

Latent-Identity Tuning in Text-to-Image Personalization Models

Daniel Garibi¹ Ronen Kamenetsky¹ Hadar Averbuch-Elor² Daniel Cohen-Or¹ Or Patashnik¹

¹Tel Aviv University

²Cornell University

<https://garibida.github.io/IdentityTuning/>

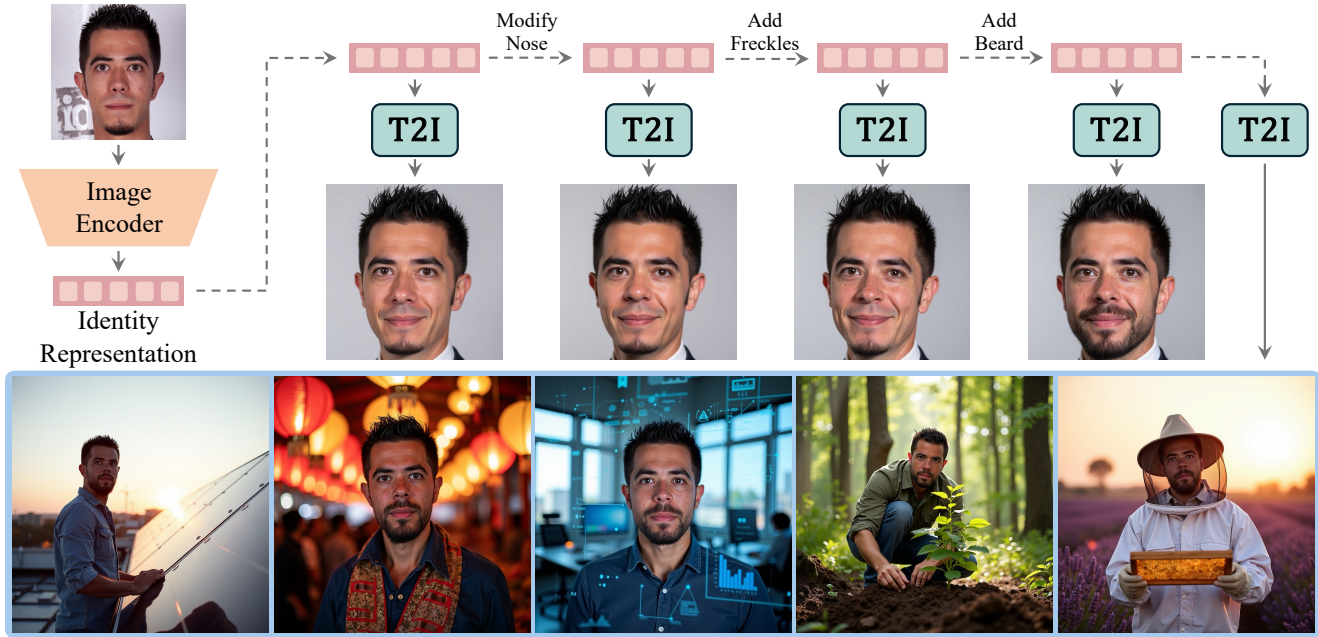


Figure 1. We present methods for directly tuning the identity tokens of a personalization encoder, enabling fine-grained control of facial attributes, for example modifying the nose, adding freckles, or a beard (top). The edited identity can then be used across diverse prompts to generate the same tuned subject consistently in new scenes (bottom).

Abstract

Generating and editing a person’s face demands high precision, as even minor modifications can significantly alter a subject’s perceived identity. Current personalization and editing methods built on general-purpose text-to-image models, however, often lack the precision required for fine-grained facial edits. We present a method for fine-grained identity tuning in text-to-image personalization models. Unlike standard image editing, which operates on a given image, identity tuning modifies the latent representation of a specific identity, enabling the generation of diverse images that consistently depict the same edited identity. To enable fine-grained latent identity tuning, we explore the latent space of a pre-trained, frozen encoder for text-to-image personalization. Our approach requires no additional training. Instead, it leverages the existing architecture of a frozen encoder to uncover latent semantic directions. This space consists of a set of latent tokens that play distinct roles in capturing different aspects of an identity and often correspond to specific spatial or semantic facial

regions. We show that meaningful directions can be identified within this space and within subspaces defined by selected tokens, enabling localized, fine-grained, and semantically coherent edits. We validate our approach through qualitative and quantitative experiments that demonstrate diverse localized facial edits while preserving cross-image identity consistency.

1. Introduction

Advances in text-to-image generation [33, 46, 61, 67] have enabled highly personalized synthesis, where models can generate new images of a person from a single or a few reference images. Such personalization methods allow the depiction of an individual in various contexts that vary in background, style, and pose [19, 40, 63, 77]. However, while existing methods faithfully reproduce identity, they offer little control over modifying it, limiting users’ ability to refine, customize, or creatively reinterpret how a given identity is depicted. For instance, a person may wish to appear with a beard, freckles, or a modified nose, requiring

this edited appearance to remain consistent across all subsequent generations (Fig. 1).

We define identity tuning as the task of modifying the latent representation that encodes a person’s identity in a pre-trained text-to-image personalization model. In contrast to image editing, which manipulates individual images, identity tuning operates on the latent identity representation itself. This ensures that the modified attributes remain consistent across all generated samples, establishing a persistent, altered identity that can be reliably deployed across varied prompts and settings [17, 52]. As this task involves human faces, it requires exceptional precision, as even minute differences can alter the perception of the identity. Achieving this level of precision requires enabling localized and continuous control over how the identity is portrayed or intended to appear.

Existing personalization methods typically learn an embedding or a set of latent tokens that capture a person’s identity from one or more reference images [29, 59, 77]. While these representations are primarily used to reproduce identity during generation, their internal structure has remained largely unexplored. In this work, we analyze the identity latent space of a personalization encoder, and show that it contains meaningful editing directions that can be explicitly identified. These directions enable fine-grained and semantically coherent tuning of a person’s identity across generated samples.

Specifically, the identity latent space consists of a set of latent tokens that play distinct roles in representing different generative aspects of facial identity, with some corresponding to specific spatial or semantic regions of the face [42, 59], as illustrated in Fig. 2. To enable control within this space, we investigate how manipulating individual identity tokens or all tokens jointly affects the encoded identity, and introduce a method to identify the tokens most relevant to a given localized facial attribute. We then apply several supervised and unsupervised techniques to the selected tokens to discover semantic directions in the latent space. These directions enable targeted and continuous fine-grained edits, ranging from localized feature modifications to smooth interpolations between identities.

We demonstrate the effectiveness of our framework across a range of identity tuning scenarios and different personalization models. In particular, we show that our approach enables fine-grained and semantically meaningful edits that are difficult to achieve through alternative text-guided methods, such as controlling the size or openness of the eyes, adding a specific beard, or transferring eyebrows from one identity to another. Moreover, by producing consistent identity representations across diverse images, our framework allows users to adjust an identity once and obtain results that align with their preferred appearance.

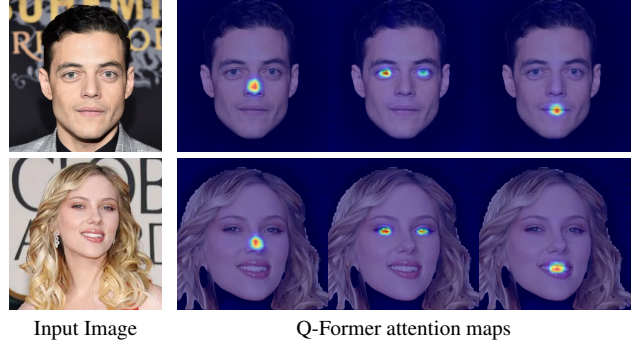


Figure 2. Token behavior in the face encoder. Q-Former attention maps for three learned query tokens reveal localized semantics: one focuses on the eyes, one on the nose, and one on the lips. The same token keeps the same role across both identities, showing consistent token-to-region mapping.

2. Related Work

Text-to-Image Personalization Given an image or a small set of images depicting a single concept, text-to-image personalization enables generating new images of that concept, guided by a text prompt that determines style, composition, and interactions with other subjects. Early works achieved this through per-subject optimization [4, 7, 8, 19, 25, 38, 58, 62, 69, 72]. Later works trained encoders that directly map the input images to concept embeddings, which are then used to condition the denoising process [6, 11, 20, 39–41, 48, 63, 68, 77].

Personalization of human faces has been a central focus [21, 29, 43, 57, 59, 73, 74], as it enables generating personalized depictions of individuals but also introduces unique challenges. Faithfully representing identity requires learning fine-grained details to which humans are highly sensitive, and users often find that personalized generations do not fully align with how they perceive themselves or wish to appear. We introduce a method for localized, continuous tuning of the identity representation, enabling fine-grained refinement or modification of a specific identity.

Image Editing Recent advances in generative models have substantially improved image editing capabilities. Early works showed that the latent spaces of GANs [28, 36] are organized in a semantically meaningful and disentangled manner, enabling controlled manipulation of diverse semantic attributes across images [1, 2, 31, 45, 55, 60, 66]. While many GAN-based editing methods explored human faces, their latent representations correspond to individual images rather than persistent identities, making them more suitable for single-image edits such as expressions or poses. Text-to-image diffusion models further broadened editing capabilities, enabling manipulation across diverse domains and direct control through text prompts [3, 10, 13, 15, 18, 24, 26, 32, 34, 37, 40, 49, 50, 54, 56, 71, 76]. Although these methods support intuitive natural-language interfaces and powerful manipulations, achieving fine-grained, local-

ized, and continuous control remains challenging and often requires specialized approaches [12, 22, 23, 27, 35, 53, 64, 75, 78]. In particular, controlling identity-related attributes through textual interfaces remains a challenge.

Identity Tuning Compared to image editing, identity editing has received limited attention. Weights2Weights [17] explores the weight space of LoRA-based personalized models, showing that it can be used to invert images, sample new identities, and edit existing ones. However, it requires per-image optimization to obtain the identity representation. In contrast, our method builds on a personalization encoder and operates without any optimization. PreciseControl [52] uses a GAN encoder [70] for personalization alongside a text-to-image model, performing per-image optimization to improve expressiveness and identity preservation. By aligning the GAN latent space with the personalization space, it enables GAN-based edits within text-to-image personalization. However, this approach is constrained by the editability of the chosen GAN and incurs significant cost due to the per-image optimization. IP-Composer [16] also explores a pretrained personalization encoder, leveraging CLIP image-text alignment for semantic composition. However, it does not address fine-grained facial editing or discover interpretable directions in the identity space.

3. Preliminaries: Q-Formers

Introduced under different names, specifically as Querying Transformers (Q-Formers) in BLIP-2 [42] and the Perceiver Resampler in Flamingo [5], these architectures act as intermediate Transformer modules designed to adapt a source token stream for a downstream model. For clarity, we adopt the Q-Former terminology in this work. A Q-Former maintains a small set of learnable query tokens that, via cross-attention, attend to a source sequence (e.g., tokens from a frozen image encoder) and produce a fixed number of output tokens. Each output token summarizes the information most relevant to its learned query, yielding a compact and structured representation while leaving the source model frozen.

In text-to-image personalization, given one or more images of a subject $\{I_i\}_{i=1}^k$, the goal is to extract a representation $Z \in \mathbb{R}^{N \times C}$, where N and C denote the number of tokens and hidden dimension of each token, that captures the subject’s identity and can condition the diffusion model during generation. IP-Adapter [77] follows this approach by first extracting features $F \in \mathbb{R}^{HW \times d}$ with a pretrained image encoder, mapping them to a set of tokens Z through an adapter network, and injecting them into the diffusion model via decoupled cross-attention layers. Q-Formers naturally serve as adapters for this task, using learnable queries $Q \in \mathbb{R}^{N \times C'}$ to distill information from the input image(s) into a fixed-length set of tokens $Z \in \mathbb{R}^{N \times C}$. Later



Figure 3. Identity interpolation. The outer columns show two identities (A and B), where intermediate columns are generated by linearly interpolating between their identity embeddings and applying the mixed embedding to all tokens.

works [29, 30, 57, 59, 74] extends this framework to face-based personalization by introducing a backbone encoder tailored for human identity and employing a Q-Former as the adapter. In the case of human personalization, we refer to the Q-Former’s output tokens Z as identity tokens.

Previous face-based personalization methods employing a Q-Former as the adapter analyzed the internal structure of the identity tokens and observed that different learned queries attend to distinct facial regions, such as the eyes, mouth, ears, and eyeglasses [57, 59]. These findings suggest that Q-Former-based adapters learn a structured and partially disentangled latent space representing human facial semantics. In this paper, we further characterize the organization of the identity tokens space and show that it enables fine grained, localized identity tuning.

4. Identity Tuning

In this section, we analyze the identity token space and introduce several methods for tuning the identity representation $Z \in \mathbb{R}^{N \times C}$. We observe that the manipulation of the identity representation can be applied to a single token $Z_n \in \mathbb{R}^{1 \times C}$, to all identity tokens Z , or to a selected subset of them $Z_S \in \mathbb{R}^{|S| \times C}$, where $S \subseteq \{1, \dots, N\}$. To identify the subset of tokens S most relevant to a given target attribute, we propose a dedicated token selection method. We then present several latent identity tuning techniques and demonstrate each of them using different token selections. An overview of our approach is shown in Fig. 4.

4.1. Global Identity Interpolation

We begin by exploring controlled interpolation between two identities. Let Z^A and Z^B denote the identity representations of two subjects, and let z^A and z^B be the corresponding variables to be interpolated. We construct an interpolated representation:

$$z^{blend} = (1 - \beta) z^A + \beta z^B, \quad (1)$$

where $0 < \beta < 1$. Setting $z = Z$ (global interpolation) produces a smooth transition of the identity from subject A toward subject B , as illustrated in Fig. 3. Since the personalization encoder is not fully disentangled between identity

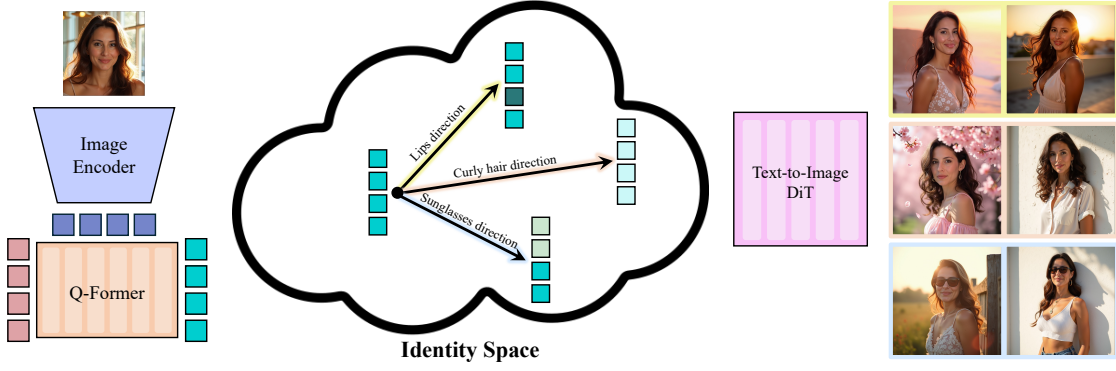


Figure 4. Overview of our identity tuning framework. An input image is encoded into a set of identity tokens, each capturing distinct aspects of the face. These tokens reside in an identity space, where meaningful directions (e.g., for eyes, age, or beard) can be discovered and traversed to achieve targeted edits. The modified identity tokens are then passed to a text-to-image model, which generates images of the edited identity in diverse contexts.

and image composition, changes in layout or pose may appear across the endpoints. However, our goal often requires fine-grained control over specific facial parts, such as interpolating only the eyes, rather than the entire face.

4.2. Token Selection

To isolate specific facial regions for editing, we must identify the relevant subset of tokens S . A straightforward approach to tuning the identity representation Z is to treat it as a single high-dimensional vector, $\text{vec}(Z) \in \mathbb{R}^{N \times C}$, and perform edits directly in this space, as demonstrated in the previous section with global interpolation. However, as shown earlier in Fig. 2, individual tokens typically correspond to semantically localized regions in the image. Therefore, as we will demonstrate later, treating all tokens as a single vector and manipulating them jointly is generally effective for global identity edits, but not for localized edits. Another option is to manipulate individual tokens $Z_n \in \mathbb{R}^{1 \times C}$, which can be regarded as vectors in $\mathbb{R}^C$. This results in highly localized changes that, as we will show, are often too limited for most tuning objectives. The third option we explore is to manipulate a subset of tokens, $S \subseteq \{1, \dots, N\}$, represented as $\text{vec}(Z_S) \in \mathbb{R}^{|S| \times C}$. Next, we describe two methods for identifying S given a target facial region to edit.

Localized Attributes In the first method, we construct paired images that differ primarily in one facial part, and then measure the differences in their corresponding identity representations. We construct these pairs by applying simple, pixel-based localized image manipulations. Specifically, given a facial part we aim to edit (e.g., lips) and a face image I , we crudely paste a patch (e.g., lips from another image) onto the corresponding region, resulting in a modified image I_{patch} . Examples of (I, I_{patch}) pairs are shown in Fig. 5. For each pair (I, I_{patch}) that differs in a specific region, we compute the difference: $\Delta Z = Z(I_{\text{patch}}) - Z(I) \in \mathbb{R}^{N \times C}$. We quantify the contribution of each token n to the given attribute by the magnitude of its difference $\|\Delta Z_n\|_2$, and construct a subset S by selecting the top k tokens for that pair. By aggregating just a small number of pairs (e.g.,



Figure 5. We form I_{patched} by pasting a new facial region onto I . Comparing their encoded identity representations reveals the tokens capturing the patched region’s influence.

five) corresponding to the same facial part, we obtain a stable subset S containing the tokens that most frequently appear among the top k across all pairs. This process is performed once per facial region and is reused for different edits of that region across all subjects (the same subset S is applied to any new identity). By directly manipulating specific facial regions, we enable control over parts of the face that are difficult to specify through text.

Global Attributes Beyond tokens that attend to localized facial regions, some tokens encode global appearance attributes (e.g., skin tone, freckles) that span multiple regions. These tokens cannot be isolated with localized patch edits, as modifying such attributes affects large portions of the face. To identify such tokens, we leverage labeled data for specific global attributes. Given a global attribute a and a labeled dataset $\{(I_i, y_i^a)\}_{i=1}^M$, we encode each image to the identity space $Z(I_i) \in \mathbb{R}^{N \times C}$. For each token index $n \in \{1, \dots, N\}$, we train a linear SVM $f_{n,a}$ on the per-token features $\{Z_n(I_i)\}$ to predict the attribute label y_i^a . Let $m_{n,a}$ denote the validation accuracy of $f_{n,a}$. We then select the subset of token indices whose classifiers achieve validation accuracy above a threshold τ : $A_a = \{n \mid m_{n,a} \geq \tau\}$. Intuitively, if a classifier trained on a single token reliably predicts attribute a , that token encodes information about a and can thus be used to control it. This selection is performed once per attribute and reused for downstream edits.

4.3. Localized Identity Transfer

Having defined a method to isolate relevant tokens, we can now perform localized identity manipulations. Setting $z = Z_S$ in Eq. (1) (region transfer) localizes the transition



Figure 6. Blending identities by part-level transfer. Our approach enables the transfer of a chosen facial component from a source identity to a target. In each grid, the target identity on the left receives a specific part from the source shown at the top. The left grid demonstrates periocular (eyelids+eyebrows) transfer, and the right grid demonstrates lip transfer. All other attributes and the background are preserved.

to the facial part corresponding to subset S . We achieve this by replacing the tokens in Z^A corresponding to the desired facial region with z^{blend} . This keeps the subject aligned with A outside the selected region while gradually adopting the corresponding trait from B within it. Fig. 6 presents results of transplanting the periocular region (eyebrows and eyelets) from different identities.

4.4. Linear Token Manipulation

Beyond interpolation, we introduce a set of methods for manipulating the identity tokens via linear directions in the latent space. To apply these linear operations, we treat the target representation as a flattened vector. Given an image I , we denote by $z(I) \in \mathbb{R}^M$ its corresponding identity representation vector to be edited. Depending on the desired level of granularity, z can represent the entire identity $\text{vec}(Z)$ ($M = N \cdot C$), an individual token Z_n ($M = C$), or a subset of tokens $\text{vec}(Z_S)$ ($M = |S| \cdot C$).

4.4.1. Supervised Linear Directions

We first learn supervised directions associated with specific facial attributes. As before, we begin by extracting identity representations from the images in our dataset, resulting in the set $\{z(I)\}_{I \in \mathcal{D}}$. Given binary attribute labels $y(I) \in \{-1, +1\}$ (either from annotated datasets or predicted by an off-the-shelf classifier), we estimate a linear direction in the latent space that corresponds to the given attribute using two complementary ways.

First, we use a mean-difference estimator. We define μ_+ and μ_- as the mean identity representations of the positive and negative samples, $\mu_+ = 1/|\mathcal{D}_+| \sum_{I \in \mathcal{D}_+} z(I)$, $\mu_- =$

$1/|\mathcal{D}_-| \sum_{I \in \mathcal{D}_-} z(I)$, where $\mathcal{D}_+ = \{I \in \mathcal{D} \mid y(I) = 1\}$ and $\mathcal{D}_-$ is defined similarly. The editing direction is then: $d_{\text{attr}} = \mu_+ - \mu_-$.

Second, we train a linear SVM on $\{z(I), y(I)\}$ to obtain a weight vector w and bias b . The learned hyperplane $w^\top x + b = 0$ separates images with and without the attribute. This formulation is inspired by GAN-based latent space editing methods, where linear separators are learned for attributes and the hyperplane normal is used as an edit direction [66]. We use the unit-norm normal of this hyperplane as an edit direction in the token space, $v_{\text{svm}} = \frac{w}{\|w\|_2} \in \mathbb{R}^M$, where w is the hyperplane normal, and choose its sign so that a small positive step increases the classifier score on a held out set.

Given this direction and a chosen scale, we apply edits in a norm-preserving manner to approximately preserve distributional scale. Let $z \in \mathbb{R}^M$ be the vector to edit, $\bar{z} = \frac{1}{|\mathcal{D}|} \sum_{I \in \mathcal{D}} z(I)$ its dataset mean, and v a unit-normalized direction. For a scale $\alpha \in \mathbb{R}$, we define:

$$e = (z - \bar{z}) + \alpha \cdot v, \quad z' = \bar{z} + e \cdot |z - \bar{z}|_2 / |e|_2. \quad (2)$$

This operation centers the vector, adds a scaled direction, and rescales to preserve the original distance from the mean.

Fig. 7 presents qualitative results obtained using both supervised approaches, the mean-difference and SVM-based estimators. We observe that the SVM-based method is more effective for fine-grained edits, such as the rosy cheeks attribute, whereas the mean-difference estimator performs better for more pronounced modifications, such as bald or beard. This difference can be explained by the nature of the

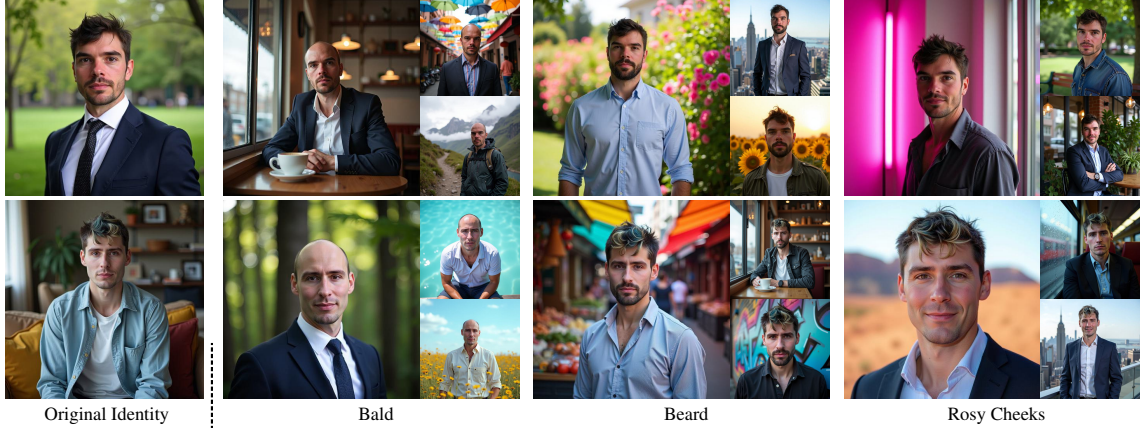


Figure 7. Each row shows a different source identity (leftmost column) and its edits along three directions: bald, beard, and rosy cheeks. The images in each column are generated by adding the same direction found by SVM. For each edit, we show the main result with two alternative prompts stacked vertically to the right.

learned directions: for substantial changes, the SVM classifier may rely on a subset of highly discriminative dimensions rather than all dimensions relevant to the attribute, effectively learning shortcuts that do not generalize to strong global edits. In contrast, the mean-difference estimator captures a broader, more global direction but lacks the precision required for subtle, fine-grained attributes.

4.4.2. Unsupervised Directions

In addition to supervised directions, we also explore unsupervised manipulation of identity tokens by identifying linear directions in the identity space. Traversing along such directions enables fine-grained control over the strength of the edit. Following prior works that discover meaningful edit directions in GAN latent spaces using Principal Component Analysis (PCA) [31, 65], we apply a similar approach to our identity representation.

Specifically, given a dataset $\mathcal{D}$ of images, we extract for each image $I \in \mathcal{D}$ its identity representation $z(I)$. We then perform PCA on the set of vectors $\{z(I)\}_{I \in \mathcal{D}}$, obtaining a set of principal directions. Given a PCA direction v , we apply the same norm-preserving update described in Eq. (2).

We show qualitative results of this method in Fig. 8. Setting $z = \text{vec}(Z)$ (Global PCA) captures correlations across tokens and produces coherent global edits, for example smooth changes in apparent age, but the resulting directions often entangle multiple facial parts and therefore provide limited locality and limited fine-grained control.

Setting $z = Z_n$ (Per token PCA) often reveals tokens that specialize to particular facial regions and thus provides locality and interpretability, e.g., a token that controls eye color. However, controlling most individual tokens has only a minor effect, suggesting information is distributed across several tokens and it is not sufficient to control only one.

By setting $z = \text{vec}(Z_S)$ and computing S using our token selection method for the eye region, we achieve smooth and localized edits of the eyes, enabling control over their

openness. Similarly, by selecting the tokens relevant to the lips, we can control their color.

5. Experiments

In this work, we utilize Omni-ID [59] and PuLID [29] as encoders for the Flux.dev [9] text-to-image model. This entire framework is built upon the IP-Adapter architecture [77]. To compute PCA over identity tokens, we use the FFHQ dataset [36], which contains 70,000 high-quality face images. For supervised attribute directions, we use the CelebA dataset [47], which provides 202,599 images annotated with 40 binary attribute labels. Additional implementation details in Sec. A.

5.1. Qualitative Results

We present qualitative results in Figures 1, 6, 7, and 8. Fig. 1 illustrates identity tuning: given a source image, our approach enables iterative tuning of the identity, highlighting the advantage of operating directly on the identity representation. The tuned identity produces consistent generations across diverse prompts, and the figure includes edits that are difficult to express via text, such as modifying the shape of the nose. Fig. 7 presents supervised attribute-aligned directions, which similarly transfer across different identities. more results in Sec. B.1. Fig. 6 illustrates identity tuning by injecting facial parts from another person. Identity tuning applies an edit direction in the identity latent space, scaled by a user-selected intensity. See Sec. B.2 and Sec. B.4 for continuous tuning and further applications of patch editing beyond paired-data creation, such as inserting specific visual elements like a beard style.

Ablation on Token Granularity Fig. 8 evaluates edit granularity across three control levels: a single token, a targeted token subset, and the global token space. Manipulating a single token yields a localized but minimal effect, useful mainly for isolated changes like eye color. Conversely, applying directions globally produces pronounced changes

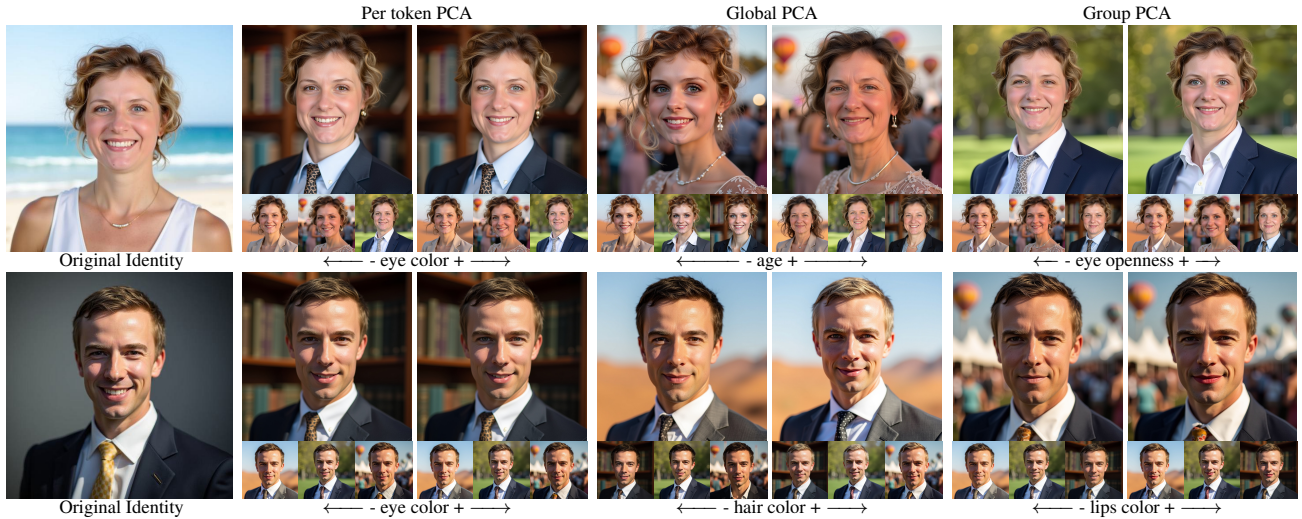


Figure 8. PCA-based controls across three regimes. Depending on the section, we vary a single principal direction for a single token, across all tokens, or within a targeted token group. The arrow labels name the observed attribute (eye color, age, eye openness, lip color). For every direction we display identity edits from negative to positive, and for each edited setting we render four prompts to show that the subject identity stays consistent while only the targeted attribute changes.

but loses fine-grained control over specific facial regions. Thus, most attributes require modifying a selected token subset. This demonstrates that tailoring token granularity to the target attribute is essential for precise, localized edits.

5.2. Comparisons

Identity tuning is largely overlooked in the existing literature, meaning few methods are directly applicable to this task. The most relevant works are Precise Control [52] and Weights2Weights [17], for which we use the authors’ official implementations and provided editing directions. In addition, we compare against the state-of-the-art *image editing* model, Flux Kontext [40]. Since this model does not natively support consistent identity tuning, we adapt it in two ways. First (Kontext Direct), we input the person’s reference image and use a single prompt that combines both the desired modification (e.g., “add a beard”) and the target scene. Second, we use a two-step procedure (Kontext Sequential). We first apply the edit to the reference image, and then use that edited image as the new reference to generate the final result using the target prompt.

Our approach supports a broad range of localized and global identity edits, including patch-based changes, token directions, interpolations, and unsupervised edit directions that cannot be expressed in text. However, not all baselines support all of these edit types. To ensure a fair comparison, we evaluate each method only on the edits it can perform. Our identity tuning evaluation focuses on three aspects: (i) the ability of each method to perform the required edit, (ii) the identity consistency of the tuned identity across different generations, and (iii) the prompt adherence.

Qualitative Comparisons We begin with a comparison to methods that support identity-tuning. Fig. 9 presents results for PreciseControl and Weights2Weights (W2W). For each method, we generate four images under diverse



Figure 9. Qualitative comparison with identity-tuning methods. Each row shows an identity edit of one attribute, with four prompts as columns. Our approach consistently applies the target attribute across diverse prompts while preserving a coherent edited identity.

prompts to evaluate identity consistency across prompts. While these methods often apply the requested edit, they do not reliably preserve the same person across prompts. In addition, W2W frequently struggles to follow the input text prompt. Additional examples and results, including those using both Flux Kontext baselines, are provided in Sec. B.5. Since Flux Kontext supports a broader range of edits than the identity-tuning baselines, we also compare it on more challenging, fine-grained edits shown in Fig. 10. Both Kontext variants (Direct and Sequential) struggle to apply the modifications locally and often introduce global artifacts. In contrast, our method produces localized, semantically aligned changes while maintaining a coherent identity across prompts.



Figure 10. Qualitative comparison of fine-grained identity. While FluxKontext struggles with precise local control and overall consistency, our approach achieves accurate, fine-grained, region-specific identity edits with stronger consistency.

Quantitative Comparisons To evaluate identity consistency, edit adherence, and prompt adherence, we construct a custom benchmark. It spans basic edits and advanced edits requiring finer control (unsupported by PreciseControl and partially supported by W2W), yielding over 3,000 generated images per method. See Sec. B.5 for details. We evaluate along three axes. *Identity consistency*: For each identity-edit pair, we compute ArcFace [14] embeddings for the 13 generated images, measure their mean pairwise cosine similarity, and average over all pairs. *Edit adherence*: Following [44], a VQA-based judge scores adherence to the requested edit. *Prompt adherence*: Using the same protocol, we score alignment with the target prompt.

Tab. 1 reports results for both edit types. For Precise Control and Weights2Weights, which support identity tuning, our approach achieves substantially higher identity consistency, higher edit adherence, and higher prompt adherence. For Flux Kontext Direct, although edit adherence is comparable, our method attains significantly higher identity consistency. Flux Kontext Sequential achieves identity consistency close to ours but underperforms in edit adherence, especially on advanced edits requiring fine-grained control. Both baselines also obtain lower prompt-adherence scores. Finally, while the metrics indicate that Kontext adheres to the requested edits, in practice it introduces many artifacts that the metrics do not capture, as shown in Fig. 10.

User Study To complement our quantitative analysis, we conducted a user study comparing our method with baselines that support advanced edits. In a pairwise comparison, participants saw (i) a source image of a person, (ii) a textual edit instruction, and (iii) edited results from our method and a competing method. For each method, we displayed two

Table 1. Quantitative Comparison. We assess ID consistency, edit and prompt adherence for methods on basic and advanced edits.

	Basic Edits			Advanced Edits		
	ID Consis. ↑	Edit Adher. ↑	Prompt Adher. ↑	ID Consis. ↑	Edit Adher. ↑	Prompt Adher. ↑
PreciseControl	0.25	0.66	0.75	N/A	N/A	N/A
W2W	0.31	0.64	0.27	0.32	0.71	0.22
Kontext Direct	0.37	0.81	0.85	0.39	0.78	0.81
Kontext Seq.	0.51	0.74	0.82	0.49	0.71	0.80
Ours (OmniID)	0.47	<u>0.79</u>	0.91	0.47	0.78	0.89
Ours (PuLID)	<u>0.49</u>	0.81	<u>0.88</u>	<u>0.48</u>	<u>0.77</u>	<u>0.86</u>

Table 2. User Study. Pairwise win rate of our approach vs. others.

	W2W	Kont. Direct	Kont. Seq.
ID Preservation	78%	75%	70%
ID Coherent	85%	82%	76%
Edit Adher.	94%	78%	83%
Prompt Adher.	96%	88%	85%
Overall Prefer.	94%	85%	83%

results generated with different prompts. Participants provided pairwise preferences on five criteria: Identity Preservation, Identity Consistency, Edit Adherence, Prompt Adherence, and Overall Preference. See the question example in Sec. B.5. In total, we collected 250 pairwise judgments. As summarized in Tab. 2, our method was preferred over all baselines across all categories. These results indicate that users judged our approach to better preserve the original identity, maintain consistency across prompts, and adhere to both the specified edit and the target prompt. That indicates the superiority of our approach for the task of identity tuning.

6. Conclusions

In this work, we explored the latent identity space of a Q-Former-based personalization encoder and introduced a framework for fine-grained, localized, continuous identity tuning. We showed that the identity tokens form a structured, spatially meaningful space capturing distinct facial attributes. Leveraging this structure, we identified region-relevant tokens and extracted editing directions within their subspaces, enabling precise, region-aware manipulation.

While prior work focused on coarse identity preservation, our framework enables subtle, continuous adjustments, such as eye openness, freckle texture, and facial-hair density. Our results demonstrate that careful exploration of the identity space enable these nuanced edits while maintaining consistency across generated images.

Overall, our findings reveal that the identity token space is both interpretable and controllable at fine facial granularity, allowing users to refine delicate aspects of visual identity. Future work may pursue perceptually aligned metrics for subtle edits and interactive workflows that support precise user-guided adjustments.

Acknowledgments

We would like to thank Omer Dahari, Etai Sella, Shelly Golan, Sara Dorfman, Kfir Aberman, and Jackson Wang for their early feedback and insightful discussions. This research was supported in part by the Israel Science Foundation (grants no. 2492/20 and 1473/24), Len Blavatnik, and the Blavatnik Family Foundation. We also thank NVIDIA for their generous support through the NVIDIA Academic Grant Program, which provided GPU hours via Brev for this research.

References

- [1] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to embed images into the stylegan latent space? In *Proceedings of the IEEE international conference on computer vision*, pages 4432–4441, 2019. 2
- [2] Rameen Abdal, Peihao Zhu, Niloy Mitra, and Peter Wonka. Styleflow: Attribute-conditioned exploration of stylegan-generated images using conditional continuous normalizing flows, 2020. 2
- [3] Yuval Alaluf, Daniel Garibi, Or Patashnik, Hadar Averbuch-Elor, and Daniel Cohen-Or. Cross-image attention for zero-shot appearance transfer, 2023. 2
- [4] Yuval Alaluf, Elad Richardson, Gal Metzger, and Daniel Cohen-Or. A neural space-time representation for text-to-image personalization, 2023. 2
- [5] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning, 2022. 3
- [6] Moab Arar, Rinon Gal, Yuval Atzmon, Gal Chechik, Daniel Cohen-Or, Ariel Shamir, and Amit H. Bermano. Domain-agnostic tuning-encoder for fast personalization of text-to-image models. In *SIGGRAPH Asia 2023 Conference Papers*, New York, NY, USA, 2023. Association for Computing Machinery. 2
- [7] Moab Arar, Andrey Voynov, Amir Hertz, Omri Avrahami, Shlomi Fruchter, Yael Pritch, Daniel Cohen-Or, and Ariel Shamir. Palp: Prompt aligned personalization of text-to-image models. *arXiv preprint arXiv:2401.06105*, 2024. 2
- [8] Omri Avrahami, Kfir Aberman, Ohad Fried, Daniel Cohen-Or, and Dani Lischinski. Break-a-scene: Extracting multiple concepts from a single image. In *SIGGRAPH Asia 2023 Conference Papers*, page 1–12. ACM, 2023. 2
- [9] Black Forest Labs. Flux, <https://github.com/black-forest-labs/flux>, 2024. 6
- [10] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18392–18402, 2023. 2
- [11] Shengqu Cai, Eric Ryan Chan, Yunzhi Zhang, Leonidas Guibas, Jiajun Wu, and Gordon Wetzstein. Diffusion self-distillation for zero-shot customized image generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 18434–18443, 2025. 2
- [12] Omer Dahary, Benaya Koren, Daniel Garibi, and Daniel Cohen-Or. On-the-fly repulsion in the contextual space for rich diversity in diffusion transformers. *arXiv preprint arXiv:2603.28762*, 2026. 3
- [13] Yusuf Dalva, Kavana Venkatesh, and Pinar Yanardag. Fluxspace: Disentangled semantic editing in rectified flow models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 13083–13092, 2025. 2
- [14] Jiankang Deng, Jia Guo, Jing Yang, Niannan Xue, Irene Kotisa, 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. 8
- [15] Gilad Deutch, Rinon Gal, Daniel Garibi, Or Patashnik, and Daniel Cohen-Or. Turboedit: Text-based image editing using few-step diffusion models. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–12, 2024. 2
- [16] Sara Dorfman, Dana Cohen-Bar, Rinon Gal, and Daniel Cohen-Or. Ip-composer: Semantic composition of visual concepts. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pages 1–11, 2025. 3
- [17] Amil Dravid, Yossi Gandelsman, Kuan-Chieh Wang, Rameen Abdal, Gordon Wetzstein, Alexei A Efros, and Kfir Aberman. Interpreting the weight space of customized diffusion models. *Advances in Neural Information Processing Systems*, 37:137334–137371, 2024. 2, 3, 7
- [18] Dave Epstein, Allan Jabri, Ben Poole, Alexei Efros, and Aleksander Holynski. Diffusion self-guidance for controllable image generation. *Advances in Neural Information Processing Systems*, 36:16222–16239, 2023. 2
- [19] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion, 2022. 1, 2
- [20] Rinon Gal, Moab Arar, Yuval Atzmon, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. Encoder-based domain tuning for fast personalization of text-to-image models. *ACM Transactions on Graphics (TOG)*, 42(4):1–13, 2023. 2
- [21] Rinon Gal, Or Lichter, Elad Richardson, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. Lcm-lookahead for encoder-based text-to-image personalization. In *European Conference on Computer Vision*, pages 322–340. Springer, 2024. 2
- [22] Rohit Gandikota, Joanna Materzyńska, Tingrui Zhou, Antonio Torralba, and David Bau. Concept sliders: Lora adaptors for precise control in diffusion models. In *European Conference on Computer Vision*, pages 172–188. Springer, 2024. 3
- [23] Rohit Gandikota, Zongze Wu, Richard Zhang, David Bau, Eli Shechtman, and Nick Kolkin. Sliderspace: Decomposing the visual capabilities of diffusion models. In *Proceedings of*

- the *IEEE/CVF international conference on computer vision*, 2025. [arXiv:2502.01639](#). [3](#)
- [24] 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. [2](#)
 - [25] Daniel Garibi, Shahar Yadin, Roni Paiss, Omer Tov, Shiran Zada, Ariel Ephrat, Tomer Michaeli, Inbar Mosseri, and Tali Dekel. Tokenverse: Versatile multi-concept personalization in token modulation space. *ACM Transactions On Graphics (TOG)*, 44(4):1–11, 2025. [2](#)
 - [26] Songwei Ge, Taesung Park, Jun-Yan Zhu, and Jia-Bin Huang. Expressive text-to-image generation with rich text. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. [2](#)
 - [27] Shelly Golan, Michael Finkelson, Ariel Bereslavsky, Yotam Nitzan, and Or Patashnik. Paretoslider: Diffusion models post-training for continuous reward control, 2026. [3](#)
 - [28] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. [2](#)
 - [29] Zinan Guo, Yanze Wu, Chen Zhuowei, Peng Zhang, Qian He, et al. Pulid: Pure and lightning id customization via contrastive alignment. *Advances in neural information processing systems*, 37:36777–36804, 2024. [2](#), [3](#), [6](#), [14](#)
 - [30] Yue Han, Junwei Zhu, Keke He, Xu Chen, Yanhao Ge, Wei Li, Xiangtai Li, Jiangning Zhang, Chengjie Wang, and Yong Liu. Face-adapter for pre-trained diffusion models with fine-grained id and attribute control. In *European conference on computer vision*, pages 20–36. Springer, 2024. [3](#)
 - [31] Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. Ganspace: Discovering interpretable gan controls. *arXiv preprint arXiv:2004.02546*, 2020. [2](#), [6](#)
 - [32] 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. [2](#)
 - [33] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. [1](#)
 - [34] Inbar Huberman-Spiegelglas, Vladimir Kulikov, and Tomer Michaeli. An edit friendly ddpn noise space: Inversion and manipulations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12469–12478, 2024. [2](#)
 - [35] Ronen Kamenetsky, Sara Dorfman, Daniel Garibi, Roni Paiss, Or Patashnik, and Daniel Cohen-Or. Saedit: Token-level control for continuous image editing via sparse autoencoder. *arXiv preprint arXiv:2510.05081*, 2025. [3](#)
 - [36] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4401–4410, 2019. [2](#), [6](#)
 - [37] Vladimir Kulikov, Matan Kleiner, Inbar Huberman-Spiegelglas, and Tomer Michaeli. Flowedit: Inversion-free text-based editing using pre-trained flow models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19721–19730, 2025. [2](#)
 - [38] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1931–1941, 2023. [2](#)
 - [39] Nupur Kumari, Xi Yin, Jun-Yan Zhu, Ishan Misra, and Samaneh Azadi. Generating multi-image synthetic data for text-to-image customization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16524–16534, 2025. [2](#)
 - [40] Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space, 2025. [1](#), [2](#), [7](#), [15](#), [16](#)
 - [41] Dongxu Li, Junnan Li, and Steven Hoi. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. *Advances in Neural Information Processing Systems*, 36:30146–30166, 2023. [2](#)
 - [42] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. [2](#), [3](#)
 - [43] Zhen Li, Mingdeng Cao, Xintao Wang, Zhongang Qi, Ming-Ming Cheng, and Ying Shan. Photomaker: Customizing realistic human photos via stacked id embedding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8640–8650, 2024. [2](#)
 - [44] Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation. In *European Conference on Computer Vision*, pages 366–384. Springer, 2024. [8](#)
 - [45] Huan Ling, Karsten Kreis, Daiqing Li, Seung Wook Kim, Antonio Torralba, and Sanja Fidler. Editgan: High-precision semantic image editing. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. [2](#)
 - [46] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling, 2023. [1](#)
 - [47] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2015. [6](#)
 - [48] Jian Ma, Junhao Liang, Chen Chen, and Haonan Lu. Subject-diffusion: Open domain personalized text-to-image generation without test-time fine-tuning. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–12, 2024. [2](#)
 - [49] 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. *arXiv preprint arXiv:2108.01073*, 2021. [2](#)

- [50] 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. 2
- [51] OpenAI. Chatgpt (gpt-5). <https://chat.openai.com/>, 2025. 16
- [52] Rishubh Parihar, VS Sachidanand, Sabariswaran Mani, Tejan Karmali, and R Venkatesh Babu. Precisecontrol: Enhancing text-to-image diffusion models with fine-grained attribute control. In *European Conference on Computer Vision*, pages 469–487. Springer, 2024. 2, 3, 7
- [53] Rishubh Parihar, Or Patashnik, Daniil Ostashev, Venkatesh Babu Radhakrishnan, Daniel Cohen-Or, and Kuan-Chieh Jackson Wang. Kontinuous kontext: Continuous strength control for instruction-based image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 37929–37939, 2026. 3
- [54] Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. Zero-shot image-to-image translation. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Proceedings*. ACM, 2023. 2
- [55] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2085–2094, 2021. 2
- [56] Or Patashnik, Daniel Garibi, Idan Azuri, Hadar Averbuch-Elor, and Daniel Cohen-Or. Localizing object-level shape variations with text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 2
- [57] Or Patashnik, Rinon Gal, Daniil Ostashev, Sergey Tulyakov, Kfir Aberman, and Daniel Cohen-Or. Nested attention: Semantic-aware attention values for concept personalization, 2025. 2, 3
- [58] Ryan Po, Guandao Yang, Kfir Aberman, and Gordon Wetstein. Orthogonal adaptation for modular customization of diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7964–7973, 2024. 2
- [59] Guocheng Qian, Kuan-Chieh Wang, Or Patashnik, Negin Heravi, Daniil Ostashev, Sergey Tulyakov, Daniel Cohen-Or, and Kfir Aberman. Omni-id: Holistic identity representation designed for generative tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8786–8795, 2025. 2, 3, 6, 14
- [60] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: a stylegan encoder for image-to-image translation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2
- [61] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1
- [62] 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. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22500–22510, 2023. 2
- [63] 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. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6527–6536, 2024. 1, 2
- [64] Etai Sella, Yoav Baron, Hadar Averbuch-Elor, Daniel Cohen-Or, and Or Patashnik. Looserope: Content-aware attention manipulation for semantic harmonization, 2026. 3
- [65] Yujun Shen and Bolei Zhou. Closed-form factorization of latent semantics in gans. *arXiv preprint arXiv:2007.06600*, 2020. 6
- [66] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of gans for semantic face editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9243–9252, 2020. 2, 5
- [67] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2020. 1
- [68] Zhenxiong Tan, Songhua Liu, Xingyi Yang, Qiaochu Xue, and Xinchao Wang. Ominicontrol: Minimal and universal control for diffusion transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14940–14950, 2025. 2
- [69] Yoad Tewel, Rinon Gal, Gal Chechik, and Yuval Atzmon. Key-locked rank one editing for text-to-image personalization. In *ACM SIGGRAPH 2023 conference proceedings*, pages 1–11, 2023. 2
- [70] Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. Designing an encoder for stylegan image manipulation. *ACM Transactions on Graphics (TOG)*, 40(4): 1–14, 2021. 3
- [71] 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. 2
- [72] Andrey Voynov, Qinghao Chu, Daniel Cohen-Or, and Kfir Aberman. P+: Extended textual conditioning in text-to-image generation, 2023. 2
- [73] Kuan-Chieh Wang, Daniil Ostashev, Yuwei Fang, Sergey Tulyakov, and Kfir Aberman. Moa: Mixture-of-attention for subject-context disentanglement in personalized image generation. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–12, 2024. 2
- [74] Qixun Wang, Xu Bai, Haofan Wang, Zekui Qin, Anthony Chen, Huaxia Li, Xu Tang, and Yao Hu. Instantid: Zero-shot

- identity-preserving generation in seconds. *arXiv preprint arXiv:2401.07519*, 2024. [2](#), [3](#)
- [75] Alon Wolf, Chen Katzir, Kfir Aberman, and Or Patashnik. Continuous control of editing models via adaptive-origin guidance, 2026. [3](#)
- [76] Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025. [2](#)
- [77] Hu Ye, Jun Zhang, Sibio Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models, 2023. [1](#), [2](#), [3](#), [6](#)
- [78] Arman Zarei, Samyadeep Basu, Mobina Pournemat, Sayan Nag, Ryan A Rossi, and Soheil Feizi. Slidedit: Continuous image editing with fine-grained instruction control. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14430–14439, 2026. [3](#)

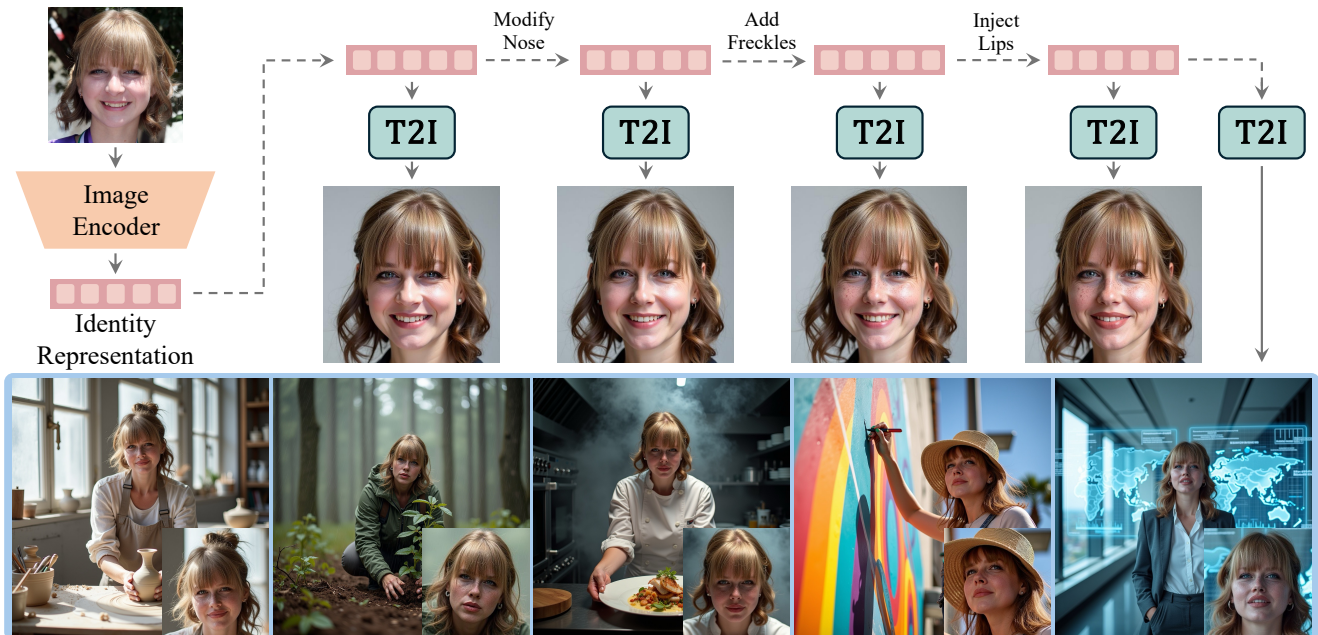


Figure 11. We present methods for directly tuning the identity tokens of a personalization encoder, enabling fine-grained control of facial attributes, for example modifying the nose, adding freckles, or a beard (top). The edited identity can then be used across diverse prompts to generate the same tuned subject consistently in new scenes (bottom).

Appendix

A. Additional Implementation Details

Token Selection We construct the token groups S as follows. For localized facial attributes, we create 5-10 (I, I_{patch}) pairs per attribute and apply the localized attribute selection method. For each pair, we select the top $k = 20$ tokens ranked by $\|\Delta Z_n\|_2$. The final set S for each attribute consists of the tokens that appear most frequently in these top- k selections across pairs. To select tokens for global attributes, we train one linear SVM per token and retain those whose validation accuracy exceeds a threshold $\tau = 0.7$. The selection process is computed once per region or attribute and reused for all subsequent edits.

PCA over identity tokens We encode the entire FFHQ dataset with the Omni-ID encoder to obtain token sequences Z . For a chosen set of token indices S , we concatenate the corresponding token embeddings for each image into a single vector $x_S = \text{vec}(Z_S)$. We center each coordinate across the dataset and fit scikit learn `IncrementalPCA` using default parameters. On our hardware the fit takes about 10 minutes on a CPU. To identify salient components and their roles, we apply the top principal direction to a small set of example identities with a scale of 100 and inspect the outcomes. We then evaluate the effect of each component either manually or with a vision and language model, retaining those that produce meaningful identity tuning, including attributes that are not easily described in words.

Supervised attribute directions via linear SVM We encode the CelebA dataset to obtain Z for all images. For each target attribute and token set S , we form $x_S = \text{vec}(Z_S)$, balance the training data to have equal positive and negative examples using the CelebA binary labels, and train a linear classifier with hinge loss using the scikit learn `SGDClassifier` with default parameters. In our runs this training completes in about 10 minutes on a CPU and can be accelerated with a GPU.

B. Additional Experiments

B.1. Identity Tuning Results

Fig. 11 illustrates identity tuning: given a source image, our approach enables iterative tuning of the identity, highlighting the advantage of operating directly on the identity representation. The tuned identity produces consistent generations across diverse prompts, and the figure includes edits that are difficult to express via text, such as modifying the shape of the nose.

Fig. 12 provides additional examples of supervised directional editing applied to facial identities. These results demonstrate that the modifications remain consistent across various text prompts while being strictly localized to the target attribute.

B.2. Continuous Tuning

To tune an identity, we perform vector addition in the identity latent space, allowing us to control the intensity of the edit using a scale parameter. Fig. 13 (Omni-ID) and Fig. 14 (PuLID) illustrate a smooth transformation obtained by in-



Figure 12. Each row shows a different source identity (leftmost column) and its edits along different directions. The images in each column are generated by adding the same direction found by SVM. For each edit, we show the main result with two alternative prompts stacked vertically to the right.



Figure 13. Continuous identity space via interpolation. The leftmost and rightmost columns use the encoding of two different identities (A and B). Intermediate columns are generated by linearly interpolating between their identity embeddings and applying the mixed embedding to all tokens. The identity changes gradually without abrupt jumps, illustrating the continuity of the learned identity representation across varied scenes.

terpolating between two identities. Fig. 15 shows three scale levels for each edit direction, derived either in an unsupervised manner via PCA (changing eyebrows) or in a supervised manner via attribute-aligned mean-difference (adding bangs) or SVM-based directions (adding freckles). These directions enable fine-grained, monotonic control over the strength of each edit.

B.3. Q-Former Attention Maps

Fig. 18 presents the attention maps for the PuLID [29] encoder, which exhibit patterns similar to those of Omni-ID [59] shown in Figure 2 in the main paper. In both cases, each learned token consistently focuses on a specific semantic region of the face, extracting localized information from areas such as the cheeks, forehead, eyebrows, nose, and eyes.

B.4. Patch Editing

We explore identity tuning through a patch-based operation. We perform a pixel-level manipulation of an input image, as shown in Fig. 16, and use the manipulated image’s representation as a new identity representation. We find that the encoder integrates the pasted patch smoothly into the representation, producing coherent identities that differ from the original only in the modified region. This method enables local identity edits and offers fine-grained control over the appearance of the edited part. It serves as a useful tool for targeted identity refinements, allowing edits that are difficult to specify through text, such as selecting a specific pair of eyeglasses or a particular mustache style.

Fig. 16 presents three example edit types. In each, our



Figure 14. Continuous identity space via interpolation. The left-most and rightmost columns use the encoding of two different identities (A and B). Intermediate columns are generated by linearly interpolating between their identity embeddings and applying the mixed embedding to all tokens. The identity changes gradually without abrupt jumps, illustrating the continuity of the learned identity representation across varied scenes.



Figure 15. Slider-based results. Each row shows the gradual nature of our edits.

method performs a localized modification of the identity, enabling consistent generation of the tuned identity across



Figure 16. Patch-based identity editing across multiple attributes and prompts. Top: we modify the source image by pasting different facial patches (eyes, lips, and a Van Dyke beard). Bottom: each modified image is encoded into the identity space to produce an edited identity representation, which is then used to generate diverse images under different prompts.

diverse contexts. The appearance of the pasted attributes is accurately preserved, for example in the style of the beard, while the remaining facial characteristics stay unchanged.

B.5. Comparisons

Qualitative Comparisons Fig. 17 presents a qualitative comparison across methods. For each method, we show results for two identities, applying a different edit to each. We generate four images per edited identity under diverse prompts to assess identity consistency across prompts.

Competing methods often apply the requested edit yet fail to preserve identity across prompts. In Flux Kontext Direct [40] this is expected: without an explicit identity embedding, attributes are not applied consistently from one generation to the next. The identity often drifts even between the input and its edited output, especially for large edits such as adding a beard. Flux Kontext Sequential improves identity coherence relative to Direct, but in cases when an edit fails (for example, bangs) it fail across all samples. We also observe spillover into unrelated regions, such as unintended skin changes. In contrast, our approach preserves a coherent identity while reliably expressing the intended edit across prompts.

Benchmark The benchmark contains 20 identity images sampled from FFHQ, five basic identity edits (the common edits supported by Precise Control and Weights2Weights),

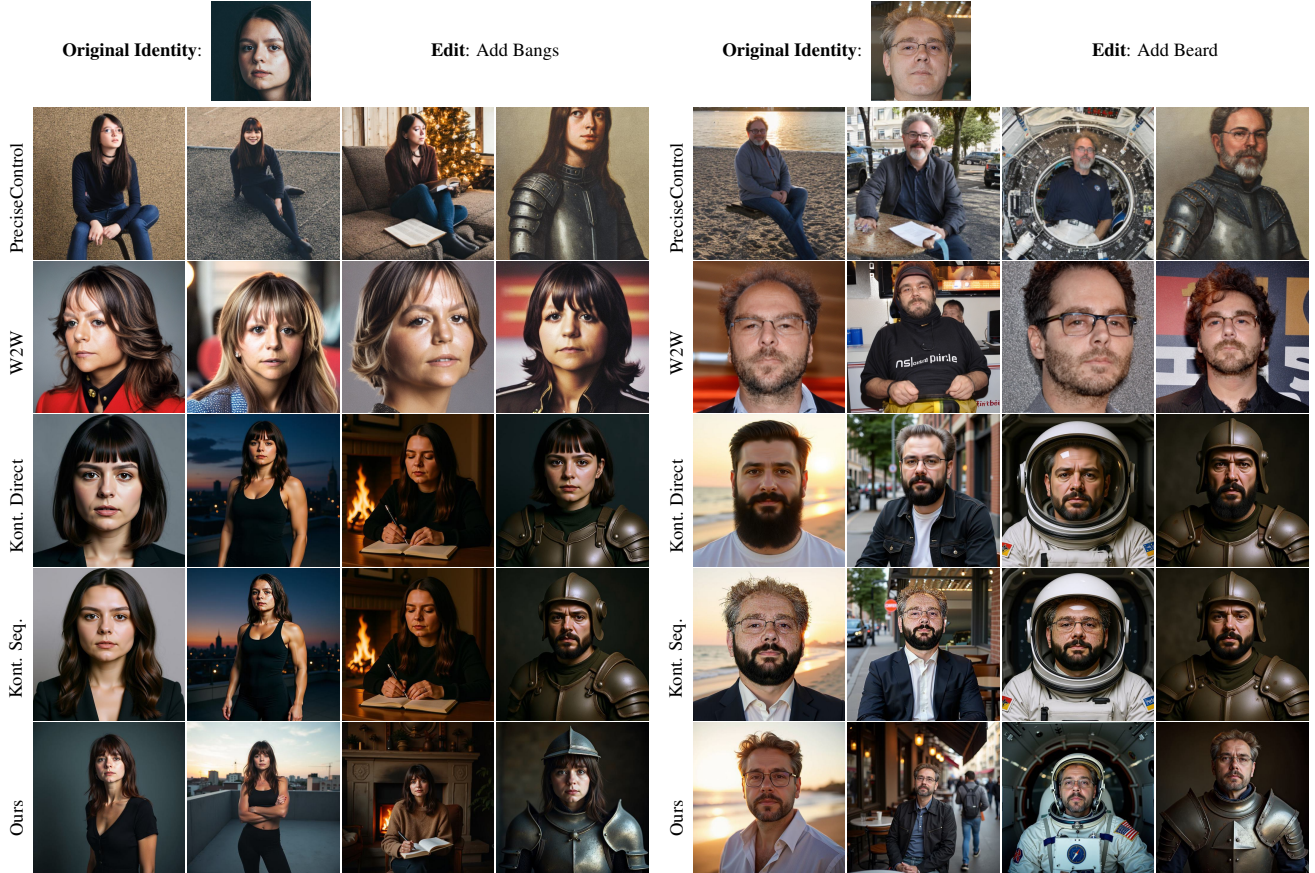


Figure 17. Qualitative comparison. Each grid presents the results of an identity tuning of a single attribute, the left grid adds bangs and the right grid adds a beard. Each grid shows results for four prompts per method as columns. Our method consistently applies the target attribute across diverse prompts while keeping the edited identity coherent.

and seven advanced edits that require finer control. We use 13 prompts generated by an LLM [51], each describing a scene that includes a person. Applying all edits to each identity yields 100 basic identity edits and 140 advanced edits. For each method, we render 13 images per edit, resulting in 1,300 basic and 1,820 advanced images per method. Tab. 3 presents the prompts which used to generate the results images, the basic and the advanced edit instructions.

As summarized in Table 2 in the main paper, our method was preferred over all baselines across all categories. These results indicate that users judged our approach to better preserve the original identity, maintain consistency across prompts, and adhere to both the specified edit and the target prompt. That indicates the superiority of our approach for the task of identity tuning.

B.6. Flux Kontext implementation details

We evaluate two Flux Kontext baselines [40] that we defined for qualitative and quantitative analyses:

Kontext Direct (single step). Given a single reference image and a target scene prompt, Kontext applies the edit

and renders the scene in one pass. Template: $\{\text{EDIT}\}$ this person and generate an image of $\{\text{BASE_PROMPT}\}$. Example (freckles): Add freckles to this person and generate an image of $\{\text{BASE_PROMPT}\}$.

Kontext Sequential (two steps). We separate the local edit from the scene rendering. Step 1 edits the portrait in isolation; Step 2 conditions on the edited image to render the scene. Step 1: $\{\text{EDIT}\}$ this person $\rightarrow$ Step 2: Generate an image of this person $\{\text{BASE_PROMPT}\}$. Example (freckles): Add freckles to this person $\rightarrow$ Generate an image of this person $\{\text{BASE_PROMPT}\}$.

Settings. Official checkpoint with default inference parameters, the reference portrait is the only image input.

C. Limitations

A limitation of our approach relates to the token capacity of the Q-Former. Our localized editing performs optimally when a sufficient number of queries are used, enabling the

Table 3. The complete set of source prompts and edit instruction used for our quantitative evaluation and user study.

Prompts
Professional headshot of one person on a beach at golden hour, centered composition, eyes looking at camera, crisp focus
Studio fashion portrait of one person on a seamless backdrop, centered composition, three-point lighting, tack-sharp details
Street portrait of one person at a sidewalk café, centered composition, natural light, subtle bokeh
Athletic portrait of one person on a rooftop at dusk, centered composition, dynamic stance, clean background separation
Renaissance-inspired oil painting of one person, half-length, centered composition, soft chiaroscuro, ornate textures
Cyberpunk portrait of one person under neon signs, centered composition, teal-magenta glow, shallow DOF
Outdoor lifestyle portrait of one person in a meadow, centered composition, airy pastel tones, gentle bokeh
Astronaut portrait of one person inside a spacecraft, centered composition, visor lifted, panel reflections, high realism
Cozy indoor portrait of one person journaling by a fireplace, centered composition, warm tungsten glow, soft textures
Surreal portrait of one person hovering above desert dunes, centered composition, pastel sky, tilt-shift effect
Hip-hop album cover portrait of one person by a classic car at night, centered composition, colored gels, subtle film grain
Minimalist editorial portrait of one person in a red coat against a white wall, centered composition, strong negative space
Medieval knight portrait of one person in weathered armor, centered composition, raised visor, cinematic realism
Basic Edits
Add/Remove beard
Make the person bald
Change the person gender
Add/Remove eyeglasses
Add/Remove bangs
Advanced Edits
Add/Remove freckles
Make the eyebrows arched
Make the eyebrows bushy
Add/Remove big lips
Add/Remove mustache
Make the skin pale
Add rosy cheeks

natural disentanglement of facial regions. Conversely, a low token count forces the model to encode multiple facial aspects into a single token. This compression prevents isolated editing through single or group token manipulation. While our discovered semantic directions (both supervised and unsupervised) can still navigate this compressed space to edit specific attributes without affecting others, direct token swapping is heavily penalized. Consequently, operations like local blending between identities produce entangled results if the underlying tokens contain information from multiple distinct facial parts. Ultimately, these findings suggest that future personalization models should train

the Q-Former with an adequate token count to preserve natural disentanglement, thereby unlocking greater controllability for downstream editing tasks.

D. Broader Impacts

Our work advances the fundamental understanding of identity representation in generative models, enabling more precise and controllable editing. While this offers significant benefits for creative applications and personalized media, it also carries potential societal risks. Similar to other identity manipulation techniques, fine-grained control over facial attributes could be misused to generate misleading content or alter a person’s likeness without consent. To mitigate these risks, we encourage the continued development of robust detection algorithms and responsible deployment practices, such as requiring watermarks for generated media and adhering to strict consent protocols when utilizing personal reference images.

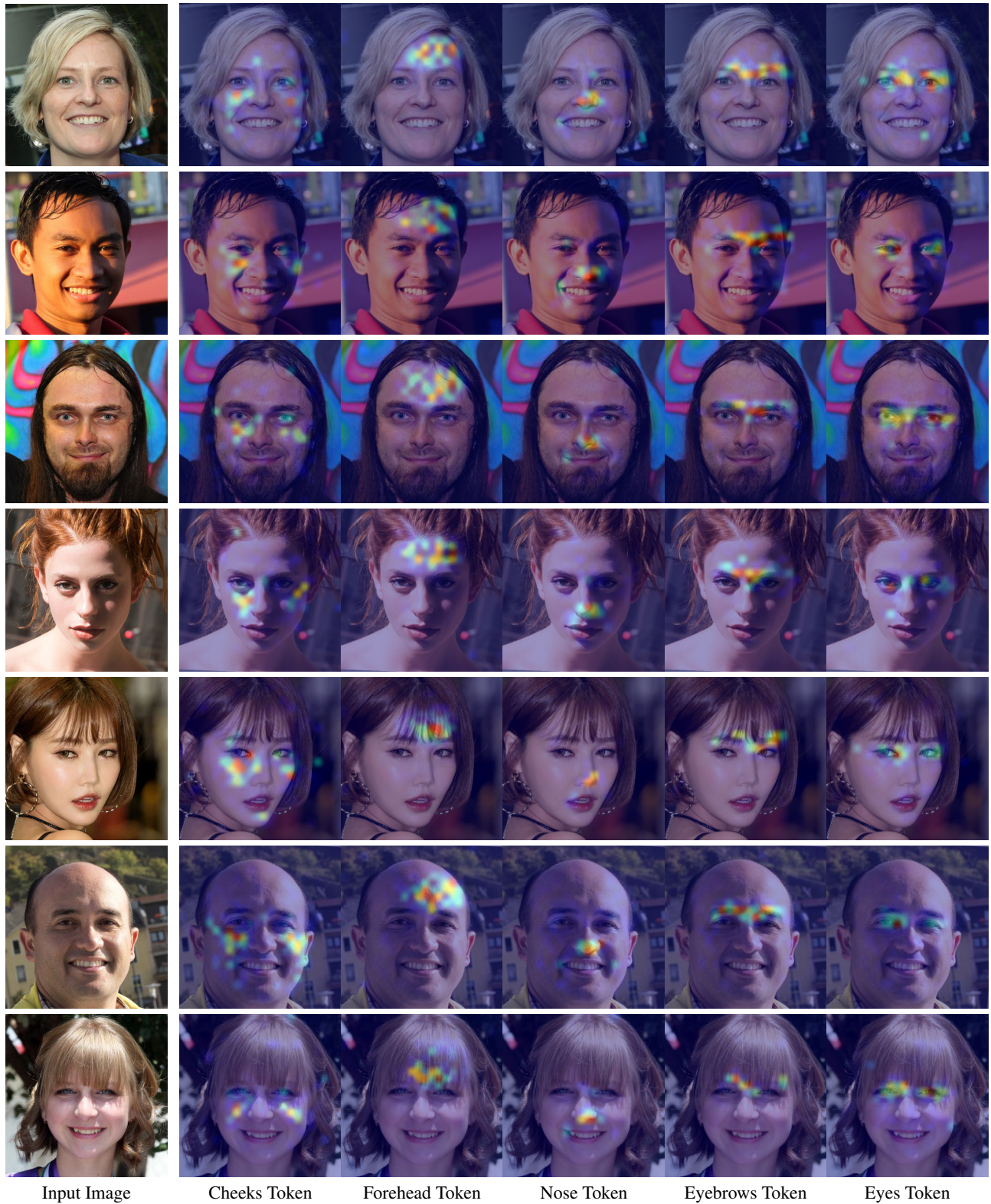
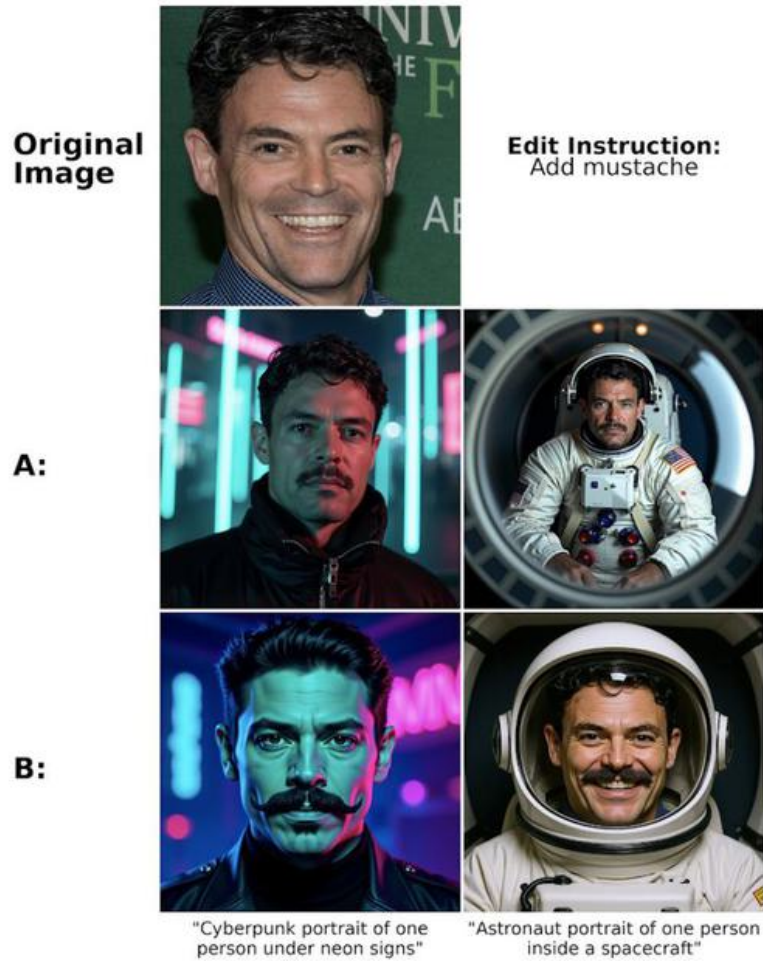


Figure 18. Token behavior in the face encoder. Q-Former attention maps for five learned query tokens reveal localized semantics. The same token keeps the same role across both identities, showing consistent token-to-region mapping.



	A	B
Which method better preserves the identity from the original image?	☐	☐
Which method shows a more consistent identity across its two images?	☐	☐
Which method performs the edit instruction more accurately and clearly?	☐	☐
Which method better matches the text prompts under the images?	☐	☐
Overall, which method do you prefer for this task?	☐	☐

Figure 19. An example of a question in the user study