👶🔁🧓 Face Time Traveller : Travel Through Ages Without Losing Identity

Supplementary Material

Purbayan Kar[1] Ayush Ghadiya[1] Vishal Chudasama[1] Pankaj Wasnik[1] C.V. Jawahar[2]
[1]Sony Research India [2]IIIT Hyderabad
{purbayan.kar, ayush.ghadiya, vishal.chudasama1, pankaj.wasnik}@sony.com, jawahar@iiit.ac.in

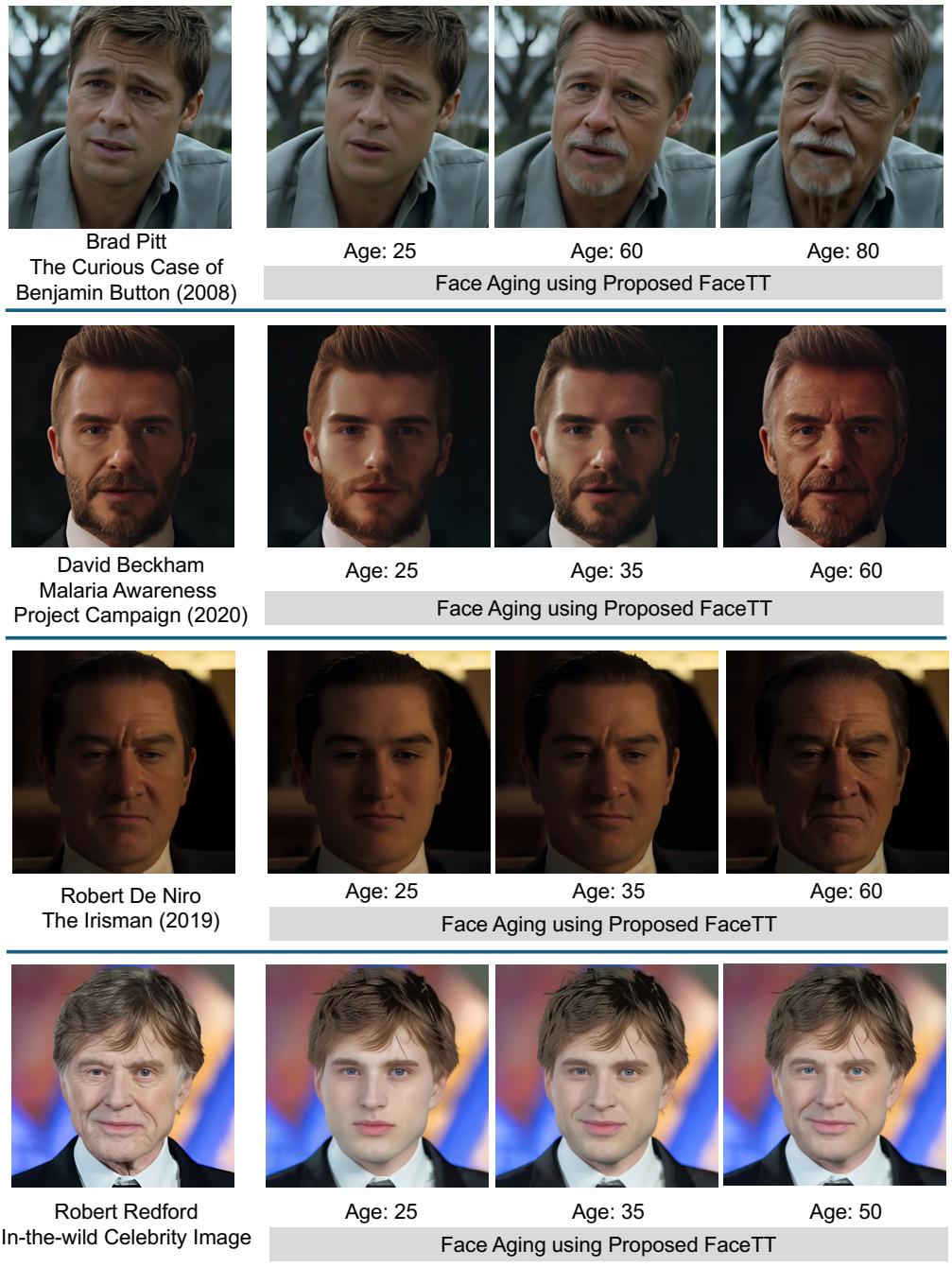

Figure S1. Results of our proposed method FaceTT in Film/Media application. FaceTT produces realistic and identity-preserving age progression on movie and in-the-wild celebrity images. These results demonstrate visually coherent transitions across young, middle-aged, and elderly appearances.

This supplementary presents below contents which we could not include in the main paper due to space constraints:

Table of Contents

S1. Analysis on Cross- and Self-Attention

We investigate how cross-attention and self-attention in Stable Diffusion influence text-guided image editing through the U-Net model's core components. Following the standard from [8], spatial features are linearly projected to form queries (Q), while text features are transformed into keys (K) and values (V) for the cross-attention module. In the self-attention module, K and V are derived from spatial features. The attention mechanism is defined as: $\text{Attention}(K, Q, V) = MV = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V$, where $M_{i,j}$ represents the attention weights for aggregating the j-th token's value at pixel i, with d being the dimensionality of K and Q.

Semantic modifications from natural language are categorized as rigid or non-rigid. Rigid changes, like background or visual element alterations, are handled using cross-attention [8], while non-rigid changes, such as adding/removing objects or changing actions, are managed with self-attention [11]. To evaluate whether the cross- and self-attention maps in face images contain meaningful semantic information, we employ a probing approach inspired by natural language processing (NLP) techniques [4, 12], as demonstrated in [11].

S1.1. Probing Analysis:

Following [11], we design and train a task-specific classifier consisting of a two-layer MLP to predict semantic categories from the attention maps. Using the prompt `Photo of a <2/10/30/50/70> years old <man/woman/boy/girl>`, we examine the effectiveness of the attention mechanisms. In this probing analysis, we avoid using our Face-Attribute-Aware refined prompts because they introduce additional semantic cues (e.g., skin texture, lifestyle conditions) that go beyond age and gender. Including these details would make the probe rely on prompt semantics rather than the intrinsic information encoded in the attention maps. Using simple age-and-gender prompts ensures a controlled setting where we can directly evaluate what the model's native attention layers capture without external influence.

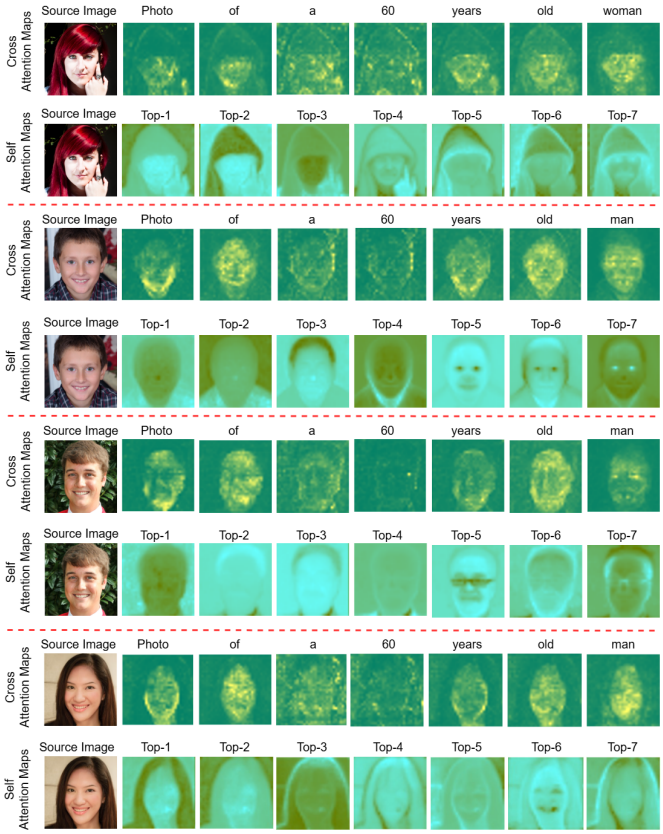

Figure S2. Visualization of Cross- & Self-attention maps for age progression to 60-year-old appearance. Cross-attention maps highlight regions in the source image (e.g., eyes, nose, mouth) that align with target prompt `<Photo of a 60 years old woman/man>`. While self-attention visualization showcases top-7 components derived from singular value decomposition (SVD) [17], which is important to maintain identity and consistency.

Table S1. Probing accuracy of cross-attention maps w.r.t. different tokens. L denotes layers of the U-Net model.

Class	L-3	L-6	L-9	L-10	L-12	L-14	L-16	Avg.
Age 2	0.29	0.42	0.69	0.81	0.83	0.85	0.75	0.77
Age 10	0.60	0.20	0.70	0.92	0.90	0.90	0.63	0.69
Age 30	0.57	0.92	0.80	0.88	0.89	0.89	0.78	0.79
Age 50	0.50	0.87	0.80	0.72	0.87	0.87	0.41	0.81
Age 70	0.54	0.41	0.82	0.71	0.63	0.55	0.52	0.59
man	0.67	0.77	0.86	0.83	0.81	0.81	0.90	0.80
woman	0.62	0.80	0.83	0.78	0.76	0.82	0.84	0.78
boy	0.60	0.80	0.90	0.92	0.96	0.75	1.0	0.85
girl	0.50	0.70	0.92	0.84	0.87	0.70	0.90	0.78

Cross-Attention Maps: We explore the information captured by cross-attention maps by visualizing the attention patterns associated with each word in a prompt, as dis-

Table S2. Probing analysis of cross-attention maps w.r.t. difference tokens. The upper part shows the classification accuracy corresponding to the token `Photo`, and the lower shows results for `old`. L denotes layers of the U-Net model.

Class	L-3	L-6	L-9	L-10	L-12	L-14	L-16	Avg.
Age 2	0.85	0.78	0.89	0.80	0.93	0.85	0.91	0.87
Age 10	0.80	0.71	0.80	0.93	0.50	0.70	0.50	0.71
Age 30	0.76	0.85	0.86	0.90	0.75	0.96	0.71	0.83
Age 50	0.71	0.73	0.70	0.83	0.70	0.55	0.66	0.70
Age 70	0.61	0.67	0.62	0.70	0.66	0.50	0.60	0.62
Age 2	0.90	0.83	0.80	0.92	0.84	0.92	0.93	0.88
Age 10	0.85	0.90	0.87	0.80	1.0	0.75	0.70	0.84
Age 30	0.80	0.82	0.79	1.00	0.70	0.69	0.75	0.79
Age 50	0.83	0.80	0.70	0.79	0.63	0.62	0.61	0.71
Age 70	0.76	0.71	0.59	0.83	0.52	0.70	0.58	0.67

Table S3. Probing accuracy of self-attention maps w.r.t. different tokens. L denotes layers of the U-Net model.

Class	L-3	L-6	L-9	L-10	L-12	L-14	L-16	Avg.
Age 2	0.00	0.00	0.00	0.20	0.00	0.00	0.00	0.028
Age 10	0.18	0.06	0.70	0.35	0.80	0.25	0.90	0.46
Age 30	0.00	0.00	0.00	0.00	0.5	1.0	0.00	0.21
Age 50	0.00	0.1	0.00	0.00	0.66	0.23	0.10	0.16
Age 70	0.00	0.00	0.82	0.05	0.25	0.00	0.23	0.19
man	0.36	0.68	0.45	0.90	0.80	0.30	0.60	0.58
woman	0.33	0.41	0.83	0.58	0.63	0.10	0.81	0.53
boy	0.30	0.21	0.80	0.87	0.91	0.65	0.80	0.76
girl	0.47	0.10	0.00	0.69	0.70	0.41	0.00	0.34

played in Figure S2. Each map highlights areas related to specific words, indicating that cross-attention retains rigid semantic details. In the context of the facial re-aging task, these rigid details include wrinkles and changes in skin texture, variations in the fullness of the cheeks and lips, as well as the development of age-related marks such as fine lines or sagging skin. Table S1 presents the classification performance of our trained classifier, demonstrating high accuracy, especially for the `age` and `person` categories.

We also investigate whether cross-attention maps associated with non-edited words contain rigid information. This analysis is essential because text embeddings generated by transformer-based encoders [10, 15] preserve the contextual details of sentences. By using prompts like `Photo of a <age> years old <person>`, our experiments (shown in Table S2) reveal that both the tokens `Photo` and `old` exhibit significantly high classification accuracy. This finding indicates that both tokens carry rigid semantic information related to faces.

Self-Attention Maps: In contrast, self-attention maps exhibit different behaviors and are analyzed similarly. Table S3 shows that the classifier struggles to recognize categories (`age` and `person`) accurately, indicating that self-attention maps encode non-rigid information, or structural rather than categorical details. For our task, this structural information includes the relative positions of the eyes, nose, mouth, and the overall shape of the face (e.g., jawline). Figure S2 illustrates that self-attention maps preserve the original structure of the image.

This show that cross-attention maps encode semantic features for category identification, while self-attention maps preserve structural integrity for image coherence.

S2. Additional Visual Comparisons

This section offers a visual comparison between our FaceTT and the state-of-the-art (SOTA) methods i.e., HRFAE [18], CUSP [6], and FADING [3]. From Figure S3 to S5 , one can observe that HRFAE preserves identity but struggles with exaggerated features. CUSP provides smooth transitions; however, it comes at the cost of losing fine details in extreme aging. FADING generates plausible results but often produces overly smooth textures or inconsistencies. In contrast, our method maintains identity and delivers highly realistic aging effects across diverse facial structures and expressions.

S3. Analysis on Identity Preservation

S3.1. Cyclic / Reference based Aging Analysis:

We conducted Cyclic and Reference-based Identity Similarity evaluation on a five-celebrity in-the-wild age-progression test set. Results for one celebrity (Tom Cruise) are shown in the main paper (Figure 2), while the remaining four—Brad Pitt, Leonardo DiCaprio, Matt Damon, and Robert Downey Jr.—are presented in Figure S6. Across all age cycles, our method consistently yields higher identity similarity than HRFAE, CUSP, and FADING, demonstrating stronger face aging with identity preservation in both forward and reverse age transformations.

S3.2. Short-Range Aging Analysis

To further assess the identity-preservation capability of our method, we perform a Short-Range Aging Analysis that evaluates identity stability under fine-grained age perturbations. Figure 9 in the main paper shows examples for two identities, demonstrating consistent identity retention across closely spaced target ages. Additional results for two more identities are provided in Figure S7, further highlighting the robustness of our approach in maintaining facial structure and appearance under small age variations.

S4. Additional Ablation Analysis

S4.1. Visual Analysis on Hyper-Parameters

Figure S8 analyzes the impact of the hyper-parameter ξ used in *Angular Inversion*. Low ξ values (specifically, $\xi = 0.1, 0.5$) result in minimal changes, allowing the input identity to be preserved but producing weaker re-aging effects. Conversely, high ξ values (such as $\xi = 2.0$) lead to

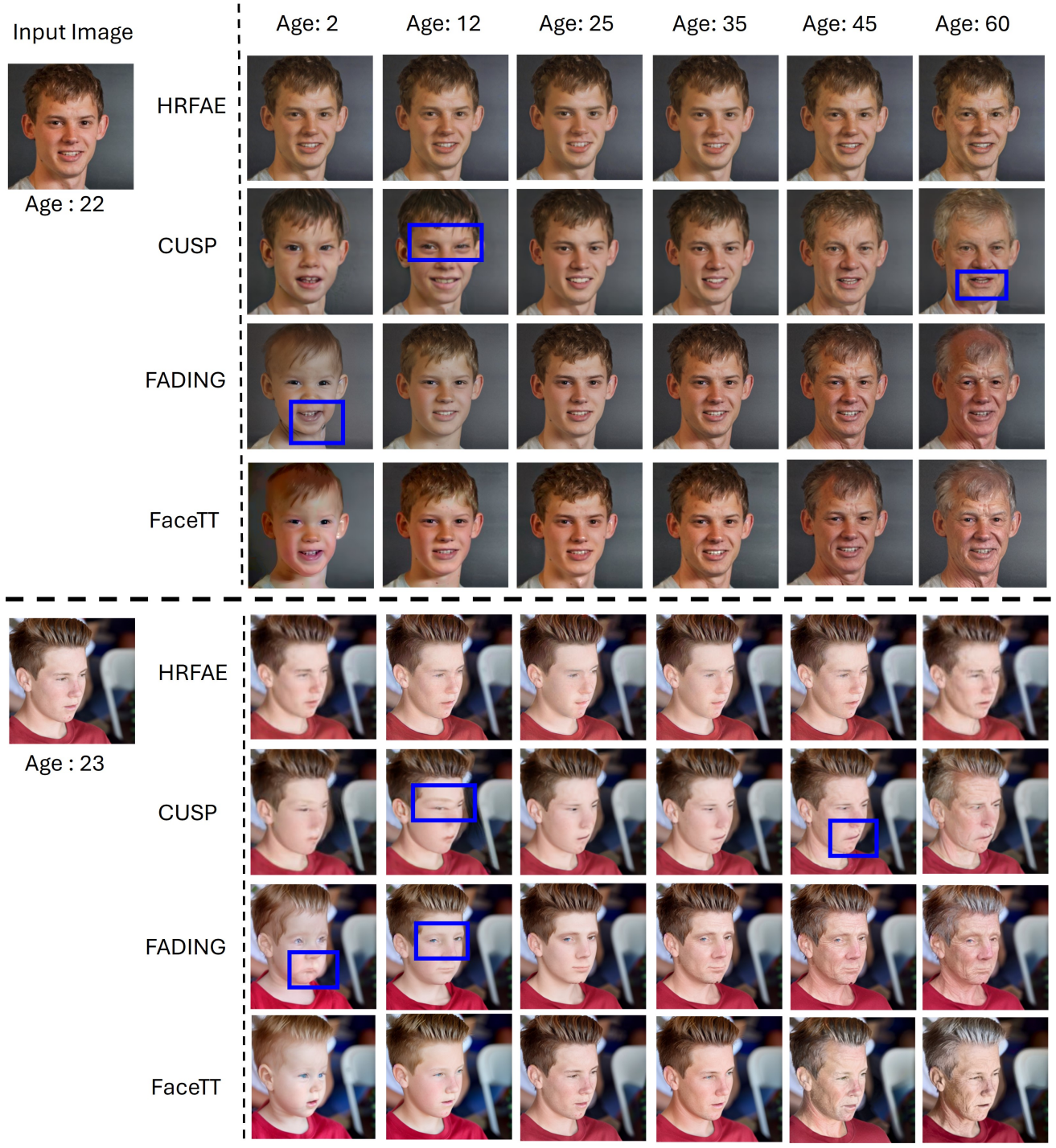

Figure S3. Comparison of facial re-aging across different age targets using HRFAE, CUSP, FADING, and our FaceTT method (Part 1). Across both examples, FaceTT produces the most consistent identity preservation and the most realistic aging trajectories from early childhood (Age: 2) to older adulthood (Age: 60). Competing methods exhibit notable artifacts: HRFAE often fails to introduce meaningful age changes and tends to output visually similar faces across age ranges; CUSP frequently produces structural distortions and texture inconsistencies (blue boxes), especially in the eyes and mouth regions; and FADING, while capable of stronger aging effects, suffers from identity drift and unnatural facial modifications (blue boxes). In contrast, FaceTT generates smooth, coherent age transitions with preserved facial geometry and identity cues, while accurately reflecting age-specific features such as fuller cheeks in youth, subtle mid-life changes, and realistic wrinkle patterns at older ages.

Figure S4. Additional visual comparison of facial re-aging methods. (Part 2, continued from Figure S3)

excessive alterations, causing distortions and a loss of identity. The optimal balance is found at $\xi = 1.2$, which provides realistic re-aging while maintaining identity, as shown in the first and third rows. This highlights $\xi = 1.2$ as the preferred choice for high-quality results.

Figure S9 examines the effect of η_{th} in *Adaptive Attention Control*, which determines the transition between self-attention and cross-attention adaptation. When η_{th} is too low (e.g., 0.03), the attention mechanism primarily preserves source characteristics, leading to insufficient aging

Figure S5. Additional visual comparison of facial re-aging methods. (Part 3, continued from Figure S3)

effects. Conversely, setting η_{th} too high (e.g., 0.06) results in excessive modification of details, introducing artifacts. The best balance occurs at $\eta_{th} = 0.05$, enabling natural aging transformations while maintaining identity consistency.

Figure S10 evaluates the impact of (τ_1, τ_2), which regulates the selection of either cross-attention or self-attention or mixture of both (based on the value of η_{th}) in *Adaptive Attention Control*. Low τ_1 values (e.g., 25) retain too many source attributes, preventing effective aging transformations. On the other hand, high values (e.g., 40) lead to

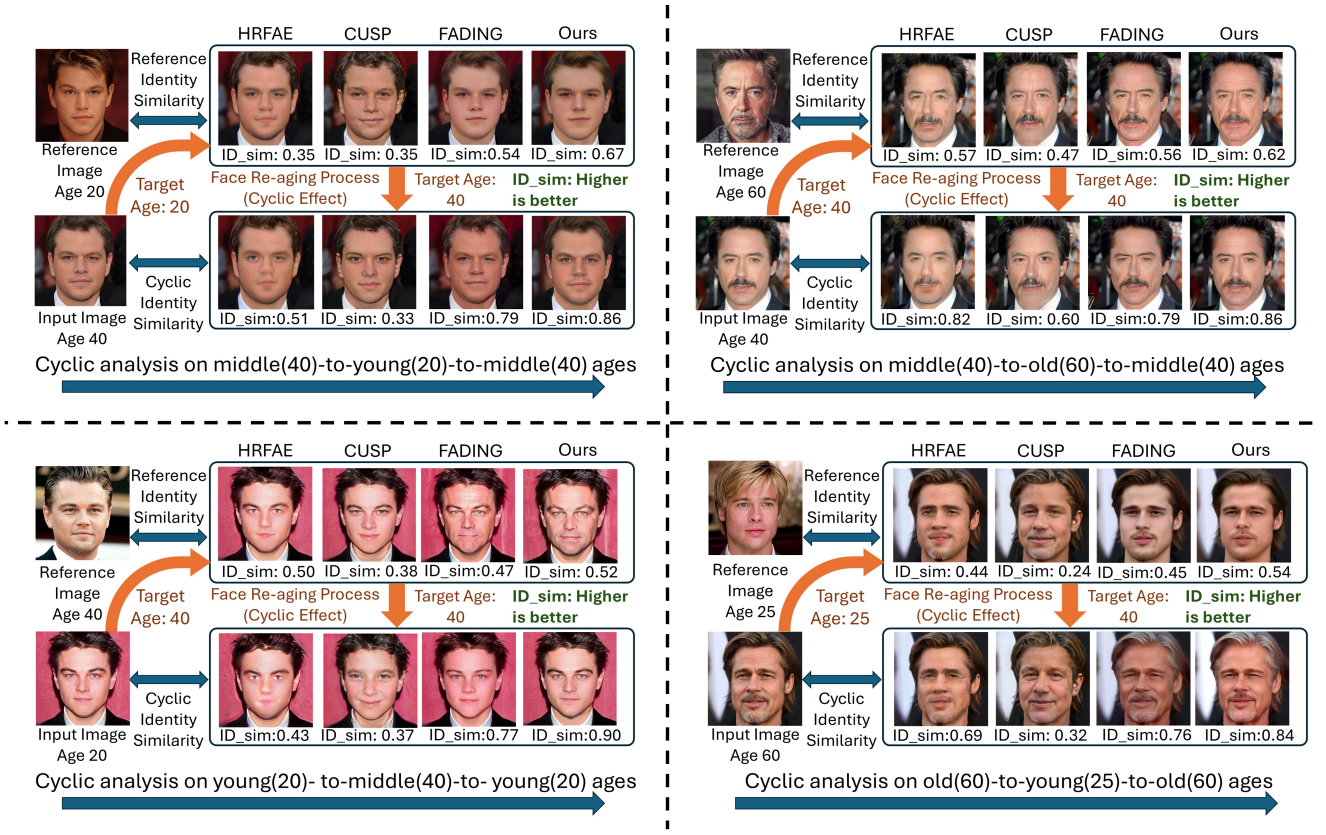

Figure S6. Cyclic Identity Similarity protocol for evaluating identity preservation across age transitions. For each input celebrity image, we perform age progression and regression cycles (e.g., 40→20→40 (top left), 40→60→40 (top right), 20→40→20 (bottom left), 60→25→60 (bottom right)) and compute both Reference Identity Similarity (between the re-aged output and a real image at the target age) and Cyclic Identity Similarity (between the input and its cyclic reconstruction). Across all age cycles, our method consistently achieves higher identity similarity compared to HRFAE, CUSP, and FADING, demonstrating more faithful identity retention during both forward and reverse age transformations.

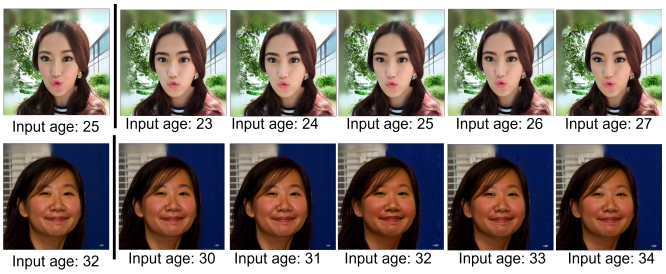

Figure S7. Short-range aging analysis for identity preservation. Each row shows aging results within a narrow age window (±2 years) around the input age. Our method preserves identity exceptionally well across all nearby target ages, producing nearly indistinguishable facial structure, expression, and texture from the original input. This demonstrates that even under fine-grained age perturbations—where inconsistencies are easiest to detect—FaceTT maintains stable and coherent identity features without drifting or introducing artifacts.

exaggerated changes, distorting facial features. Similarly, improper selection of τ_2 affects the smoothness of the transformation. The optimal configuration, (τ_1, τ_2) = (35,15), provides the best trade-off between structural consistency and realistic aging effects, as demonstrated in Figure S10.

S4.2. Statistical Analysis of Hyper-Parameter

To further validate our approach, we incorporated a statistical analysis in Fig. S11, where the mean ± standard deviation of the predicted ages is reported for different age groups across various hyperparameter configurations. This evaluation was conducted on the CelebA-HQ dataset to assess the consistency and accuracy of the generated age transformations. The statistical trends observed support the robustness of our chosen hyperparameters, ensuring realistic and demographically appropriate age progression across the dataset.

Biometric Matching Analysis: Inspired by [1], we perform a biometric verification study on the FFHQ dataset.

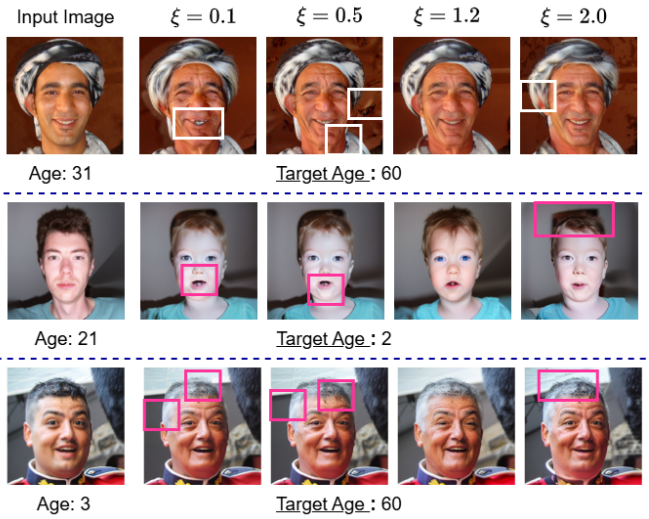

Figure S8. Ablation analysis on different ξ values being used in the proposed *Angular Inversion* technique. The results demonstrate how different ξ values influence the balance between preserving identity and achieving accurate transformations to match the target prompts (e.g., age 2 or 60 years). Among the tested values, ξ = 1.2 provides the best results, striking an optimal trade-off between transformation fidelity and identity preservation.

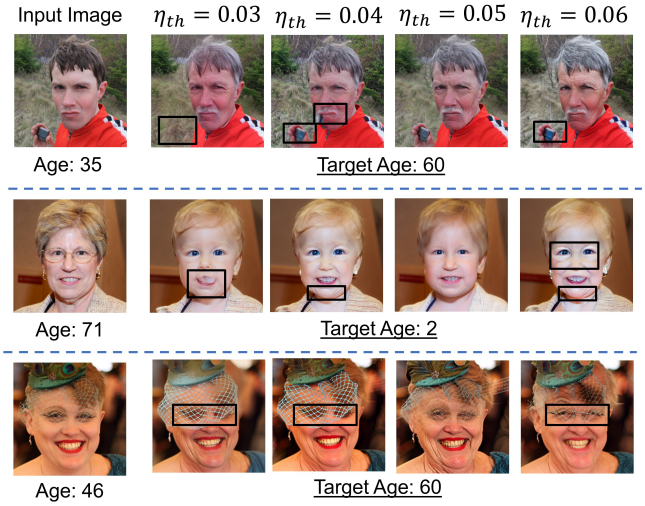

Figure S9. Ablation analysis on different η_{th} values in the proposed *Adaptive Attention Control* technique. The results illustrate how varying η_{th} impacts the trade-off between maintaining structural consistency and achieving accurate aging transformations. Lower values (e.g., η_{th} = 0.03) retain excessive source attributes, leading to incomplete aging effects, while higher values (e.g., η_{th} = 0.06) introduce artifacts and distortions in fine details (highlighted in black boxes). The optimal value, η_{th} = 0.05, provides the best balance, ensuring realistic aging transformations while preserving important identity and contextual details.

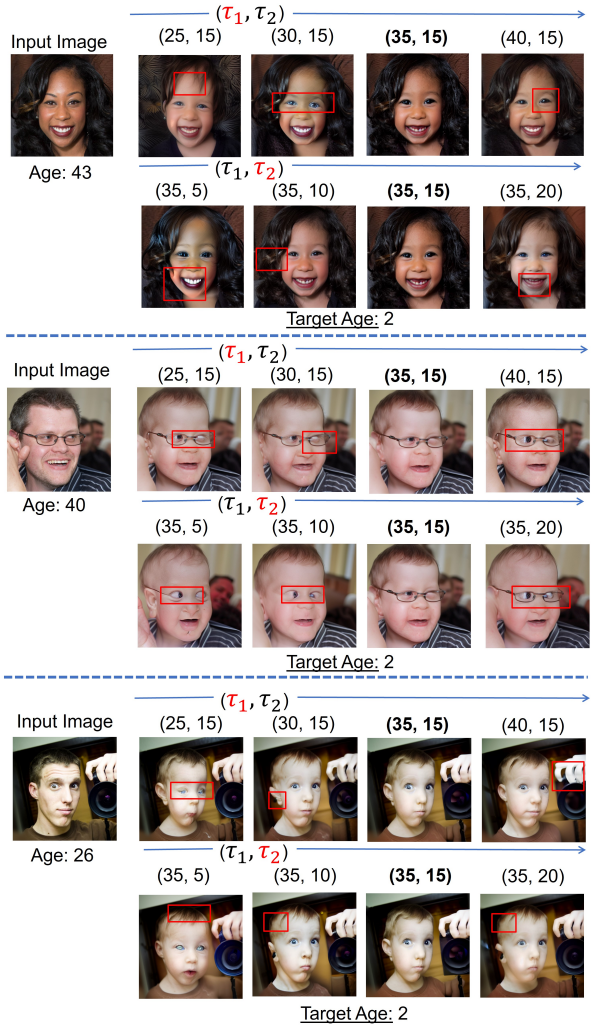

Figure S10. Ablation analysis on different (τ_1, τ_2) values in the *Adaptive Attention Control* technique, investigating their effect on balancing structural preservation and non-rigid transformations. Lower values of τ_1 (e.g., 25) retain more source characteristics but limit the effectiveness of the transformation, while higher values (e.g., 40) can lead to excessive modifications, causing distortions in fine details (highlighted in red boxes). Similarly, an improper choice of τ_2 can result in either overly rigid transformations or undesirable warping effects. The best balance is achieved with (τ_1, τ_2) = (35,15), ensuring natural age transformation while preserving key identity features.

Table S4. Biometric matching results between original and re-aged images on FFHQ dataset. The metrics are False Non Match Rate (FNMR) at False Match Rate (FMR) = 0.01/0.1%. Lower is better.

Method	2	12	25	35	45	60
CUSP	0.52/0.28	0.22/0.11	0.31/0.13	0.27/0.09	0.23/0.07	0.45/0.10
AgeTransGAN	0.73/0.51	0.52/0.25	0.52/0.38	0.47/0.41	0.36/0.30	0.38/0.24
ADFD	0.97/0.91	0.79/0.69	0.77/0.68	0.68/0.63	0.69/0.60	0.77/0.67
FADING	0.55/0.49	0.11/0.07	0.26/0.09	0.25/0.07	0.25/0.07	0.33/0.12
FaceTT	0.60/0.49	0.18/0.10	0.06/0.06	0.02/0.01	0.03/0.02	0.13/0.07

Following the standard verification protocol, we report the *False Non-Match Rate (FNMR)* at two operating points of *False Match Rate (FMR)* = 0.01 and 0.1%. This experi-

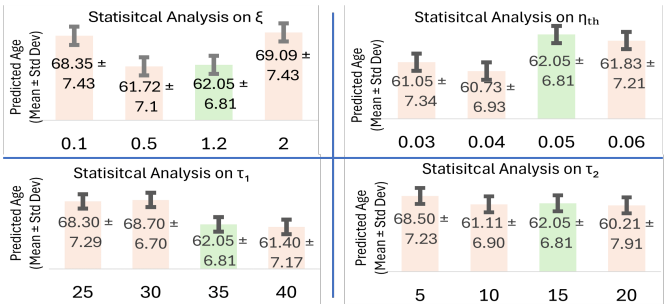

Figure S11. Statistical analysis. Targeted predicted age: 65.14±4.86

Table S5. Comparative analysis on CelebA-HQ dataset for young-to-60 task. Best score: in red, second-best score: in green.

Techniques	Predicted age	Blur ↓	Gender ↑	Smiling ↑	Neutral ↑	Happy ↑
Analysis on Inversion techniques						
Real Images	65.14 ± 4.86	1.73	—	—	—	—
Null-Text [14]	69.88 ± 6.20	2.18	98.44	76.17	56.54	73.19
Negative [13]	69.35 ± 5.51	2.69	98.23	74.35	61.77	69.19
Proximal [7]	68.39 ± 6.74	2.49	98.13	73.99	63.94	71.98
Direct [9]	61.84 ± 6.80	2.18	99.71	73.74	64.65	66.23
Angular (Proposed)	62.05 ± 6.81	2.18	99.79	78.31	62.67	72.17
Analysis on editing techniques						
Real Images	65.14 ± 4.86	1.73	—	—	—	—
P2P [8]	57.31 ± 4.23	2.31	97.40	72.07	61.19	65.77
MasaCtrl [2]	57.23 ± 7.65	2.57	96.90	75.39	64.25	66.32
PnP [16]	69.35 ± 6.26	2.30	98.67	78.22	62.15	69.85
FPE [11]	68.45 ± 6.32	2.27	97.53	73.39	62.36	67.20
AAC (Proposed)	62.05 ± 6.81	2.18	99.79	78.31	62.67	72.17

ment measures how well aged faces remain recognizable to a pretrained face recognition model [5]. Table S4 presents FNMR@FMR across six aging distances (2, 12, 25, 35, 45, and 60 years) for CUSP, FADING, and our FaceTT. Lower FNMR indicates stronger identity retention. FaceTT consistently achieves the lowest FNMR (particularly for the middle and older age groups). Notably, at a 35-year aging distance, FaceTT achieves an FNMR of 0.02/0.01, representing a significant improvement over competing methods.

The additional analysis is illustrated in Figures S12, which shows a visual comparison between the proposed *Angular Inversion* and existing inversion techniques [7, 9, 13, 14]. This comparison shows that the Proximal Inversion [7] achieves partial alignment with target prompts but struggles with fine details, resulting in inconsistent transformations. While Negative-Prompt Inversion [13] enhances alignment, it introduces distortions, particularly in extreme age changes. Null-Text inversion [14] provides better prompt alignment but compromises identity and realism, leading to overly smoothed features. Direct Inversion [9] maintains facial structure but lacks the adaptability needed for extreme transformations, which results in artifacts. In contrast, our *Angular Inversion* method captures fine details and preserves semantic alignment and identity, resulting in realistic and coherent age transformations. Its balanced performance proves its superiority over existing methods.

Figures S13 show an additional comparison of the proposed *AAC* with existing image editing techniques, including P2P [8], PnP [16], MasaCtrl [2], and FPE [11]. It shows that P2P struggles to balance alignment and identity, often resulting in exaggerated or distorted features. While PnP is more consistent, it is still prone to artifacts, especially in extreme cases. MasaCtrl improves prompt fidelity but tends to overemphasize features, resulting in unnatural transformations. FPE balances structure and identity but lacks adaptability in extreme cases, resulting in overly smooth results. In contrast, our *AAC* outperforms these methods by modulating attention mechanisms to achieve semantic fidelity and identity preservation. It delivers realistic transformations, handling challenging cases like extreme age regression or progression with natural, coherent outputs.

S5. Ethical Concerns

The development of facial re-aging technology raises several ethical considerations that must be addressed to ensure responsible usage. The ability to manipulate facial features across different age groups brings up concerns about privacy, identity misuse, and potential exploitation. For example, re-aging technology could be misused to create deceptive content, such as deepfakes, which might lead to misinformation or harm an individual's reputation. This challenge is applied to all image editing methods in general. However, advancements in detecting and mitigating malicious edits are evolving quickly. We believe our work will support these efforts by providing insights and access of the proposed image editing and generation process.

S6. Baseline Details

Our proposed framework for facial re-aging is compared with three SOTA methods [3, 6, 18]. Details of these baselines are as follows:

- **HRFAE [18]**, a hybrid model that uses feature-aligned encoders to preserve identity and achieve photorealistic re-aging effects. However, HRFAE often struggles with extreme age transformations, leading to minor inconsistencies in older age predictions.
- **CUSP [6]** combines cycle-consistent adversarial networks with spatial priors to achieve smooth age progression and regression. Despite its ability, CUSP sometimes fails to capture fine-grained aging details accurately.
- **FADING [3]** uses null-text inversion [14] and attention control for facial image editing with a pre-trained diffusion model. Despite high-quality results, it faces challenges with identity consistency in extreme age edits and is computationally intensive for real-time applications.

To validate the effectiveness of our proposed *angular inversion* technique, we compare it with existing inversion techniques [7, 9, 13, 14].

- **Null-Text Inversion [14]** uses pivot tuning with null-text prompts for better alignment but suffers from inefficien-

Figure S12. Comparison of different inversion methods for facial re-aging across multiple target ages (2 → 60 years). Proximal, Null-Text, Direct, and DDPM inversion each introduce distortions or identity drift during the re-aging process. Proximal Inversion often produces inconsistent geometry and blurry structural changes, leading to unnatural aging transitions. Null-Text Inversion exhibits significant shape warping and texture artifacts (blue boxes), particularly in early-age predictions. Direct Inversion suffers from identity inconsistency and incorrect facial proportions across ages, while DDPM Inversion frequently alters key identity features (blue boxes), causing the re-aged outputs to deviate from the input subject. In contrast, *Angular Inversion* yields stable, high-fidelity reconstructions with smooth and realistic age progression, maintaining consistent identity cues from childhood to old age. The results highlight that *Angular Inversion* achieves the most reliable and artifact-free initialization for age editing, enabling accurate transformations across the entire age spectrum.

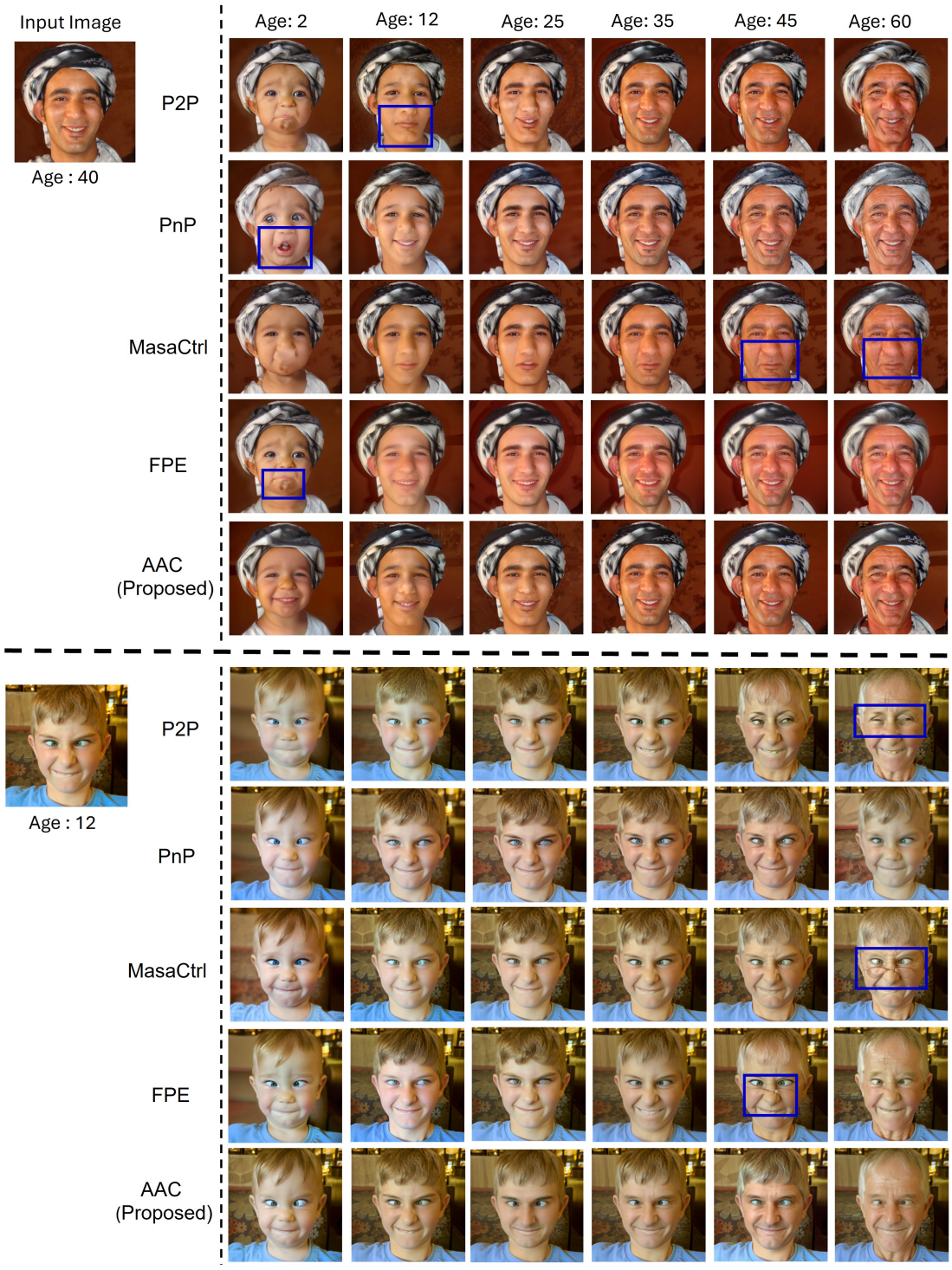

Figure S13. Comparison of age editing using P2P, PnP, MasaCtrl, FPE, and our *Adaptive Attention Control (**AAC**)* across target ages 2→60. Existing attention-based editing methods exhibit noticeable artifacts and identity inconsistencies: P2P often produces over-smoothed faces and distorted geometry (blue boxes), especially for younger ages; PnP struggles with preserving identity and frequently introduces incorrect mouth and eye structures; MasaCtrl shows unstable attention modulation, leading to drifted identity and inconsistent aging patterns; and FPE tends to hallucinate textures or distort facial regions (blue boxes), particularly in older targets. In contrast, *AAC* achieves smooth and coherent age progression while consistently preserving identity, facial structure, and attribute integrity across the entire age spectrum. The results demonstrate that Adaptive Attention Control provides more precise, stable, and semantically aligned guidance than prior attention-editing approaches.

cies in computation.

- **Negative-Prompt Inversion [13]** speeds up inversion by approximating DDIM inversion, but sacrifices reconstruction quality.
- **Proximal Inversion [7]** improves Negative-Prompt Inversion by using proximal guidance, but struggles with latent space disruptions, particularly for complex transformations.
- **Direct Inversion [9]** separates source and target branches for better content preservation but faces challenges with complex edits, especially when large attribute shifts are required.

We also validate the proposed *AAC* editing technique with existing baselines [2, 8, 11, 16].

- **Prompt-to-Prompt (P2P) [8]** adjusts cross-attention to maintain spatial consistency in edits. Although effective in many scenarios, it struggles with transformations requiring significant structural changes.
- **Plug-and-Play (PnP) [16]** modifies attention maps for text-driven image editing but introduces minor artifacts.
- **MasaCtrl [2]** enables mutual self-attention control but generates noticeable artifacts in challenging transformations.
- **FPE [11]** modifies self-attention maps for stable edits but struggles with precise age-specific effects.

References

[1] Sudipta Banerjee, Govind Mittal, Ameya Joshi, Chinmay Hegde, and Nasir Memon. Identity-preserving aging of face images via latent diffusion models. In *2023 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–10. IEEE, 2023. 7

[2] Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22560–22570, 2023. 9, 12

[3] Xiangyi Chen and Stéphane Lathuilière. Face aging via diffusion-based editing. In *BMVC*, 2023. 3, 9

[4] Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D Manning. What does bert look at? an analysis of bert's attention. *arXiv preprint arXiv:1906.04341*, 2019. 2

[5] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019. 9

[6] Guillermo Gomez-Trenado, Stéphane Lathuilière, Pablo Mesejo, and Óscar Cordón. Custom structure preservation in face aging. In *European Conference on Computer Vision*, pages 565–580. Springer, 2022. 3, 9

[7] Ligong Han, Song Wen, Qi Chen, Zhixing Zhang, Kunpeng Song, Mengwei Ren, Ruijiang Gao, Anastasis Stathopoulos, Xiaoxiao He, Yuxiao Chen, et al. Proxedit: Improving tuning-free real image editing with proximal guidance. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4291–4301, 2024. 9, 12

[8] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross-attention control. In *ICLR*, 2023. 2, 9, 12

[9] Xuan Ju, Ailing Zeng, Yuxuan Bian, Shaoteng Liu, and Qiang Xu. Pnp inversion: Boosting diffusion-based editing with 3 lines of code. *International Conference on Learning Representations (ICLR)*, 2024. 9, 12

[10] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, page 2. Minneapolis, Minnesota, 2019. 3

[11] Bingyan Liu, Chengyu Wang, Tingfeng Cao, Kui Jia, and Jun Huang. Towards understanding cross and self-attention in stable diffusion for text-guided image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7817–7826, 2024. 2, 9, 12

[12] Nelson F Liu, Matt Gardner, Yonatan Belinkov, Matthew E Peters, and Noah A Smith. Linguistic knowledge and transferability of contextual representations. In *Proceedings of NAACL-HLT*, pages 1073–1094, 2019. 2

[13] Daiki Miyake, Akihiro Iohara, Yu Saito, and Toshiyuki Tanaka. Negative-prompt inversion: Fast image inversion for editing with text-guided diffusion models. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2063–2072. IEEE, 2025. 9, 12

[14] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6038–6047, 2023. 9

[15] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3

[16] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1921–1930, 2023. 9, 12

[17] Michael E Wall, Andreas Rechtsteiner, and Luis M Rocha. Singular value decomposition and principal component analysis. In *A practical approach to microarray data analysis*, pages 91–109. Springer, 2003. 2

[18] Xu Yao, Gilles Puy, Alasdair Newson, Yann Gousseau, and Pierre Hellier. High resolution face age editing. In *2020 25th International conference on pattern recognition (ICPR)*, pages 8624–8631. IEEE, 2021. 3, 9