The Aging Multiverse: Generating Condition-Aware Facial Aging Tree via Training-Free Diffusion

Bang Gong[1*] Luchao Qi[1*] Jiaye Wu[2] Zhicheng Fu[3]
Chunbo Song[3] David W. Jacobs[2] John Nicholson[3] Roni Sengupta[1]
[1]UNC Chapel Hill [2]University of Maryland [3]Lenovo
gongbang@cs.unc.edu, lqi@cs.unc.edu, jiayewu@cs.umd.edu
zcfu@motorola.com, csong2@lenovo.com, dwj@umd.edu, jnichol@lenovo.com, ronisen@cs.unc.edu

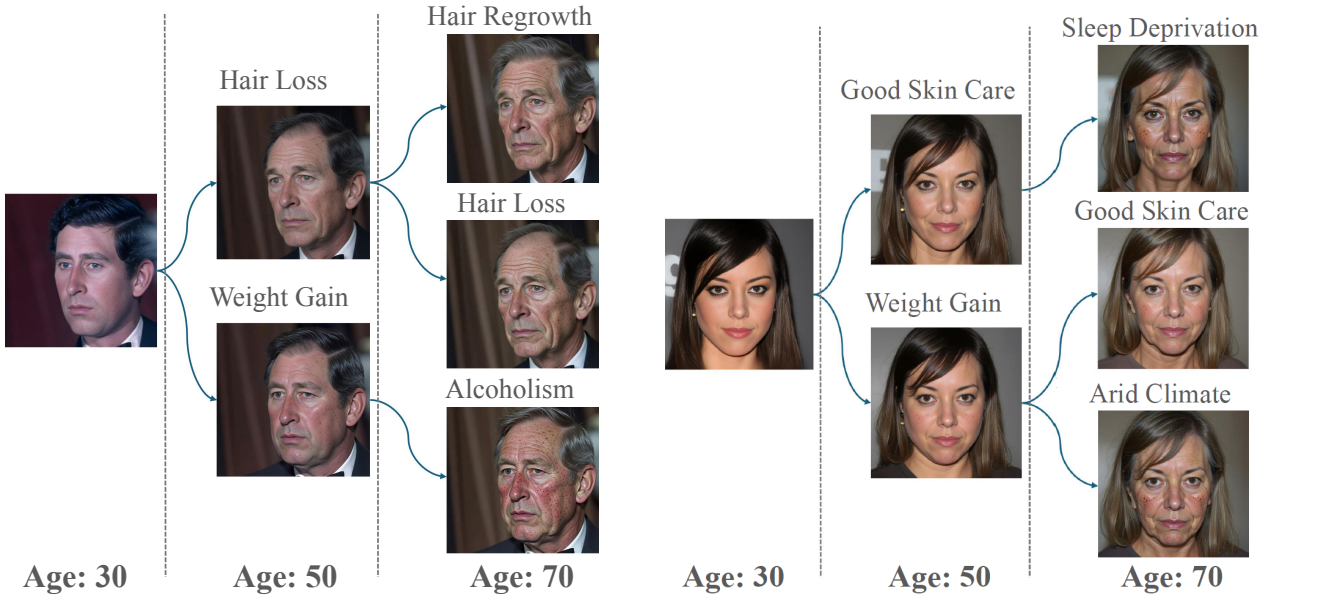

Figure 1. Given a single input image, our method generates an aging multiverse—multiple plausible aging trajectories conditioned on different external factors such as weight gain, skincare, hair loss, alcohol use, and environment. Each branch visualizes how environment, lifestyle, and health choices could shape appearance over time.

Abstract

We introduce the Aging Multiverse, a framework for generating multiple plausible facial aging trajectories from a single image, each conditioned on external factors such as environment, health, and lifestyle. Unlike prior methods that model aging as a single deterministic path, our approach creates an aging tree that visualizes diverse futures. To enable this, we propose a training-free diffusion-based method that balances identity preservation, age accuracy, and condition control. Our key contributions include attention mixing to modulate editing strength and a Simulated Aging Regularization strategy to stabilize edits. Extensive experiments and user studies demonstrate state-of-the-art performance across identity preservation, aging realism, and conditional alignment, outperforming existing editing and age-progression models, which often fail to account for one or more of the editing criteria. By transforming aging into a multi-dimensional, controllable, and interpretable process, our approach opens up new creative and practical avenues in digital storytelling, health education, and personalized visualization. Additional visual examples are available on our project website: https://agingmultiverse.github.io/

1. Introduction

What might you look like in your 60s? Would a consistent skincare routine make you appear more youthful? How would hair loss or alcohol addiction affect your appearance over time? While genetics plays a significant role in facial aging, external factors such as environmental exposure (e.g., sunlight, humidity), health conditions (e.g., stress, alcoholism, weight changes), and daily habits (e.g., skincare) can profoundly influence how we age. In this paper,

[1]* Equal contribution.

we introduce the concept of an aging multiverse—a framework for generating multiple plausible facial aging trajectories for an individual, each conditioned on different external factors. This approach enables a range of creative and practical applications. By transforming aging into a multi-dimensional, controllable, and interpretable process, the aging multiverse allows users to explore an "aging tree" of lifestyle-driven futures, empowering applications in digital storytelling, health education, and personalized visualization.

Prior portrait image age transformation methods [1, 4, 13, 15, 41] mostly focus on learning a global aging prior through pretraining on large human face datasets like FFHQ [19]. These approaches often do not consider the inherent plurality of the aging process, and when they do [23], they do not condition it on physical external factors that affect aging. In contrast to these approaches that generate only "aging line", we aim to generate an "aging tree" by developing a novel training-free conditional method.

Generating condition-aware aging paths from an input image requires strong preservation of identity, while editing aging features and introducing the specific attributes aligned with the conditions. While existing face transformation techniques [1, 4, 15] have now excelled in identity-preserving aging, they are unable to add any condition to their generation. Existing image-editing approaches [7, 37, 42], on the other hand, can handle different external conditions but struggle to balance all three criteria, e.g., RF-Solver-Edit [42] can preserve identity but struggles in aging and alignment to condition, and FlowEdit [20] can enable condition-aware aging but struggles to preserve identity. Our goal of generating an aging multiverse requires jointly editing age and conditional attributes while maintaining identity, evolving the conventional inversion-editability trade-off into a three-way balance among identity, age, and external conditions. To solve this problem, we propose a novel training-free method that can transform an input image into any target age under any condition defined in a text prompt.

Our key technical contribution lies in introducing training-free attention mixing and regularization strategies to enable multi-factorial aging while balancing identity preservation, age accuracy, and condition control. Central to our approach is the observation that the alignment between attention features for identity inversion and those for condition-aware editing determines the editability-identity trade-off. We leverage this by amplifying editing signals when these features align and attenuating them when they conflict. Building on this insight, we propose two modulation functions that operate on the Value and Key tensors of the attention blocks in a second-order Rectified Flow model [42]. Additionally, we introduce Simulated Aging Regularization, which applies unconditioned age progression to derive a stable aging trajectory, serving as a guide to further regularize attention features during condition-aware editing.

Our technique is training-free, and can be applied to any individual to simulate any age between 20-90 years old with external factors related to the environment, health, and lifestyle of the individual. This is in contrast to existing face age transformation techniques that either rely on global face aging datasets [4] or personalized datasets for training [35]. This training-free property ensures a plug-and-play framework, enabling universal compatibility with existing DiT-based models [9] and reduced computational resources.

We conduct comprehensive evaluations on both celebrity and non-celebrity images, using a combination of automated metrics and user studies to assess identity preservation, age accuracy, and consistency with the specified external conditions. Our evaluation indicates that our approach can provide the best balance across all three axes of image editing, while previous methods mostly succeed on one or two of them and fail for the rest. We further ablate the importance of our proposed attention mixing of key and value tensors, as well as the effectiveness of our attention regularization via simulated unconditional aging.

In summary, our key contributions are: (i) We formulate novel problem of generating an aging multiverse—or an aging tree—for an individual from a single image, simulating appearances across a wide age range (20–90) under varying environmental, health, and lifestyle conditions, a first to our knowledge. (ii) We propose a novel training-free framework that balances three key objectives: identity preservation, aging accuracy, and adherence to external conditions. (iii) We introduce attention mixing and attention regularization strategies that significantly improve the inversion-editability trade-off, leading to state-of-the-art performance on both celebrity and non-celebrity images.

2. Related Work

Age Transformation. Earlier approaches for face age transformation relied heavily on Generative Adversarial Networks (GANs), particularly StyleGAN2 [18], due to its disentangled latent space, leading to techniques that perform linear age-editing [30, 38] and later non-linear transformations [1, 13, 15, 35]. Recent methods have explored the Stable Diffusion model for age editing using age as a prompt in DiffAge3D [41] and FADING [4]. However, these approaches lack the ability to generate an "aging tree" by performing condition-aware aging. While PADA [23] models the stochastic nature of aging by introducing diversity in the diffusion latent space, it does not support conditional guidance like specific lifestyle or health attributes. In contrast, our method enables diverse, controllable age transformations conditioned on both target age and external factors, offering a more flexible and open framework for

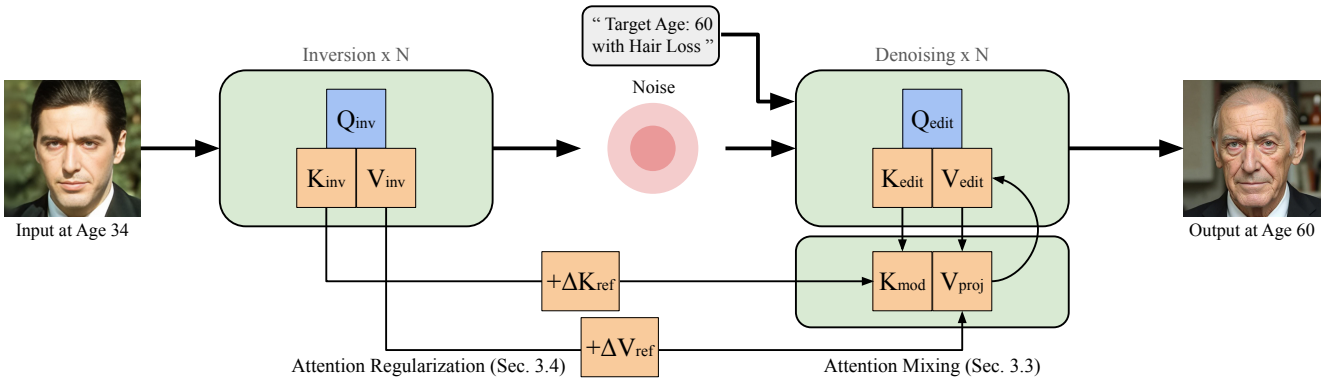

Figure 2. Overview of our training-free conditional-age progression framework. Given an input image and a textual description of external aging factors, our method leverages flow matching techniques to perform editing. Our approach balances three competing objectives—identity preservation, age accuracy, and condition alignment—enabling conditional age transformation without retraining. Our key innovations are: (i) attention mixing (§3.3) of Key and Value tensors between inversion and editing, and (ii) attention regularization (§3.3) with simulated unconditional aging to achieve the best inversion-editability trade-off.

multifaceted facial aging.

The Identity–Editability Trade-off in Diffusion Models. Diffusion-based editing method often first inverts the input image into a noisy latent and then denoises it with text prompt guidance. A key challenge lies in achieving faithful modifications to a specific attribute (e.g., age) while preserving the identity and structure of the original image. Early methods like DDIM inversion [39] and SDEdit [27] enabled diffusion-based editing but struggled with identity drift due to error accumulation. Subsequent techniques such as NTI [29], Negative-Prompt Inversion [28], and ReNoise [12] improved inversion fidelity and stability. To better balance identity and editability, recent works proposed manipulating internal representations during denoising—Prompt-to-Prompt [14], Plug-and-Play [40], and MasaCtrl [3] used attention control, while Asyrp [21] and Pix2Pix-Zero [33] introduced latent and semantic guidance strategies.

Flow-based diffusion models like Rectified Flow [25, 26] have enabled faster and more stable editing, with works such as RF-Inversion [37], FireFlow [7], and FTEdit [43] enhancing inversion quality through novel solvers, dynamic control, and fixed-point refinements. To improve the identity-editability trade-off, methods like RF-Solver-Edit [42], FireFlow [7], and KV-Edit [46] introduced feature sharing strategies, including attention reuse and memory-efficient KV caching inspired by language models.

Complementary works have also addressed architectural limits on editing fidelity: Stable Flow [2] and Head-Router [44] proposed routing and reweighting based on attention bottlenecks, while Fluxpace [5] and FlowChef [34] enabled controllable edits via latent direction extraction and optimization techniques.

While prior methods improve either fidelity or controllability, few consider age editing as a core task, and even fewer address the joint challenge of preserving identity, age accuracy, and external conditions. Our work addresses this gap with a novel training-free framework that introduces attention mixing and regularization techniques for generating condition-aware aging trajectories.

3. Method

We first introduce some preliminaries about Rectified Flow in Sec. 3.1. We then introduce key technical innovations of our pipeline in the following sections, starting with auto-generating a detailed prompt to create condition-aware aging using LLM in Sec. 3.2, followed by attention mixing for inversion-editability trade-off in Sec. 3.3, and simulated aging regularization for improving editng stability and robustness in Sec. 3.4.

3.1. Preliminaries

Rectified Flow (RF) models aim to learn a mapping between the real data distribution π_0 and a Gaussian noise distribution π_1 by modeling a velocity field v. This mapping is formulated as an ordinary differential equation (ODE):

$$\frac{dZ_t}{dt} = v(Z_t, t), \quad t \in [0, 1], \tag{1}$$

where $v(Z_t, t)$ denotes the velocity field at time t and state Z_t. In practice, this field is parameterized using DiT architectures [9, 22]. By design, the intermediate state Z_t follows a linear interpolation between $X_0 \sim \pi_0$ and $X_1 \sim \pi_1$. Formally:

$$Z_t \sim (1-t)X_0 + tX_1. \tag{2}$$

During sampling, the process begins with $Z_1 \sim \mathcal{N}(0, I)$. Given discretization steps $\{t_N, t_{N-1}, \ldots, t_0\}$, the ODE in

Eq. (1) is solved numerically as:

$$Z_{t_{i-1}} = Z_{t_i} + \int_{t_i}^{t_{i-1}} v_\theta(Z_\tau, \tau)\, d\tau, \quad i = N, N-1, \ldots, 1, \tag{3}$$

where v_θ is the learned velocity field.

To enable image editing with Rectified Flow models, prior work [37, 42] typically follows two steps: 1) **Inversion** maps the input image to the noise space using a diffusion transformer DiT as $Z_t = \text{DiT}(Z_{t-1}, t)$. 2) **Editing** performs denoising conditioned on a target text prompt (*txt*) and the inverted noise as $Z_{t-1} = \text{DiT}(Z_t,\ t,\ \text{txt})$. During inversion, the diffusion transformer consists of multiple attention blocks with the query, key, and value as $(Q_{\text{inv}},\ K_{\text{inv}},\ V_{\text{inv}})$. Similarly, during editing, the diffusion transformer has multiple attention blocks that depend on the text prompt (*txt*) with the query, key, and value as $(Q_{\text{edit}},\ K_{\text{edit}},\ V_{\text{edit}})$. To achieve editing that preserves the identity of the original image, RF-Solver-Edit [42] proposed replacing the value of editing attention layers with that of the inversion attention layer as:

$$(Q_{\text{edit}},\ K_{\text{edit}},\ V_{\text{edit}}) \leftarrow (Q_{\text{edit}},\ K_{\text{edit}},\ V_{\text{inv}}). \tag{4}$$

While this replacement improves identity and background fidelity, it often results in overfitting to the original image X^0, particularly in the facial region. As a consequence, the intended semantic edits are suppressed, as illustrated in Fig. 3, where RF-Solver-Edit [42] produces minimal visible changes to the input. Our approach also uses a similar Rectified Flow model, but proposes novel training-free attention feature modulation techniques that can balance identity preservation, age editing, and external conditions.

3.2. Prompt Refinement

While modern text-to-image (T2I) models generate highly detailed outputs, they often struggle to interpret complex or abstract conditions [17]. For instance, prompting Flux with "*a photo of a male at 40 years old addicted to alcohol*" fails to produce the expected facial characteristics. This is because high-level conditions like "alcohol addiction" are not directly grounded in visual features without additional context.

To bridge this gap, we refine prompts using GPT-4o [31], which helps translate abstract conditions into specific, low-level facial attributes. Following prior work [11], we use LLMs to expand vague inputs into detailed descriptions that guide the model more effectively. For example, we convert the original prompt into a refined version like "*a 40-year-old man with pale skin, sunken eyes, and facial wrinkles due to long-term alcohol abuse*," enabling the model to better align appearance with the intended condition.

3.3. Attention Mixing For Inversion-Editability Trade-off

Our goal is to enable text-driven age transformation while preserving input identity. A naive baseline is RF-Solver-Edit, which replaces self-attention values during denoising with those saved during inversion. This preserves identity and background well, but suppresses edits, since V_{edit}—which encodes the desired transformation—is fully discarded. As shown on the top of Fig. 3, editing an image of *Al Pacino* to appear "60 years old with hair loss" yields minimal visual change: identity is intact, but the edit fails.

Attention Value Projection. To address the above limitation, we propose a modulation-based fusion of V_{inv} and V_{edit} that allows us to retain identity while enabling stronger edits, balancing the inversion-editing trade-off. The core idea is intuitive: if the edit direction aligns with the identity features, we amplify it; if not, we suppress it to avoid identity loss. We formalize this by computing the orthogonal vector projection of V_{inv} onto V_{edit} as:

$$V_{\text{proj}} = \alpha V_{\text{edit}}, \quad \text{where } \alpha = \frac{\langle V_{\text{inv}}, V_{\text{edit}} \rangle}{\langle V_{\text{edit}}, V_{\text{edit}} \rangle} \tag{5}$$

Text Embedding Masking. Since DiT blocks jointly process text and image embeddings, we further refine our projection strategy to preserve guidance from the text prompt. Specifically, we mask out the text channels of V_{inv} when computing α in Eq. 5, ensuring that the projection only fuses image-related features. After computing α, we restore the text channels by setting their corresponding values in α to 1, thereby preserving prompt semantics from V_{edit} in V_{proj}.

Attention Key Modulation. Modifying only the attention value V can cause inconsistencies. In particular, since Q_{edit} and K_{edit} do not carry information from the inversion branch, the attention weights may misalign with the modulated V_{proj}, leading to unrealistic results such as distorted face or out-of-distribution images, as shown in ablation studies Fig. 8. To address this, we also adjust the key tensor K_{edit} using features from K_{inv}:

$$K_{\text{mod}} = K_{\text{edit}} + g \cdot (A \cdot K_{\text{inv}}), \quad \text{where } A = \text{softmax}\left(\frac{K_{\text{edit}} K_{\text{inv}}^T}{\sqrt{d_K}}\right) \tag{6}$$

Here, g is a scaling factor that controls the inversion-editing trade-off and A serves as an attention alignment matrix. This update ensures that the attention computation reflects both the editing goal and the identity constraint, improving consistency in the final output. Finally, attention layers during editing using DiT is computed using $(Q_{\text{edit}}, K_{\text{mod}}, V_{\text{proj}})$.

Method	External Condition		Age Accuracy		ID preservation	
	CLIP-T(↑)	Human Eval.(↑)	Age_{MAE}(↓)	Human Eval.(↑)	ID_{sim}(↑)	Human Eval.(↑)
FADING*	-	-	8.6	3.81	0.57	3.82
RF-Inversion	0.299	-	13.9	-	0.34	-
FlowChef	0.293	-	14.1	-	0.43	-
Fireflow	0.299	-	16.5	-	0.51	-
FlowEdit	0.303	3.30	13.4	3.72	0.42	3.28
RF-Solver-Edit	0.292	3.16	17.8	3.08	0.57	3.89
Ours w/o SAR	0.322	3.57	11.0	3.63	0.48	3.48
Ours	0.326	3.65	9.5	3.84	0.49	3.84

Table 1. Quantitative comparison of condition-aware age transformation on celebrity data. An asterisk (*) indicates methods that require aging-specific pre-training and cannot be conditioned on external prompts; thus, no CLIP-T score is reported for these methods. Red highlights the best result, Orange indicates the second best, and Yellow denotes the third best.

Input Age: 31 | Target Age: 60 | Condition: Hair Loss

Input Age: 20 | Target Age: 70 | Condition: Poor Skin Care

Figure 3. Visual comparison of our method with RF-Solver-Edit [42] and FlowEdit [20]. Given input images, we edit the faces with a desired aging condition. RF-Solver-Edit shows limited editing capability, yielding results very similar to the input. FlowEdit can generate some edits but leads to id drift and unrealistic skin texture. In contrast, our method achieves stronger edits that accurately reflect the prompt.

3.4. Simulated Aging Regularization in Attention Layers

Our attention mixing technique enhances editing by enabling fine-grained control over attention distributions. However, directly modifying attention can lead to unstable results or out-of-distribution generations [8, 45]. To address this, we introduce a Simulated Aging Regularization mechanism that leverages reference-based "age directions", derived by simulating age clusters, to guide attention feature modifications, resulting in a semantically grounded and stable editing.

Specifically, we generate unconditional age-progressed images from the input using GPT-4o [16], and diversify them with Arc2Face [32] to construct distinct clusters corresponding to both the input and target age ranges. From these clusters, we compute average self-attention features for representative older (e.g., age 70) and younger (e.g., age 30) groups, denoted as $\overline{V}_{\text{70 cluster}}$, $\overline{K}_{\text{70 cluster}}$, $\overline{V}_{\text{30 cluster}}$ and $\overline{K}_{\text{30 cluster}}$ respectively. We then define a semantic aging direction in the attention space as:

$$\Delta V_{\text{ref}} = \overline{V}_{\text{70 cluster}} - \overline{V}_{\text{30 cluster}} \tag{7}$$

$$\Delta K_{\text{ref}} = \overline{K}_{\text{70 cluster}} - \overline{K}_{\text{30 cluster}} \tag{8}$$

Given a target age Age_{target}, we compute a weighting factor based on the relative position between the target age Age_{target}, the upper reference bound Age_{high} (e.g., 70), and the lower reference bound Age_{low} (e.g., 30), and apply this to modulate the inversion features:

$$w = \frac{Age_{\text{target}} - Age_{\text{low}}}{Age_{\text{high}} - Age_{\text{low}}} \tag{9}$$

$$V_{\text{inv}} \leftarrow V_{\text{inv}} + w \cdot \Delta V_{\text{ref}}, \quad K_{\text{inv}} \leftarrow K_{\text{inv}} + w \cdot \Delta K_{\text{ref}} \tag{10}$$

This process effectively aligns the attention features with expected semantic changes due to aging. For instance, transforming a 30-year-old face to appear 50 years old would involve applying half the computed age direction: $V_{\text{inv}} = V_{\text{inv}} + 0.5\Delta V_{\text{ref}}$. This guidance leads to smoother, more realistic aging transformations, especially in challenging mid-life edits, while retaining the flexibility of text-driven conditioning and maintaining identity fidelity.

4. Experiments

We first discuss the evaluation setup in Sec. 4.1, including the Flux ODE solver we use, the methods we compare to, the conditions we evaluated on, and the metrics we use for numerical evaluation. In Sec. 4.2 and Sec. 4.3, we show details of numerical and visual comparisons of our method with other state-of-the-art methods on celebrity and non-celebrity images respectively. In Sec. 4.4, we present a user study that shows how humans evaluate the performance of our method against other approaches. Finally, in Sec. 4.5 we ablate the contributions of our attention mixing and regularization strategies.

4.1. Experimental Setup

Dataset. We evaluate our method on the same celebrity dataset used in the MyTM paper [35], which contains a curated set of real-life 12 celebrities spanning various age ranges. To further assess the robustness and generalization of our method, we additionally collect a set of 11 non-celebrity individuals with age annotations.

Baselines. Our method is built upon RF-Solver-Edit [42], employing its second-order RF-Solver for both inversion and denoising. We compare against several recent open-source flow-based image editing approaches: RF-Inversion [37], RF-Solver-Edit [42], FlowEdit [20], FireFlow [7], and FlowChef [34]. All baselines are evaluated using the same base model, Flux.1-dev, and under the same set of aging-related conditions: *alcoholism*, *gain weight*, *good skin care*, *poor skin care*, *hair loss*, *strong sunlight exposure*, and *living in dry windy climate*. To ensure a fair comparison, we apply the same prompt refinement strategy described in Sec. 3.2 across all models. In addition, we compare with the state-of-the-art for age transformation FADING [4] that uses Null-Text Inversion (NTI)[29] to perform text-guided re-aging from a single input image.

Metrics. We follow the evaluation framework used in Personalize Anything [10], using the *CLIP-T* score [36] to assess alignment between the edited image and the text prompt. To measure aging quality, use report *Age Mean Absolute Error (Age$_{MAE}$)*, which compares the target age to the predicted age from FP-Age [24], and *Identity Similarity (ID$_{sim}$)*, computed as the ArcFace [6] embedding similarity between the edited image and reference images of the same person. For the celebrity dataset, we use images of the subject at the target age as references. For the non-celebrity dataset, where multi-age references are unavailable, we use the input image itself as the reference. In addition to quantitative metrics, we conduct a human evaluation comparing our method against three main baselines, RF-Solver-Edit [42], FlowEdit [20], and FADING [4], to assess editing quality in terms of age accuracy, identity preservation, condition alignment, and overall preference.

4.2. Evaluation on Celebrity Dataset

We present a comparison of our method with recent flow-based image editing models and state-of-the-art age transformation methods, summarized in Fig. 4. We evaluate performance on three key metrics: *CLIP-T* for prompt alignment, *Age MAE* for aging accuracy, and *ID Similarity* for identity preservation in Table 1.

Among the flow-based editing baselines, our method achieves the best overall balance between editability and identity preservation. It ranks highest in *CLIP-T* (0.326), indicating strong alignment with the target prompt, and highest in *Age MAE* (9.5), demonstrating accurate age transformation. While its *ID Similarity* score (0.49) is lower than that of RF-Solver-Edit (0.57) and FireFlow (0.51), this is expected: those methods prioritize identity preservation by directly replacing attention features, as discussed in Sec. 3.3. As a result, they achieve near-perfect fidelity to the input image but show minimal responsiveness to editing prompts, as reflected in their significantly lower *CLIP-T* scores and higher *Age MAE* values. Visual comparisons in Fig. 3 and 4 further illustrate this trade-off: models with high ID$_{\text{sim}}$ scores tend to overfit to the input appearance, producing outputs with minimal visual change. Our method, by contrast, produces edits that are both prompt-aligned and identity-aware.

Method	CLIP-T (↑)	Age$_{MAE}$ (↓)	ID$_{sim}$ (↑)
FADING*	-	10.4	0.65
RF-Inversion	0.311	18.5	0.28
FlowChef	0.312	18.1	0.47
FlowEdit	0.322	17.6	0.40
Fireflow	0.300	23.3	0.62
RF-Solver-Edit	0.287	25.2	0.75
Ours	0.331	13.9	0.49

Table 2. Quantitative comparison of condition-aware age transformation on non-celebrity dataset. * indicates methods that require aging-specific pre-training and cannot be conditioned on external prompts; thus, no CLIP-T score is reported for these methods. Red highlights the best result, Orange indicates the second best, and Yellow denotes the third best.

Compared to FADING—an age transformation specialist—our outputs are visually comparable in terms of aging realism, even though FADING achieves slightly better *Age MAE* and *ID Similarity*. Importantly, FADING does not support conditional prompts (e.g., "60 years old with hair loss"), limiting its flexibility. Our approach not only edits toward target ages but also adapts to lifestyle or health-related conditions.

4.3. Evaluation on Non-celebrity Dataset

Similar to results in Sec. 4.2, our method achieves the best balance across all three evaluation metrics for in-the-wild non-celebrities as shown in Table 2. It achieves the highest *CLIP-T* score (0.331), indicating strong alignment with the target prompt, and ranks second in *Age MAE* (13.9), behind FADING, which trains a dedicated aging model. While its *ID Similarity* score (0.49) is not the highest, it still outperforms most flow-based baselines, which often suffer from poor identity preservation or limited editability.

For instance, RF-Solver-Edit and FireFlow achieve the highest identity scores (0.75 and 0.62, respectively), but their performance on *CLIP-T* and *Age MAE* indicates minimal responsiveness to prompts and weak aging accuracy, with visual examples shown in Fig. 5. On the other hand, FlowEdit shows strong prompt alignment and age prediction but fails to preserve identity. Our method sits at a favorable point in this trade-off space, successfully editing facial

Figure 4. Given the input celebrity image on the left and the editing context indicated below each row, we present a visual comparison of our method with RF-Solver-Edit [42], FireFlow [7], FlowEdit [20], FlowChef [34], and FADING [4]. For FADING, only the aging effect is evaluated.

attributes according to the prompt while maintaining visual consistency with the input identity. Compared to FADING, which benefits from extensive age-specific pretraining, our method achieves competitive aging performance without requiring any task-specific finetuning, underscoring its flexibility and robustness in a general editing framework.

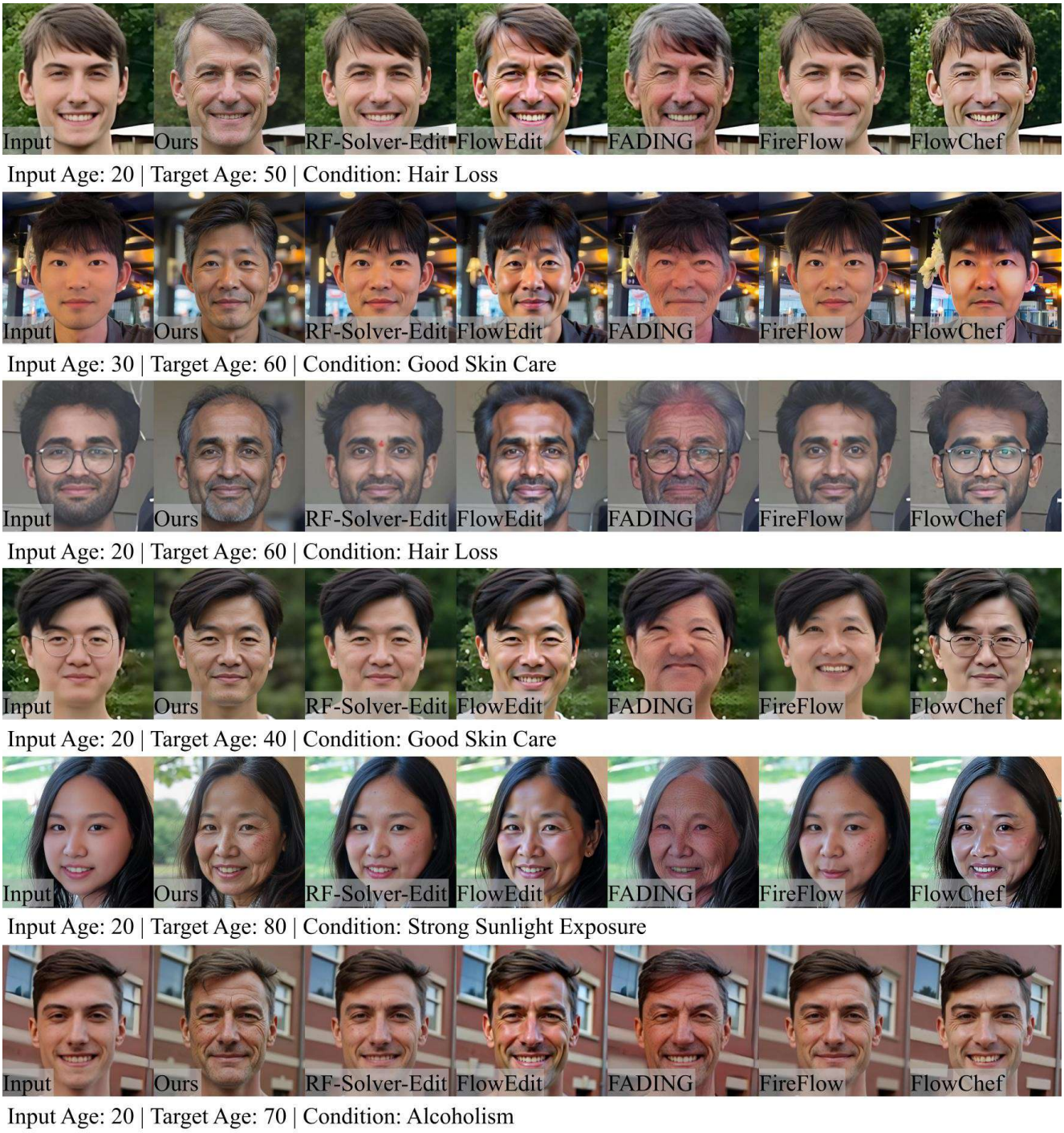

Figure 5. Given the input in-the-wild non-celebrity image on the left and the editing context indicated below each row, we present a visual comparison of our method with RF-Solver-Edit [42], FireFlow [7], FlowEdit [20], FlowChef [34], and FADING [4]. For FADING, only the aging effect is evaluated.

4.4. User Study

Given the novelty of our task, we acknowledge that existing evaluation metrics are not always well-suited to capture the perceptual quality of edits. In particular, the *ID Similarity* metric used in [35]—based on ArcFace [6] embedding similarity between the edited image and either a reference or the input image—can introduce bias against our method.

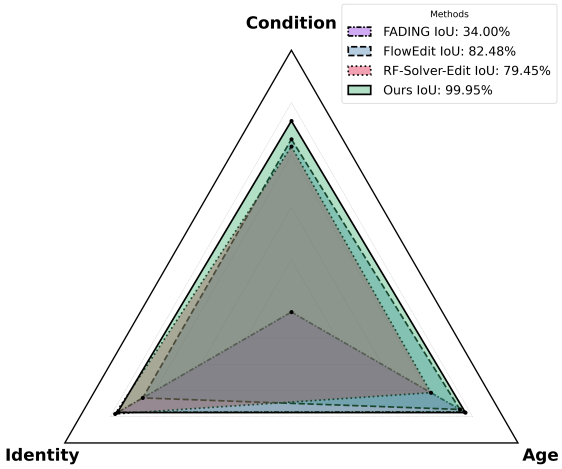

Figure 6. Visualization of human evaluation scores across three axes: external condition alignment, age accuracy, and ID preservation from Table. 1. For each method, we calculate IoU as the ratio of area covered by the method's triangle over the union of all four triangles. This indicates our approach provides the best balance across all three criterion.

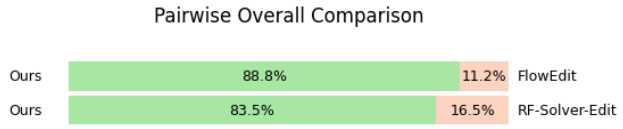

Figure 7. Pairwise user study comparing our method to FlowEdit and RF-Solver-Edit to judge overall condition-aware age editing performance. Values indicate the percentage of users preferring our method over each baseline.

For celebrity data, the reference image corresponds to the target age, while for non-celebrities, the input image is used directly. This setup penalizes edits that introduce realistic aging and external condition effects not present in the real reference. For example, in Fig. 3, editing a young image of *Al Pacino* to appear 60 years old with hair loss yields a lower similarity score when compared to actual photos of him at that age, since he did not exhibit hair loss in reality. Such discrepancies become even more pronounced under conditions like poor skin care or chronic sun exposure, which can drastically alter facial appearance in ways that diverge from available reference images.

To address these limitations and better assess perceptual quality, we conducted a user study aimed at capturing human judgments in evaluating the quality of condition-aware aging. We conducted a user study via Amazon Mechanical Turk to evaluate three key aspects of the edited images: alignment with external conditions, age accuracy, and identity preservation. Participants rated each image on a 1–5 scale across these criteria. We select 10 images of varying identity spanning 7 different external conditions for this user study. Each image was rated by 15 different users. We ask users to separately judge each of the three criterion: alignment with external conditions, age accuracy, and

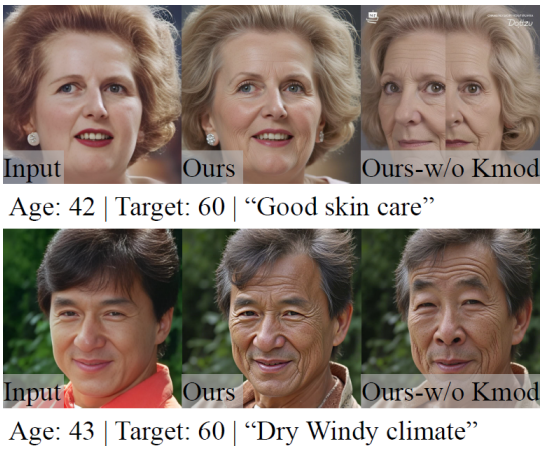

Figure 8. Visual ablation of Attention Mixing with and without K modulation. Without K modulation, the edits are not stable, often leading to distorted face or our-of-distribution image.

identity preservation in different HITs. Based on numerical results in Table 1, we identify FADING [4], FlowEdit [20] and RF-Solver-Edit [42] as our main competitors.

The average scores are presented in Table 1, alongside the corresponding quantitative metrics. Our method achieved the highest human ratings (Human Eval.) for both condition alignment and age accuracy, even slightly outperforming the aging-specialized model FADING. In terms of identity preservation, our score was slightly below that of RF-Solver-Edit, but notably closer than suggested by the embedding-based *ID Similarity* metric. This highlights a key observation: our edits are perceived by humans as identity-consistent, despite being penalized by reference-based ID similarity metrics. To visualize overall performance, we use a radar plot (Fig. 6) with three axes corresponding to the mean human scores on our evaluation criteria. Each method forms a triangle, and we compute the Intersection over Union (IoU) as the ratio of its area to the union of all method areas. Our method achieves the largest triangle with a 99.95% IoU, reflecting balanced performance across all dimensions.

While per-criterion user ratings are informative, they don't capture overall preferences—e.g., whether users favor minimal edits that preserve identity or more aggressive edits that compromise it. To assess holistic quality, we conducted a pairwise preference study comparing our method to FlowEdit and RF-Solver-Edit (excluding FADING, which cannot edit external conditions). Across 10 celebrity and 5 non-celebrity images, each rated by 10 users, participants preferred our results in 85% of cases (Fig.7), highlighting our method's strong balance between editability and identity fidelity.

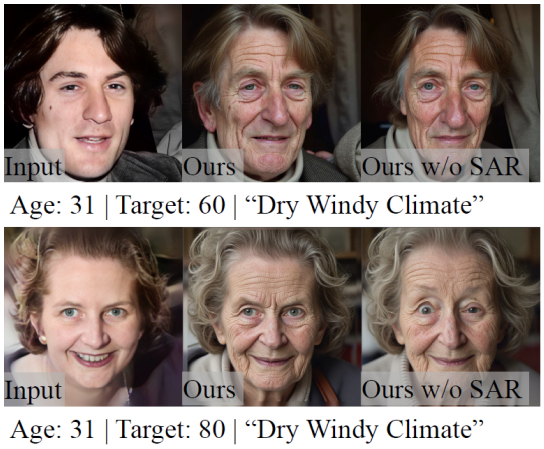

Figure 9. Visual comparison between Ours and Ours w/o SAR. The results show that our proposed aging regulation helps stabilize editing results, preventing identity drift from the input image or unrealistic face distortions.

4.5. Ablation Study

We perform a series of ablation studies to justify the effectiveness of our proposed components. Table 3 presents the quantitative results, with each row incrementally building upon the RF-Solver-Edit [42] baseline. We begin by replacing the original value tensor copying strategy between identity and editing blocks with our proposed value projection approach, without masking out text embedding channels during the computation of α. This change improves both *CLIP-T* and Age_{MAE}, indicating stronger prompt alignment and more accurate aging. As expected, ID_{sim} decreases due to the increased influence of semantic edits, consistent with our earlier observations.

Next, we introduce the text embedding masking strategy (Sec. 3) to disentangle identity and prompt information. This further boosts *CLIP-T* to 0.317 and reduces *AgeMAE* to 12.5, but again slightly lowers *ID Similarity*, as the edits become more visually distinct. However, as illustrated in Fig. 8, this increase in editing power may occasionally lead to instability or unrealistic outputs. To mitigate these issues, we first incorporate the Key modulation technique to further improve the editing quality as shown in Fig. 8. Finally, our Simulated Aging Regularization (SAR) further enhances edit stability while partially improving identity preservation, as shown in Fig. 9. This final version of our method achieves strong semantic edits with improved robustness, validating the effectiveness of our design choices.

5. Conclusion

In this paper, we define the novel task of diverse-conditioned age transformation and present a simple, training-free method that extends flow-based models like Flux to this problem. By manipulating attention during inference, our approach adds minimal computational overhead while effectively combining identity preservation with text-driven edits. It enables the generation of aging trees that reflect various lifestyle and environmental conditions. Our method broadens the scope of facial aging research, achieving strong performance across both qualitative and quantitative benchmarks, outperforming prior Flux-based editors and matching the capabilities of state-of-the-art aging models.

Method	CLIP-T↑	Age_{MAE}↓	ID_{sim}↑
RF-Solver-Edit (baseline)	0.292	17.8	0.57
+ Att. Mixing (Value only)	0.304	14.4	0.50
+ Text Embedding Masking	0.317	12.5	0.47
+ Att. Mixing (Value & Key)	0.322	11.0	0.48
+ Simulated Aging Regularization	0.326	9.5	0.49

Table 3. Ablation study validating the effectiveness of attention mixing (dynamic feature modulation), text channel masking, and attention regularization (simulated aging regularization).

Figure 10. Failure cases of our method when input image is low-quality, which often leads to identity drift from input image or desired edits not generated.

Limitation. Although we introduced multiple strategies to stabilize the attention mixing, our method is still input-sensitive. As shown in Fig. 10, our method could fail on low-quality heavily pre-processed images by changing the identity or failing to generate desired edit, especially when the edit requires significant change.

Ethical Consideration. The photorealistic facial aging trajectories generated by our method are algorithmic simulations, not deterministic predictions. While the trajectory is plausible and can be used for visual effects applications, it should not be used for facial identification purposes as it may likely raise many false positives. Rigorous human subject testing must be performed before such a tool can be deployed for health education or lifestyle choice determination.

Acknowledgments

This research was supported in part by Lenovo Research (Morrisville, NC). We gratefully acknowledge the invaluable support and assistance of the members of the Mobile Technology Innovations Lab. This work was also supported in part by the National Science Foundation under Grant No. 2213335.

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*, 40(4):1–12, 2021.

[2] Omri Avrahami, Or Patashnik, Ohad Fried, Egor Nemchinov, Kfir Aberman, Dani Lischinski, and Daniel Cohen-Or. Stable Flow: Vital Layers for Training-Free Image Editing, 2024.

[3] 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 (ICCV)*, pages 22560–22570, 2023.

[4] Xiangyi Chen and Stéphane Lathuilière. Face aging via diffusion-based editing. In *34th British Machine Vision Conference 2023, BMVC 2023, Aberdeen, UK, November 20-24, 2023*. BMVA, 2023.

[5] Yusuf Dalva, Kavana Venkatesh, and Pinar Yanardag. FluxSpace: Disentangled Semantic Editing in Rectified Flow Transformers, 2024.

[6] Jiankang Deng, Jia Guo, Jing Yang, Niannan Xue, Irene Kotsia, and Stefanos Zafeiriou. ArcFace: Additive Angular Margin Loss for Deep Face Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):5962–5979, 2022.

[7] Yingying Deng, Xiangyu He, Changwang Mei, Peisong Wang, and Fan Tang. Fireflow: Fast inversion of rectified flow for image semantic editing, 2024.

[8] Mischa Dombrowski, Hadrien Reynaud, Johanna P Müller, Matthew Baugh, and Bernhard Kainz. Tradeoffs in fine-tuned diffusion models between accuracy and interpretability. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 21037–21045, 2024.

[9] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling Rectified Flow Transformers for High-Resolution Image Synthesis, 2024.

[10] Haoran Feng, Zehuan Huang, Lin Li, Hairong Lv, and Lu Sheng. Personalize anything for free with diffusion transformer. *arXiv preprint arXiv:2503.12590*, 2025.

[11] Tsu-Jui Fu, Wenze Hu, Xianzhi Du, William Yang Wang, Yinfei Yang, and Zhe Gan. Guiding instruction-based image editing via multimodal large language models, 2024.

[12] Daniel Garibi, Or Patashnik, Andrey Voynov, Hadar Averbuch-Elor, and Daniel Cohen-Or. Renoise: Real image inversion through iterative noising. In *European Conference on Computer Vision*, pages 395–413. Springer, 2024.

[13] Guillermo Gomez-Trenado, Stéphane Lathuilière, Pablo Mesejo, and Óscar Cordón. Custom structure preservation in face aging. In *Computer Vision – ECCV 2022*, pages 565–580, Cham, 2022. Springer Nature Switzerland.

[14] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-or. Prompt-to-prompt image editing with cross-attention control. In *The Eleventh International Conference on Learning Representations*, 2023.

[15] Gee-Sern Hsu, Rui-Cang Xie, Zhi-Ting Chen, and Yu-Hong Lin. Agetransgan for facial age transformation with rectified performance metrics. In *Computer Vision – ECCV 2022*, pages 580–595, Cham, 2022. Springer Nature Switzerland.

[16] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

[17] Hao Kang, Stathi Fotiadis, Liming Jiang, Qing Yan, Yumin Jia, Zichuan Liu, Min Jin Chong, and Xin Lu. Flux already knows-activating subject-driven image generation without training. *arXiv preprint arXiv:2504.11478*, 2025.

[18] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and Improving the Image Quality of StyleGAN . In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8107–8116, Los Alamitos, CA, USA, 2020. IEEE Computer Society.

[19] Tero Karras, Samuli Laine, and Timo Aila. A Style-Based Generator Architecture for Generative Adversarial Networks . *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 43(12):4217–4228, 2021.

[20] Vladimir Kulikov, Matan Kleiner, Inbar Huberman-Spiegelglas, and Tomer Michaeli. FlowEdit: Inversion-Free Text-Based Editing Using Pre-Trained Flow Models, 2024.

[21] Mingi Kwon, Jaeseok Jeong, and Youngjung Uh. Diffusion models already have a semantic latent space. In *The Eleventh International Conference on Learning Representations*, 2023.

[22] Black Forest Labs. Flux. `https://github.com/black-forest-labs/flux`, 2024.

[23] Peipei Li, Rui Wang, Huaibo Huang, Ran He, and Zhaofeng He. Pluralistic Aging Diffusion Autoencoder. In *2023 IEEE/CVF International Conference*

on Computer Vision (ICCV), pages 22556–22566, Paris, France, 2023. IEEE.

[24] Yiming Lin, Jie Shen, Yujiang Wang, and Maja Pantic. Fp-age: Leveraging face parsing attention for facial age estimation in the wild. *IEEE Transactions on Image Processing*, pages 1–1, 2022.

[25] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow Matching for Generative Modeling, 2023.

[26] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow Straight and Fast: Learning to Generate and Transfer Data with Rectified Flow, 2022.

[27] 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, 2022.

[28] Daiki Miyake, Akihiro Iohara, Yu Saito, and Toshiyuki Tanaka. Negative-prompt inversion: Fast image inversion for editing with text-guided diffusion models, 2024.

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

[30] Yotam Nitzan, Rinon Gal, Ofir Brenner, and Daniel Cohen-Or. LARGE: Latent-Based Regression through GAN Semantics. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19217–19227, New Orleans, LA, USA, 2022. IEEE.

[31] OpenAI and Hurst et al. Gpt-4o system card, 2024.

[32] Foivos Paraperas Papantoniou, Alexandros Lattas, Stylianos Moschoglou, Jiankang Deng, Bernhard Kainz, and Stefanos Zafeiriou. Arc2Face: A Foundation Model of Human Faces, 2024.

[33] Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. Zero-shot Image-to-Image Translation, 2023.

[34] Maitreya Patel, Song Wen, Dimitris N. Metaxas, and Yezhou Yang. Steering rectified flow models in the vector field for controlled image generation. *arXiv preprint arXiv:2412.00100*, 2024.

[35] Luchao Qi, Jiaye Wu, Bang Gong, Annie N. Wang, David W. Jacobs, and Roni Sengupta. MyTimeMachine: Personalized Facial Age Transformation, 2024.

[36] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[37] Litu Rout, Yujia Chen, Nataniel Ruiz, Constantine Caramanis, Sanjay Shakkottai, and Wen-Sheng Chu. Semantic Image Inversion and Editing using Rectified Stochastic Differential Equations, 2024.

[38] Yujun Shen, Ceyuan Yang, Xiaoou Tang, and Bolei Zhou. Interfacegan: Interpreting the disentangled face representation learned by gans. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(4): 2004–2018, 2022.

[39] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising Diffusion Implicit Models, 2022.

[40] 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 (CVPR)*, pages 1921–1930, 2023.

[41] Junaid Wahid, Fangneng Zhan, Pramod Rao, and Christian Theobalt. Diffage3d: Diffusion-based 3d-aware face aging, 2024.

[42] Jiangshan Wang, Junfu Pu, Zhongang Qi, Jiayi Guo, Yue Ma, Nisha Huang, Yuxin Chen, Xiu Li, and Ying Shan. Taming Rectified Flow for Inversion and Editing, 2024.

[43] Pengcheng Xu, Boyuan Jiang, Xiaobin Hu, Donghao Luo, Qingdong He, Jiangning Zhang, Chengjie Wang, Yunsheng Wu, Charles Ling, and Boyu Wang. Unveil inversion and invariance in flow transformer for versatile image editing. *arXiv preprint arXiv:2411.15843*, 2024.

[44] Yu Xu, Fan Tang, Juan Cao, Yuxin Zhang, Xiaoyu Kong, Jintao Li, Oliver Deussen, and Tong-Yee Lee. Headrouter: A training-free image editing framework for mm-dits by adaptively routing attention heads. *arXiv preprint arXiv:2411.15034*, 2024.

[45] Jing Nathan Yan, Jiatao Gu, and Alexander M Rush. Diffusion models without attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8239–8249, 2024.

[46] Tianrui Zhu, Shiyi Zhang, Jiawei Shao, and Yansong Tang. Kv-edit: Training-free image editing for precise background preservation. *arXiv preprint arXiv:2502.17363*, 2025.