Supplementary of "DreamSalon: A Staged Diffusion Framework for Preserving Identity-Context in Editable Face Generation"

Supplementary Material

We start with Appendix 1, which shows the algorithmic description of our framework and theoretical analyses of staged editing. Next, in Appendix 2, we show some more comparisons with current SOTAs and editing on more identities. Moreover, Appendix 3 provides the effects of different stage configurations and the effects of varying text prompts. Lastly, we provide additional implementation details in Appendix 4. We show additional qualitative results in Fig. 5~9.

1. Theoretical Analysis

An overview of our method is provided in the main paper. Moreover, the algorithmic description of our method is summarized in 1.

Algorithm 1 The Framework of Staged Editing

Input: A pretrained diffusion model θ with its decoder ψ, a set of images $\mathbf{X}$ of an identity (2~4 images), source and target prompts $(\mathbf{c}_{(0)}, \mathbf{c}_{(1)})$, the default weight factor λ', and any inference image $\mathbf{x}_{\text{source}}$ of this identity

Output: The edited image $\mathbf{x}_{\text{edit}}$

```
 1: // Training a HyperNetwork $\Upsilon$ for weights generation
 2: $\hat{\Upsilon} \leftarrow \min_{\theta_\Upsilon} \mathcal{L}_{\text{HyperDreambooth}}(\mathbf{X})$;    ▷ Eq. 5
 3: // Obtain the personalized weights $\theta_{per}$ for $\mathbf{x}_{source}$
 4: $\theta_{\text{per}} \leftarrow \hat{\Upsilon}(\mathbf{x}_{\text{source}})$;
 5: // Replace attention weights in $\theta$ with $\theta_{per}$
 6: $\hat{\theta} \leftarrow \text{Loading}(\theta_{\text{per}})$
 7: // Compute the covariance guidance
 8: $\text{CovDiff} = \text{Normalize}(\max_{i \text{ or } j} |\text{Cov}_{\mathbf{c}_{(1)}} - \text{Cov}_{\mathbf{c}_{(0)}}|)$
 9: for $t \in [T, T-1, \ldots, 1, 0]$ do
10:     if $t \geq t_{\text{edit}}$ then
11:         // During the editing stage; $\mathbf{z}$ is the latent codes
12:         $\lambda_t^{\text{init}} = \text{Normalize}(\text{FFT}(\psi(\mathbf{z}_t)))$
13:         $\hat{\mathbf{c}}_t^{\text{mixed}} = \text{CovDiff} \odot \left((1-\lambda_t^{\text{init}})\mathbf{c}_{(0)} + \lambda_t^{\text{init}}\mathbf{c}_{(1)}\right)$
14:         $\mathbf{z}_{t-1} = \sqrt{\alpha_{t-1}}\mathbf{P}(\epsilon_t^{\hat{\theta}}(\mathbf{z}_t, \hat{\mathbf{c}}_t^{\text{mixed}})) + \mathbf{D}(\epsilon_t^{\hat{\theta}}(\mathbf{z}_t, \hat{\mathbf{c}}_t^{\text{mixed}}))$
15:     else
16:         // After the editing stage
17:         $\lambda_t^{\text{init}} = \lambda'$
18:         $\hat{\mathbf{c}}_t^{\text{mixed}} = (1-\lambda_t^{\text{init}})\mathbf{c}_{(0)} + \lambda_t^{\text{init}}\mathbf{c}_{(1)}$
19:         if $t > t_{\text{boost}}$ then
20:             // Same formula as in step 10
21:             $\mathbf{z}_{t-1} = \sqrt{\alpha_{t-1}}\mathbf{P}_t + \mathbf{D}_t$
22:         else
23:             // During boosting stage
24:             $\mathbf{z}_{t-1} = \sqrt{\alpha_{t-1}}\mathbf{P}_t + \mathbf{D}_t + \sigma_t \mathbf{z}_t$
25:         end if
26:     end if
27: end for
28: $\mathbf{x}_{\text{edit}} = \psi(\mathbf{z}_0)$
29: return $\mathbf{x}_{\text{edit}}$
```

1.1. Frequency of Predicted Noises

To demonstrate that images with higher frequencies can be manipulated more efficiently, we measure the change in the uniformity and smoothness of their frequency spectrum (Eq. 1) upon introducing random disturbances. Uniformity, quantified by entropy and coefficient of variation, assesses the even distribution of frequency components, and smoothness, measured by the standard deviation, reflects the gradual variation in these components across the spectrum [1]. Fig. 1 reveals that disturbances at time step 35 introduce significant changes in uniformity and smoothness, whereas disturbances at time step 5 are less impactful.

Editing Stage ($t = 35$): At this stage, the original image exhibits a high degree of noise, lacking any identifiable characteristics. The frequency spectrum's magnitude is relatively even, indicating a greater degree of randomness suitable for more impactful edits. The disturbances introduced lead to higher increases in entropy and variation, which facilitates more effective modifications without altering the image's core identity or context.

Boosting Stage ($t = 5$): As the diffusion process progresses, the magnitude spectrum shows a concentration of energy in the low-frequency region (center of the spectrum), indicating the presence of more defined features and less high-frequency noise. The introduction of disturbance at this stage results in a lower increase of entropy and variation, indicating less intense manipulations. Hence, editing at this stage should be more conservative to maintain the image's integrity and recognizable features.

$$\mathbf{z}_{fft} = \frac{\sum_{k_1=0}^{4-1} \sum_{k_2=0}^{63-1} \sum_{k_3=0}^{63-1} \mathbf{z}_{k_1,k_2,k_3} \times e^{-2\pi j \times (k_1 \times \frac{k_2}{64} \times \frac{k_3}{64})}}{d_1 \times d_2 \times d_3}, \quad (1)$$

where 4, 64, 64 are the dimensions of latent codes $\mathbf{z}$.

1.2. Gradient-Guided Boosting Stage

In a diffusion model, the process of adding noise and then reversing it is characterized by how the image changes at each timestep. For stochastic denoising, randomness is introduced into the reversal process, which can lead to the accumulation of errors if not controlled properly. To inspect how gradients with respect to noise prediction influence error accumulation, we begin with the reverse process in the

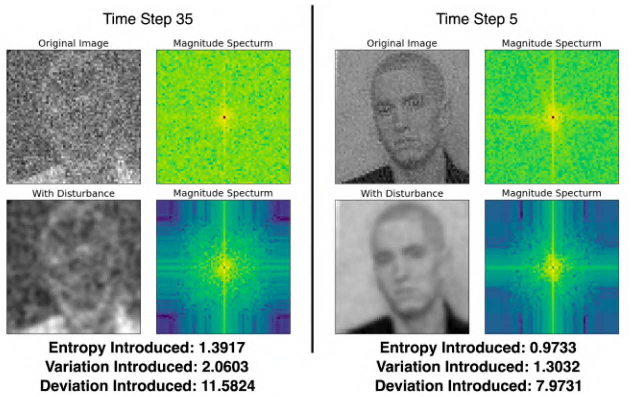

Figure 1. Frequency Spectrum Analysis upon Disturbance Introduction: This figure illustrates the effects of random disturbances on the uniformity and smoothness of the frequency spectrum at different time steps. The left panel shows time step 35 with greater changes in entropy, variation, and deviation, indicating a higher manipulation sensitivity.

presence of stochastic denoising:

$$x_{t-1} = \epsilon_\theta(x_t, t) + \sigma_t \epsilon_t, \tag{2}$$

where:

- $\epsilon_\theta(x_t, t)$: the denoising function that predicts the noise-free image from the noisy input x_t at timestep t.
- σ_t: the variance of the reverse process.
- ϵ_t: a standard Gaussian random variable, representing the stochastic component of the denoising process.

Now, consider the gradient of the predicted noise with respect to the noisy image x_t, denoted by $\nabla_{x_t} \epsilon_\theta(x_t, t)$. This gradient reflects how sensitive the noise prediction is to changes in x_t. When this gradient is large, a small change in x_t can lead to a large change in the noise prediction, which can cause significant changes in the pixel values and result in error accumulation. The error introduced at each timestep can be quantified by the expected squared difference between the denoised image x_{t-1} and the true noise-free image x_0:

$$E[\|x_{t-1} - x_0\|^2]. \tag{3}$$

The change in this error as we move from timestep t to $t-1$ can be approximated by a Taylor expansion around x_t:

$$\begin{aligned}\Delta E_t \approx\ & \nabla_{x_t} E[\|x_t - x_0\|^2]^\top (\epsilon_\theta(x_t, t) - x_t) \\ & + \frac{1}{2}\sigma_t^2 \mathrm{Tr}(\nabla_{x_t}^2 E[\|x_t - x_0\|^2]).\end{aligned} \tag{4}$$

The first term represents the deterministic part of the error change, which is directly influenced by the gradient $\nabla_{x_t} \epsilon_\theta(x_t, t)$. The second term, involving the trace of the Hessian, represents the stochastic part of the error change, which is scaled by σ_t^2. To minimize error accumulation, we want the change in error ΔE_t to be as small as possible. If $\nabla_{x_t} \epsilon_\theta(x_t, t)$ is small, then the deterministic part of ΔE_t will be small. Furthermore, by carefully choosing σ_t, we can ensure that the stochastic part of ΔE_t does not introduce significant error.

Therefore, the goal is to find the point in the diffusion process where $\nabla_{x_t} \epsilon_\theta(x_t, t)$ is small, which corresponds to the time when the predicted noise is less sensitive to changes in x_t. This point is ideal for applying stochastic denoising, as it minimizes the risk of error accumulation and ensures that content is not significantly modified during the boosting stage. This theoretical foundation underlies the practice of choosing the appropriate timing for noise introduction to maintain the fidelity of the image's content.

2. More Experimental Details

2.1. The List of Prompts

For qualitative and quantitative evaluation, we employ the editing prompts listed in Tab. 1, which contains 20 editing prompts for each subject.

2.2. More Comparisons to Current SOTAs

In the main paper, we compare the image editing performance of various methods using a collection of human face images and textual prompts. For this set of images and prompts, Fig. 5 combines them one by one to more fully demonstrate the results achieved by our method.

Extending the comparison in the main paper, Fig. 6 presents a similar evaluation with the same identities but with alternative prompts. The results are consistent with the comparisons shown in the main paper, which demonstrates that our approach is able to preserve the integrity of the identity and context while performing the requested manipulations. Additionally, Fig. 7 displays the combinations of identities and prompts *w.r.t.* Fig. 6.

2.3. Fine Editing on Diverse Identities

The effectiveness of our method is showcased by its precision in editing while maintaining the subject's identity and context. This capability is crucial for applications demanding the subject's essence to remain unchanged despite edits. Fig. 8 demonstrates this, where edits on various identities correspond to the comparison in the main paper. Each edited image retains the unique characteristics that define the individual's identity, such as facial structure, skin texture, and inherent expression, while seamlessly integrating the specified alterations.

DreamSalon's robustness across different identities and prompts highlights its sophisticated handle on diverse human features, ensuring dependable results. This generalizability confirms the method's strength and real-world relevance for personalized, context-aware image modifications.

Table 1. Text prompt list for quantitative evaluation.

Text prompts for highly semantic parts	Text prompts for low semantic parts
“a photo of a [V] face, with glasses”	“a photo of a [V] face, with red hair”
“a photo of a [V] face, with chinstrap beard”	“a photo of a [V] face, with green hair”
“a photo of a [V] face, with a mustache”	“a photo of a [V] face, with red lips”
“a photo of a [V] face, with eyes closed”	“a photo of a [V] face, with black eyeshadows”
“a photo of a [V] face, smiling”	“a photo of a [V] face, with thick eyebrows”
“a photo of a [V] face, with a smirk”	“a photo of a [V] face, with blue eyes”
“a photo of a [V] face, with fringe hairstyle”	“a photo of a [V] face, with exaggerated blush”
“a photo of a [V] face, with sunglasses”	“a photo of a [V] face, with a golden tan”
“a photo of a [V] face, with thick glasses”	
“a photo of a [V] face, with a beanie”	
“a photo of a [V] face, with a wider smile”	
“a photo of a [V] face, with a headscarf”	
“a photo of a [V] face, with tribal face paint”	
“a photo of a [V] face, with a double chin”	

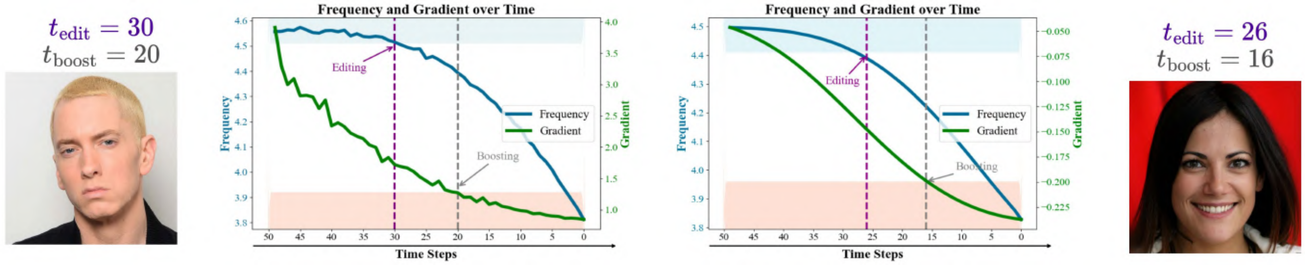

Figure 2. Stage settings and λ_t^{init} of different identities.

Overall, DreamSalon advances the field of personalized image editing, establishing a benchmark for future developments.

2.4. Editing Consistency Across Multiple Images of a Single Identity

Our approach streamlines personalization by initially training a HyperNetwork for an identity in approximately two minutes. Unique weights of an image are then generated by the HyperNetwork and employed for fast identity-context editing across various images of the same identity, taking just 26 seconds per image. The adaptability of editing various images of the single identity without re-training the HyperNetwork is crucial for fast personalization models. Fig. 9 presents edits on various images of a single identity, without re-training the HyperNetwork, emphasizing our method’s sensitivity to individual traits.

3. More Ablation Studies

3.1. Stage Configurations of Different Identities

Analyzing the frequency and initial lambda (λ_t^{init}) changes over time for different identities, as illustrated in Fig. 2, we observe that the frequency change aligns with the λ_t^{init} variation. Both identities exhibit distinct patterns in their frequency and gradient shifts, marking the transition from the editing to the boosting stages. For the first identity, editing occurs until timestep 30 and then transitions into boosting at timestep 20. In contrast, the second identity starts boosting earlier at timestep 26 after editing which concludes at timestep 16. This variance highlights the unique dynamic editing pathways that can be adapted based on individual facial characteristics.

3.2. Different Configuration of Stages

This ablation study, depicted in Fig. 3, evaluates how different configurations of editing and boosting stages influence facial image manipulation. We compare the efficacy of DreamSalon’s editing (2nd column) against four variant configurations, focusing on the timing of aggressive editing and quality boosting.

Different Editing Stages: Aggressive editing in the later diffusion stage (3rd column, timesteps 30 to 0) fails to achieve the intended edits, supporting the theory in Appendix 1.1 that early-stage editing is more effective. Continuous aggressive editing across the process (4th column, timesteps 50 to 0) impacts context and images’ quality, un-

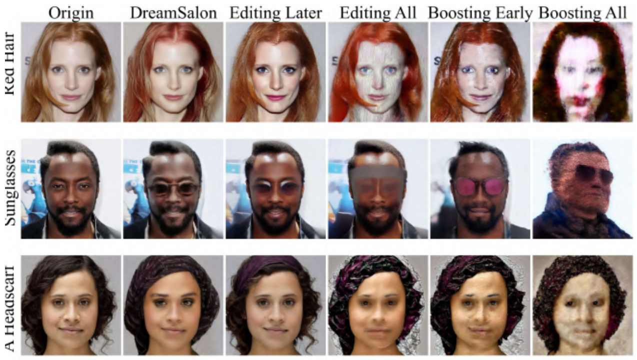

Figure 3. Impact of Stages Configurations in Facial Image Editing: From optimal identity-preserving manipulations (DreamSalon) to various degrees of editing intensity and quality boosting across the diffusion process.

derscoring that indiscriminate editing lacks the necessary subtlety and can lead to over-manipulation.

Different Boosting Stage: Early-stage quality boosting (5th column, timesteps 50 to 30) yields artifacts, indicating that such an approach is premature at a stage when the image lacks sufficient structure to benefit from the enhancement. Extensive boosting throughout the process (6th column, timesteps 50 to 0) leads to widespread artifacts and undesirable edits, correlating Appendix 1.2 that continuous introduction of stochastic noise can increase the risk of error accumulation, necessitating precise boosting timing.

3.3. Multiple Attributes in Target Prompts

In our examination of DreamSalon's adaptability to various textual prompts, we assess the impact of prompt complexity on editing outcomes, by maintaining consistent source prompts and progressively enriching the target prompts with additional attributes. As demonstrated in Fig. 4, the introduction of a single new attribute (2nd column) is effectively handled by DreamSalon, affirming its capacity for precise attribute manipulation. Furthermore, DreamSalon showcases its robustness by handling multiple attributes, whether they are independent (*e.g.*, glasses and a mustache) or exhibit correlation (*e.g.*, smiling and red lips). DreamSalon delivers compelling and coherent editing effects that capture the nuanced interplay of the combined attributes.

4. Implementation Details

The foundational concept of HyperDreambooth [4] involves the partitioning of the weight space of rank-1 LoRA residuals, introducing two novel hyperparameters: the down-rank a and the up-rank b. Consistent with the original study, we adopt the same hyperparameter settings for our experiments. HyperDreambooth innovates by utilizing a HyperNetwork to craft personalized weights, which serve as the initial attention weights for the pre-trained Latent Diffusion Model. The HyperNetwork's optimization is directed by the following loss function:

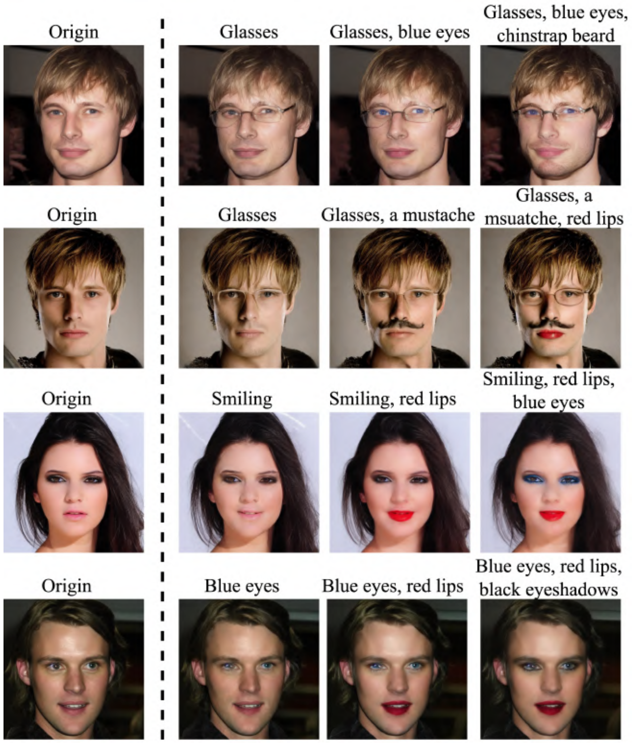
Origin
Glasses
Glasses, blue eyes
Glasses, blue eyes, chinstrap beard
Origin
Glasses
Glasses, a mustache
Glasses, a msuatche, red lips
Origin
Smiling
Smiling, red lips
Smiling, red lips, blue eyes
Origin
Blue eyes
Blue eyes, red lips
Blue eyes, red lips, black eyeshadows

Figure 4. Exploring the Robustness of DreamSalon to Complexity of Target Prompts: An evaluation of facial image edits with progressively complex prompts, demonstrating the method's proficiency in executing edits from simple single-attribute changes to more intricate multi-attribute transformations.

$$\mathcal{L}_{\text{HyperDreambooth}} = \alpha\mathbb{E}_{\epsilon,\mathbf{z},\mathbf{c}}[||\epsilon-\epsilon^{\theta}(\mathbf{z}_t,\mathbf{c})||_2^2]+||\hat{\theta}-\theta||_2^2, \quad (5)$$

where $\hat{\theta}$ are the pre-optimized weight parameters of the HyperNetwork, and α is the hyperparameter that modulates the balance between the two terms of the loss function. Following the original work, we set α to 0.1. By using 2~4 images of an identity, we obtain personalized weights θ_{per} from the HyperNetwork, enabling the generation of various customized images for that identity. In comparison to Dreambooth's requirement of approximately 10 minutes to fine-tune a Latent Diffusion Model and 8.6 GB of storage per identity, HyperDreambooth significantly reduces both the time to about 1.5 minutes and storage needs to 1.2 GB for each identity. Once the HyperNetwork is trained, editing across all images of the same identity can be accomplished in 26 seconds.

5. Limitations

One of the limitations acknowledged in our work relates to the method's dependency on the similarity between source and target prompts when utilizing differences in covariance matrices for editing. In line with practices from prior studies such as P2P [2], PnP [5], and DreamBooth [3], our approach assumes that users will typically modify attributes in the source prompts rather than make drastic changes. Hence, the covariance matrices primarily differ in these appended attributes, which our method is designed to capitalize on for effective editing. We recognize that our method may not be fully equipped to handle cases where source and target prompts are dissimilar – a scenario that has not been extensively explored in the literature and presents a new avenue for future research.

Additionally, we observed the increased brightness in the generated images and found they appear to be linked to the signal-to-noise ratio within the images. This finding suggests that the observed brightness variation is an emergent property of the editing mechanism rather than a byproduct of the boosting stage, pointing to the complex nature of visual attribute manipulations in image generation models.

References

[1] Rafael C Gonzales and Paul Wintz. *Digital image processing*. Addison-Wesley Longman Publishing Co., Inc., 1987. 1

[2] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. 5

[3] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. *arXiv preprint arXiv:2208.12242*, 2022. 5

[4] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Wei Wei, Tingbo Hou, Yael Pritch, Neal Wadhwa, Michael Rubinstein, and Kfir Aberman. Hyperdreambooth: Hypernetworks for fast personalization of text-to-image models. *arXiv preprint arXiv:2307.06949*, 2023. 4

[5] 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. 5

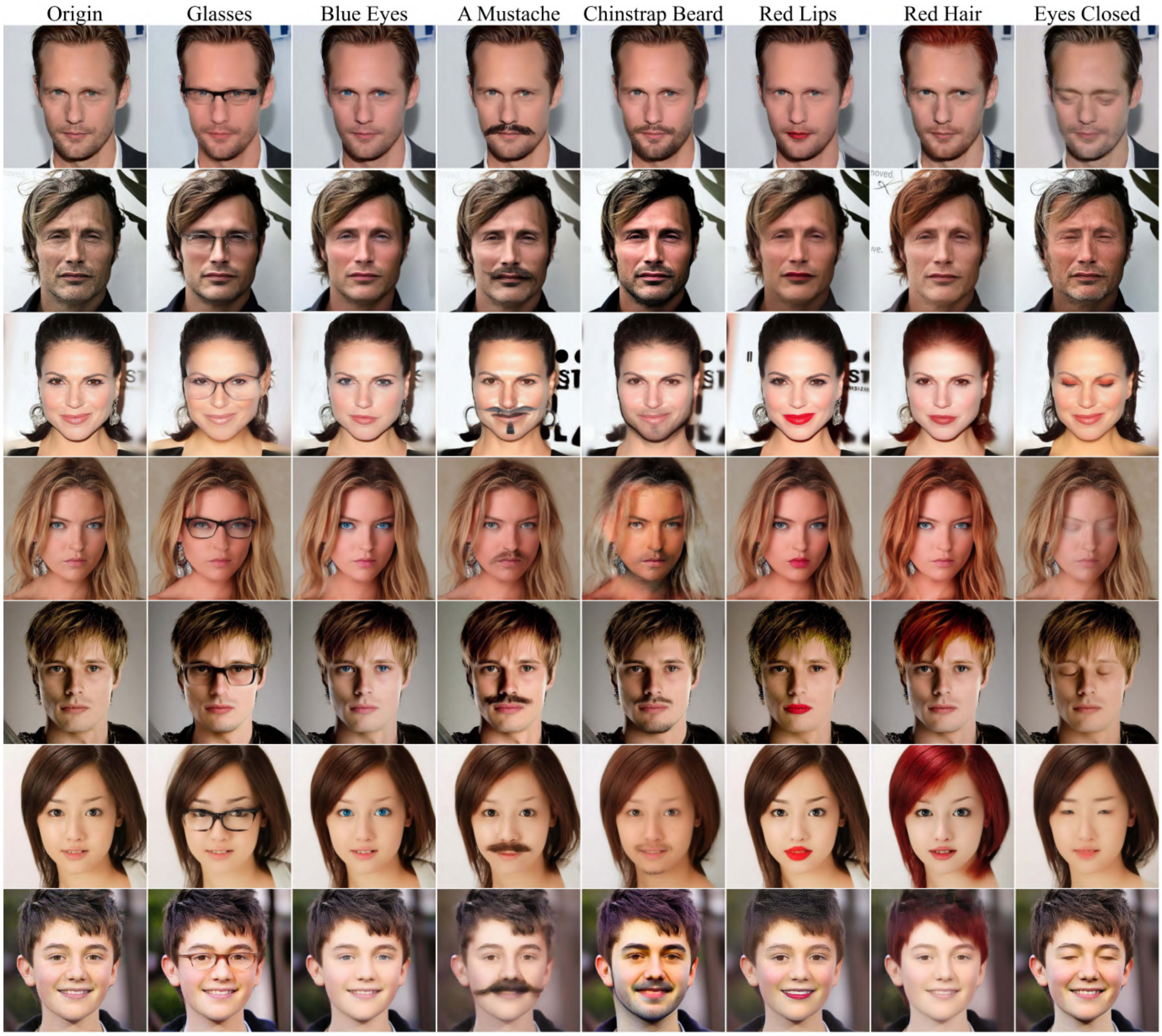

Figure 5. Comprehensive comparison of identity-preserved face editing across various methods using our approach, showcasing the consistent retention of identity and contextual elements in response to a diverse set of prompts.

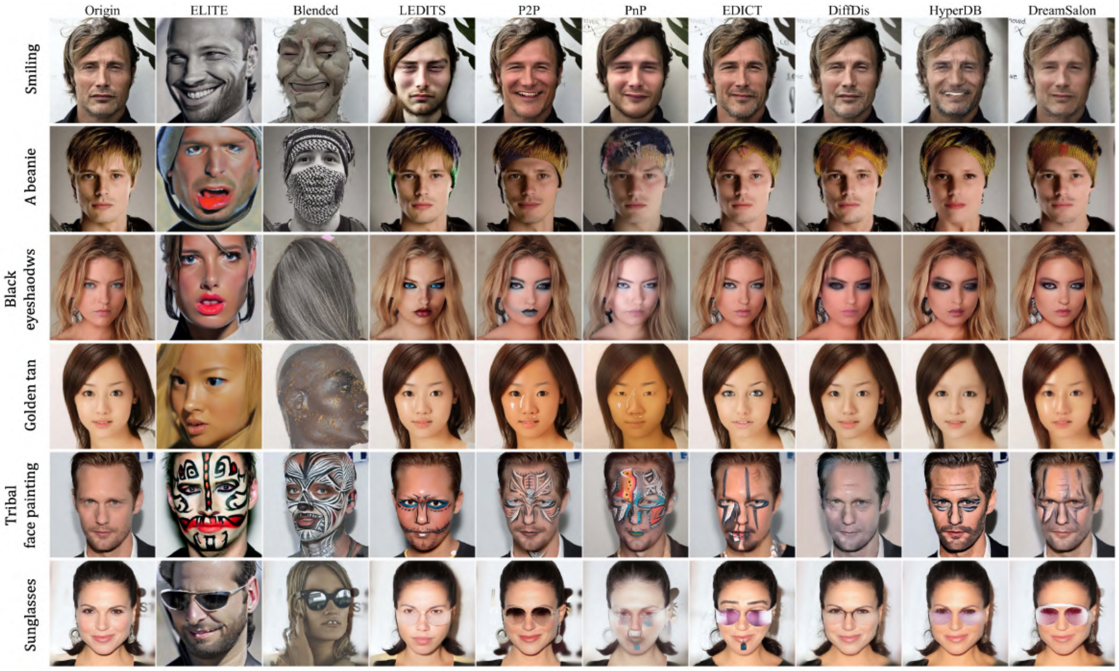

Figure 6. Evaluation of our method's editing performance on a consistent set of identities with the introduction of alternative prompts, demonstrating the method's robustness and fidelity in identity and context preservation.

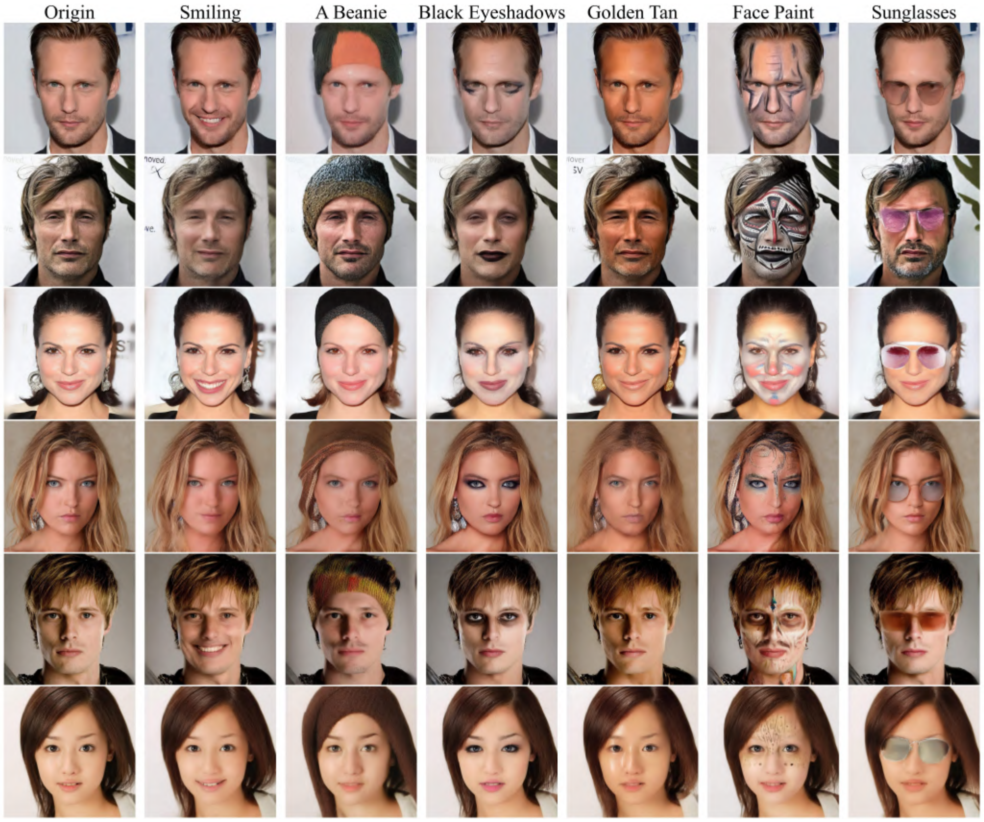

Figure 7. Visual assortment of edited facial images, detailing the interactions of different identities with various prompts, emphasizing the precision and adaptability of our editing technique.

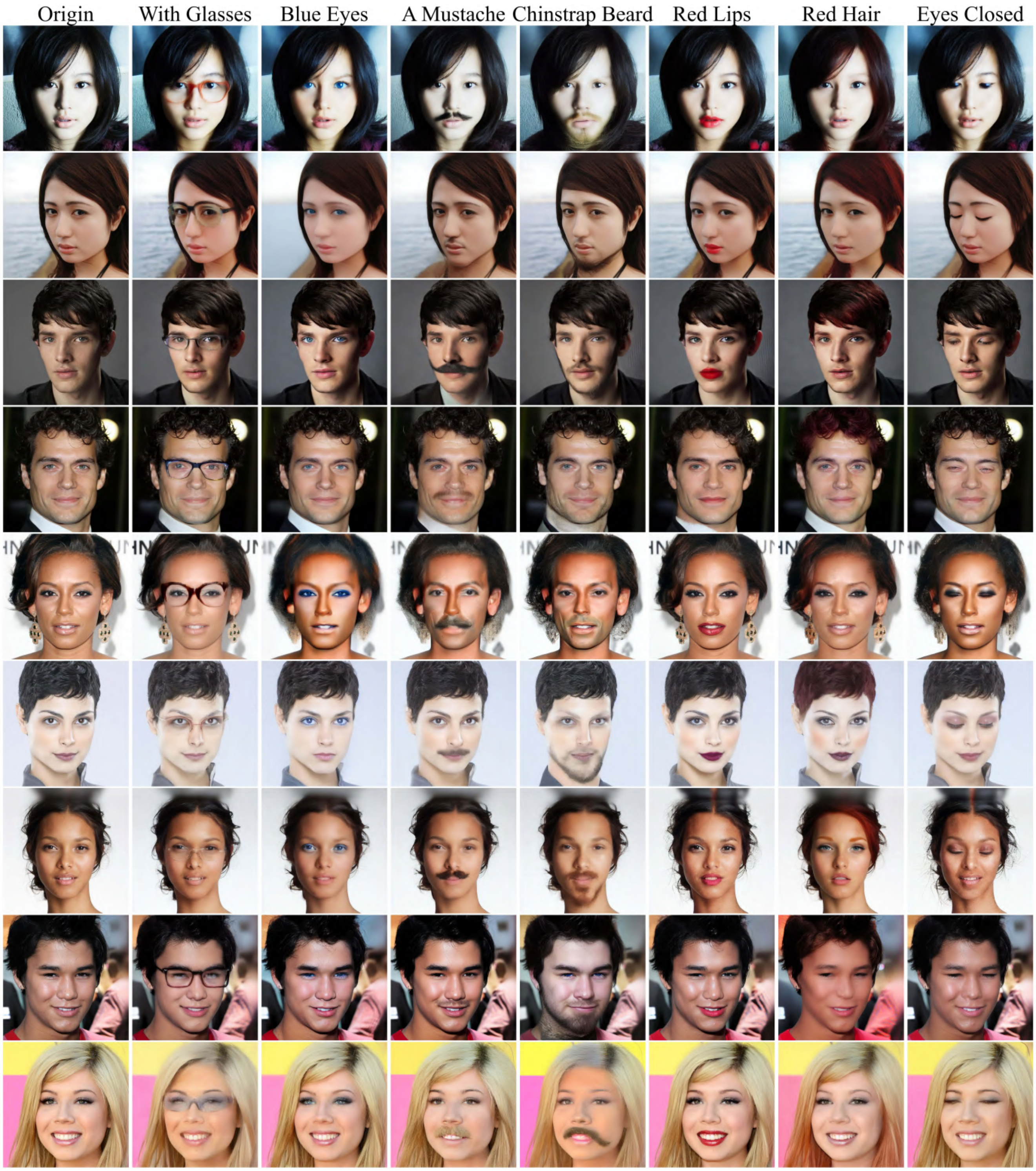

Figure 8. Display of editing outcomes on new identities using the main prompts from the study, underscoring our method's capacity to extend its personalization and editing prowess across a broader identity spectrum.

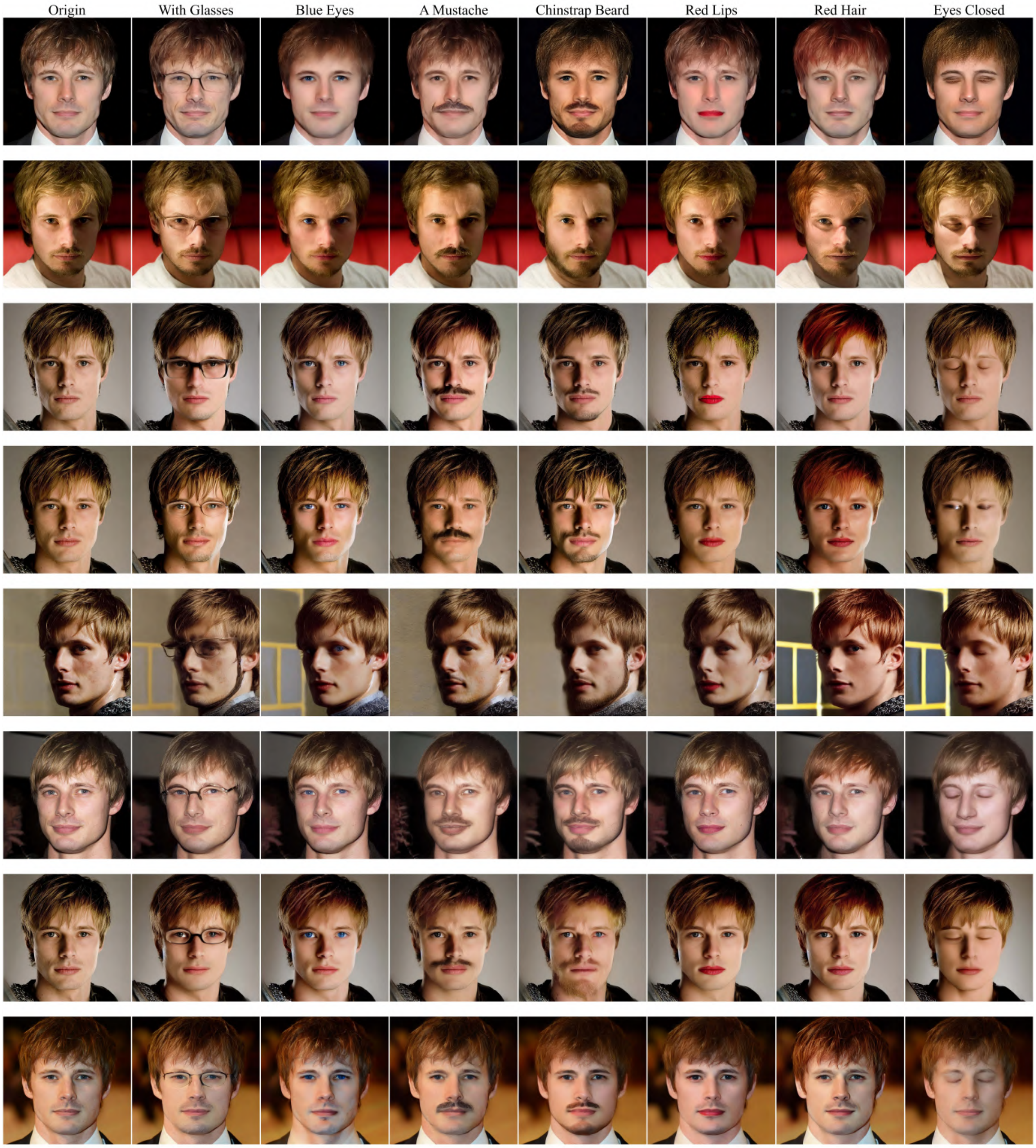

Figure 9. Exhibition of our method's personalized editing capabilities applied to multiple facial images of the same identity, each reflecting our approach's nuanced understanding of individual facial characteristics.