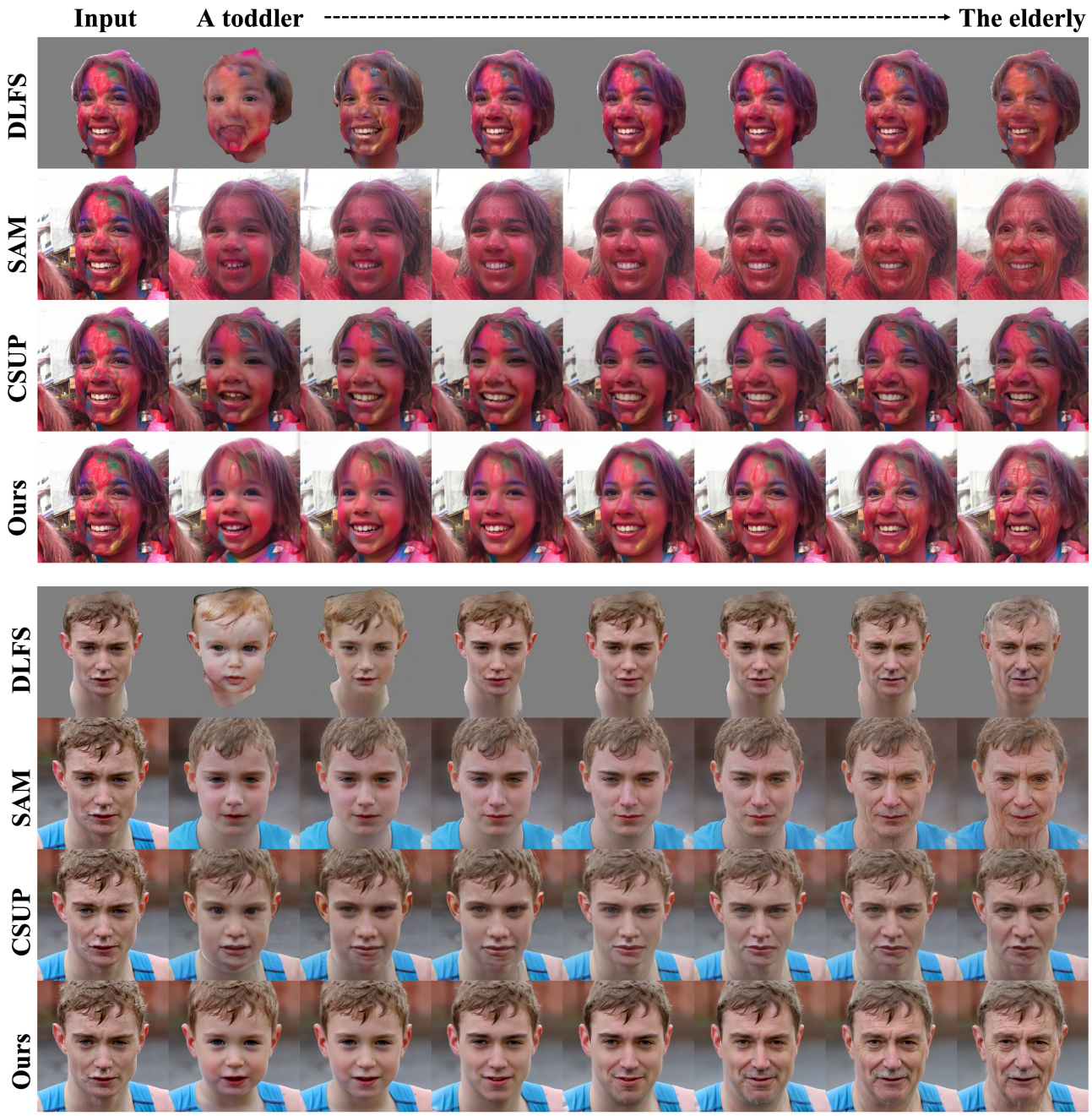


Figure 2. More comparison results with DLFS [5], SAM [1], and CUSP [4] on FFHQ-AT test set.

generate images conditioned on exact age. Additionally, its generalization ability is limited, as it fails at face aging conditioned on reference images in the wild.

## 4. Pluralistic Face Aging

We also show more reference-guided face aging results in Fig. 5 and Fig. 6. Amazingly, our PADA can generate acne marks, which cannot be achieved by current face aging methods. The text-guided face aging results are shown in Fig. 10. More results based on the open-world age descriptions or arbitrary unseen facial images are shown in Fig. 11. Although our PADA has not seen both these two variants during training, it still can generate plausible face aging results. We show more diverse face aging results with high-level varia-

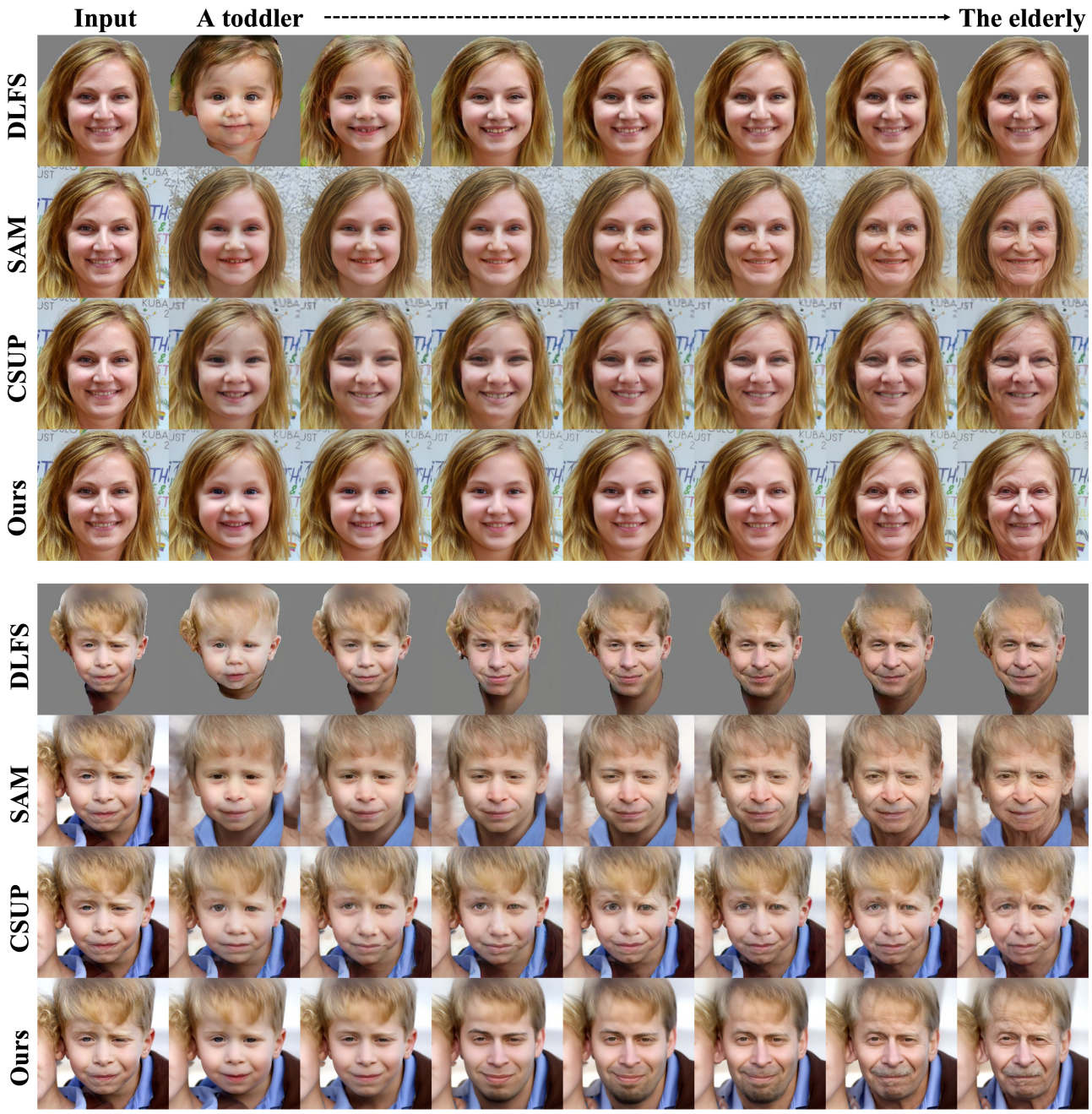


Figure 3. More comparison results with DLFS [5], SAM [1], and CUSP [4] on FFHQ-AT test set.

tions in Fig. 12 and Fig. 13. The intermediate generation results of diffusion decoder are shown in Fig. 14.

## References


[1] Yuval Alaluf, Or Patashnik, and Daniel Cohen-Or. Only a matter of style: Age transformation using a style-based regression model. *TOG*, 40(4):1–12, 2021. 2, 4, 5

[2] Eungchun Cho. Inner product of random vectors. *IJPAM*, 56(2):217–221, 2009. 1

[3] Markos Georgopoulos, James Oldfield, Mihalis A Nicolaou, Yannis Panagakis, and Maja Pantic. Enhancing facial data diversity with style-based face aging. In *CVPRW*, 2020. 3

[4] Guillermo Gomez-Trenado, Stéphane Lathuilière, Pablo Mesejo, and Óscar Cordón. Custom structure preservation in face aging. In *ECCV*. Springer, 2022. 2, 4, 5


Figure 4. Results with/without gender adjustment.

[5] Sen He, Wentong Liao, Michael Ying Yang, Yi-Zhe Song, Bodo Rosenhahn, and Tao Xiang. Disentangled lifespan face synthesis. In *ICCV*, pages 3877–3886, 2021. 2, 4, 5

[6] Chuanxia Zheng, Tat-Jen Cham, and Jianfei Cai. Pluralistic image completion. In *CVPR*, 2019. 3


Figure 5. Reference-guided aging results on FFHQ-AT test set.


Figure 6. Reference-guided aging results on FFHQ-AT test set.

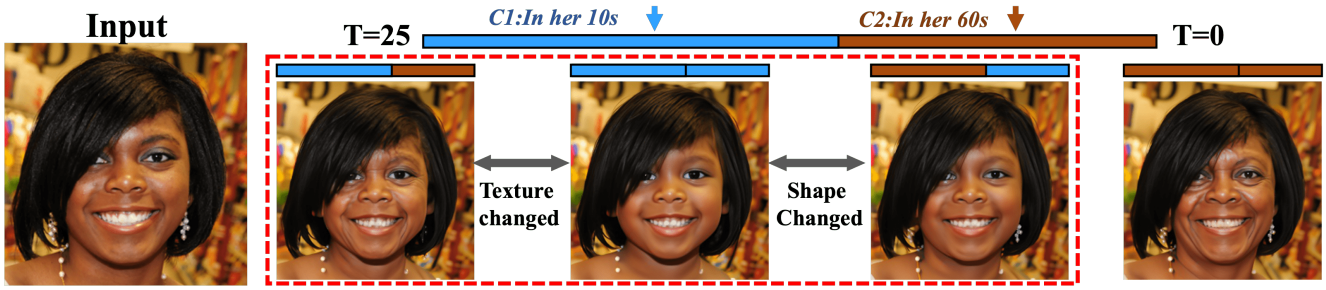


Figure 7. Manipulation at different time ($T$).

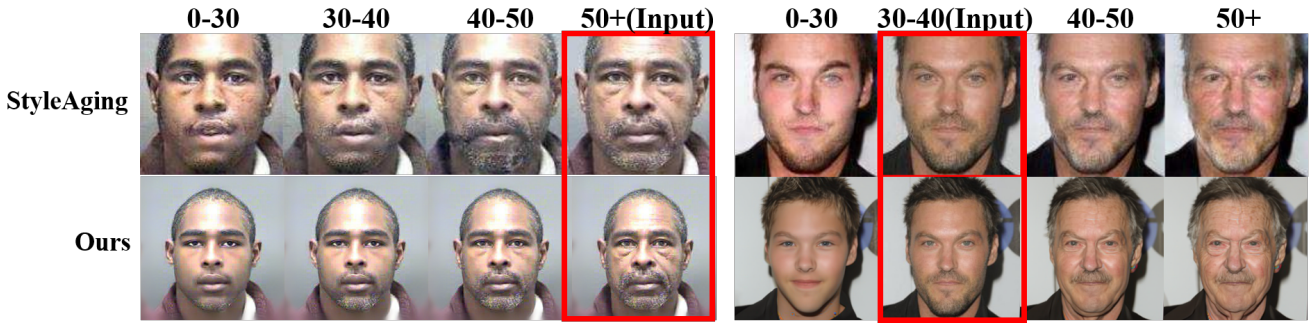


Figure 8. Comparisons on Morph(left) and CACD2000(right).

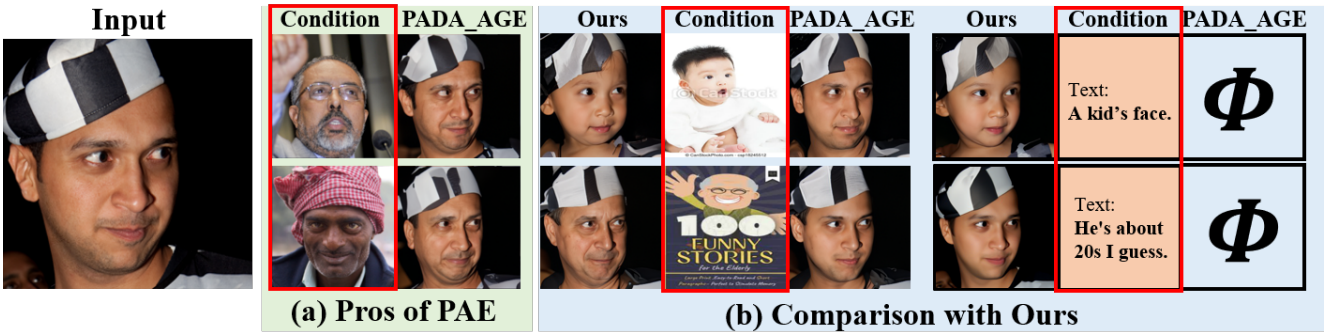


Figure 9. Compared with other feature space.

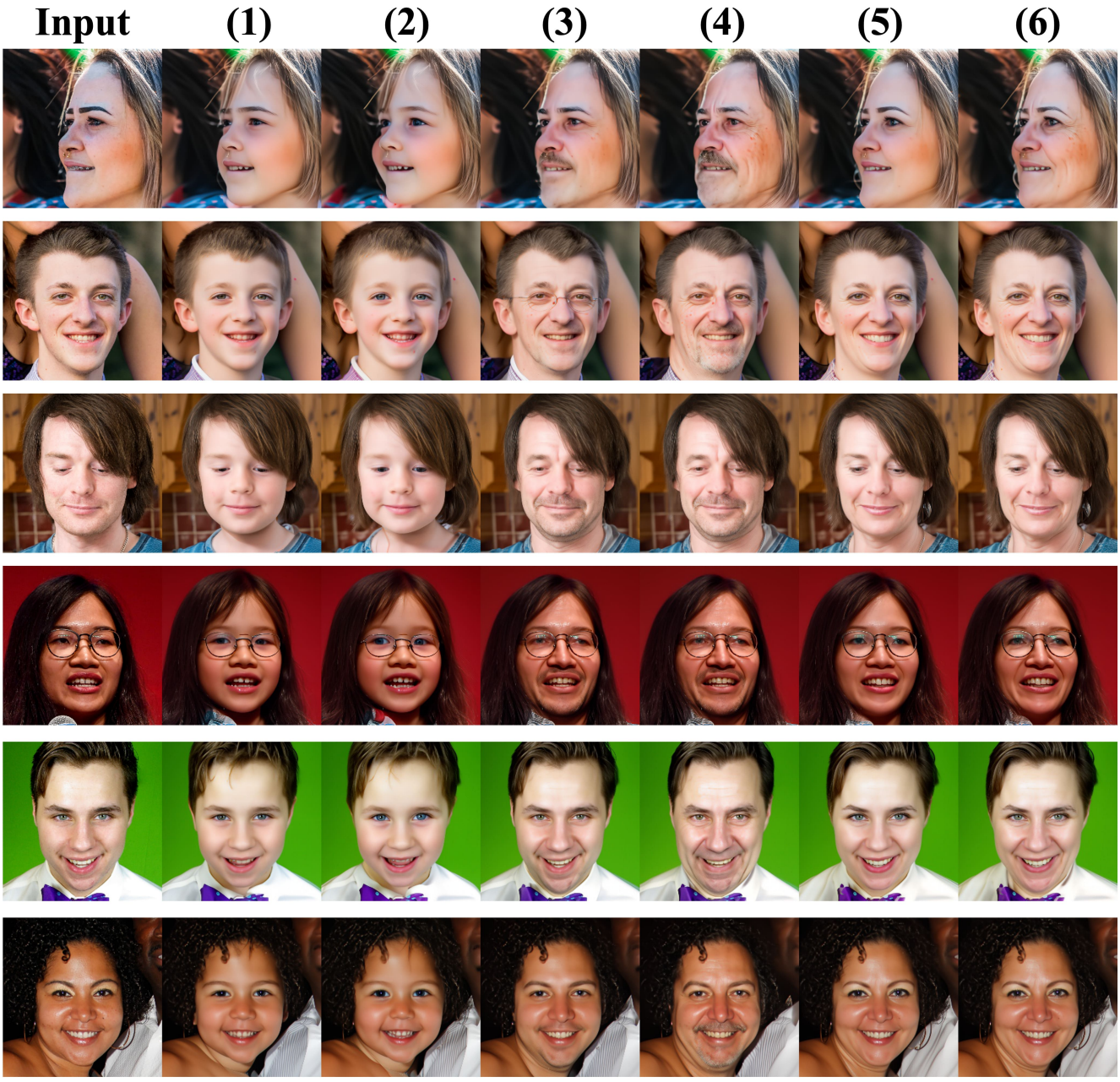


Figure 10. Text-guided aging results on FFHQ-AT test set. We apply different unseen age-related text descriptions as conditions. Concretely, (1) "a quite young boy", (2) "a daughter aged five", (3) "a face in his early forties ", (4) "a face in his late forties", (5) "a face in her early forties", (6) "a face in her late forties".

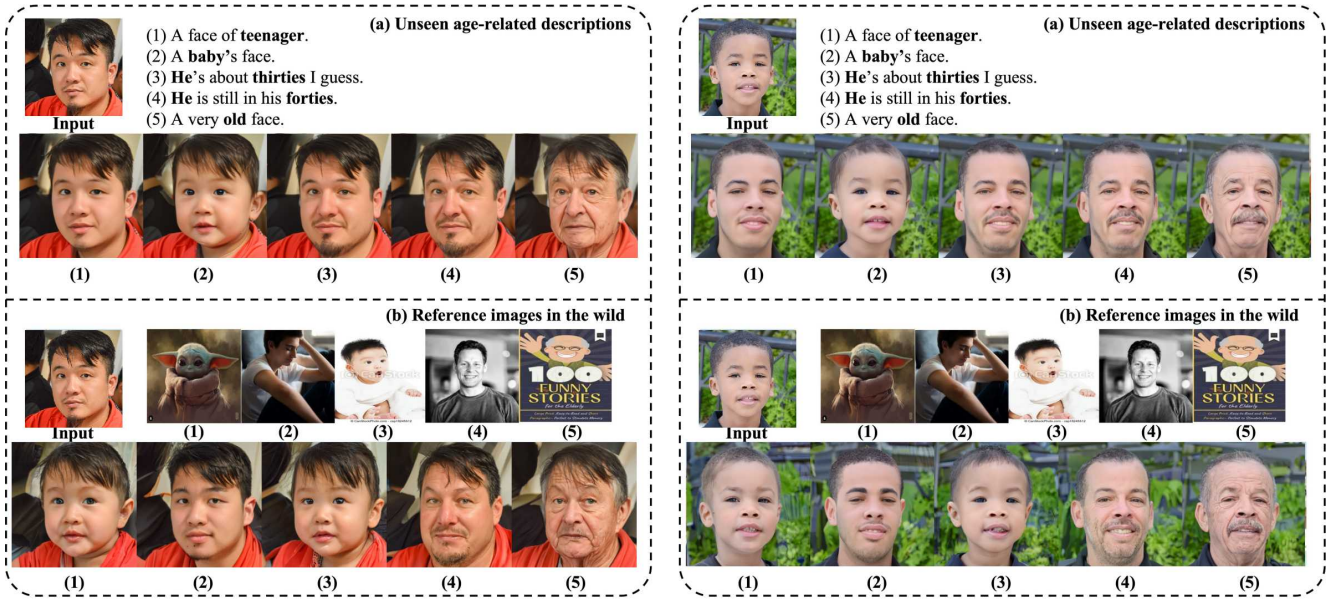


Figure 11. Face aging conditioned on unseen age-related descriptions and reference images in the wild. (a) Despite never being trained with texts of 'a very old face', our PADA still yields plausible face aging results. (b) We can utilize arbitrary reference images to guide the aging process.


Figure 12. Pluralistic aging results with high-level variations on FFHQ-AT test set.


Figure 13. Pluralistic aging results with high-level variations on FFHQ-AT test set.

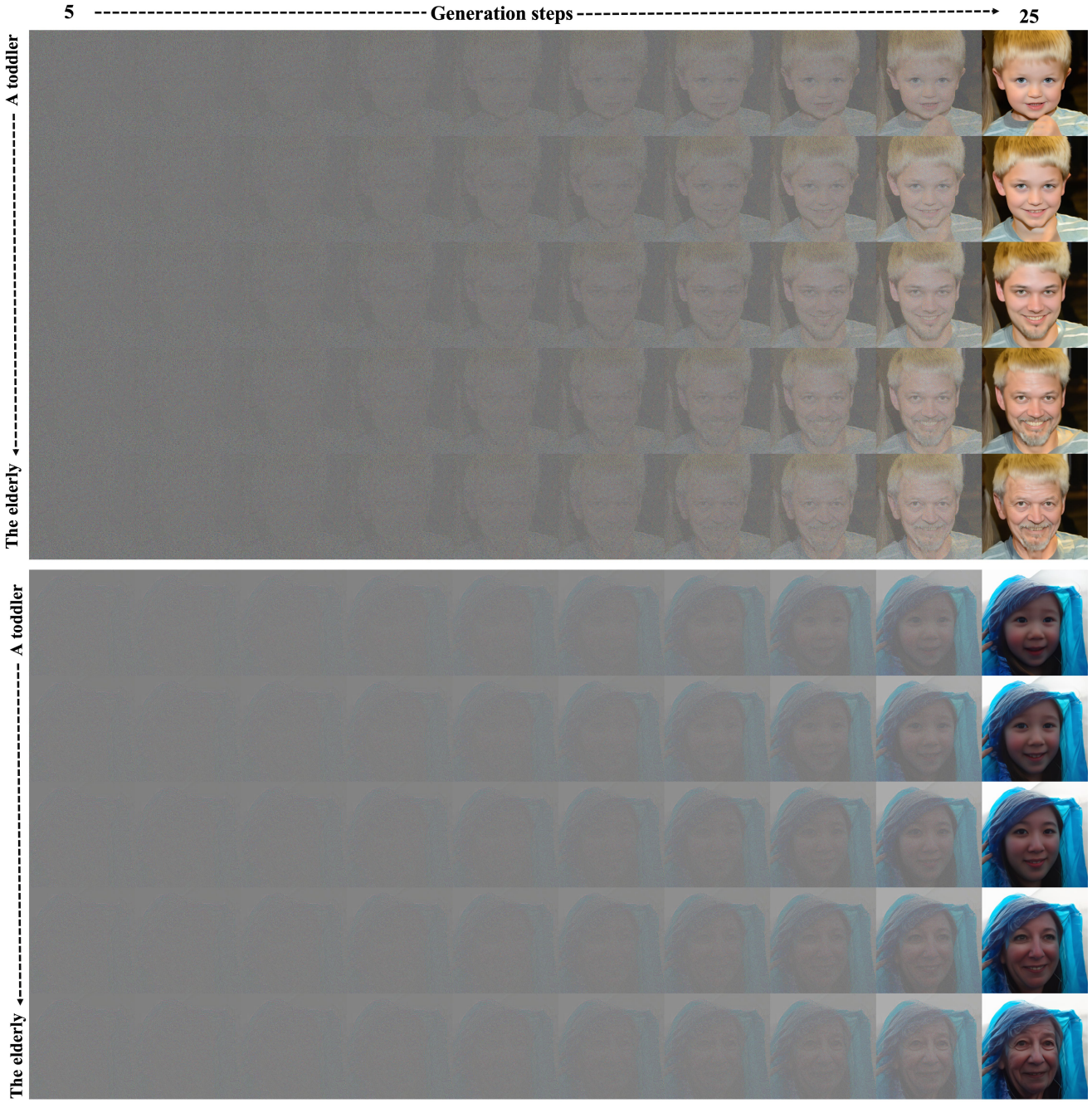


Figure 14. The intermediate generation results of diffusion decoder.