EditID: Training-Free Editable ID Customization for Text-to-Image Generation

Anonymous ACL submission

Abstract

We propose EditID, a training-free approach based on the DiT architecture, which achieves highly editable customized IDs for text-to-image generation. Existing text-to-image models for customized IDs typically focus more on ID consistency while neglecting editability. It is challenging to alter facial orientation, character attributes, and other features through prompts. EditID addresses this by deconstructing the text-to-image model for customized IDs into an image generation branch and a character feature branch. The character feature branch is further decoupled into three modules: feature extraction, feature fusion, and feature integration. By introducing a combination of mapping features and shift features, along with controlling the intensity of ID feature integration, EditID achieves semantic compression of local features across network depths, forming an editable feature space. This enables the successful generation of high-quality images with editable IDs while maintaining ID consistency, achieving excellent results in the IBench evaluation, which is an editability evaluation framework for the field of customized ID text-to-image generation that quantitatively demonstrates the superior performance of EditID. EditID is the first text-to-image solution to propose customizable ID editability on the DiT architecture, meeting the demands of long prompts and high-quality image generation.

1 Introduction

ID customization generation (Guo et al., 2024; Gal et al., 2022; Kumari et al., 2023; Ruiz et al., 2023; Liu et al., 2023), as a personalized type of text-to-image generation, integrates IDs with prompts to create specific appearances. It offers significant application value in scenarios such as story generation and character creation, and it is one of the core selling points of major text-to-image creative production platforms today.

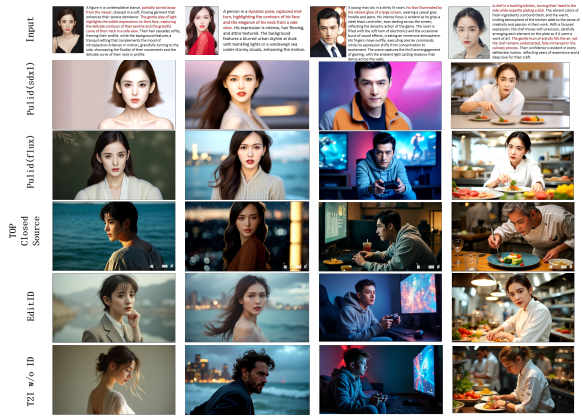

Figure 1: We introduce EditID, a training-free ID customization approach. EditID achieves better editability compared to similar methods. It demonstrates excellent editability in long prompts (where action prompts are marked in red) and aligns well with Flux's T2I.

ID customization methods include fine-tuning (Ruiz et al., 2023; Liu et al., 2023), tuning-free (Guo et al., 2024; Xiao et al., 2024; Ye et al., 2023; Zhang et al., 2024), and training-free (Tewel et al., 2024) approaches. Fine-tuning requires time-consuming, ID-specific training. Tuning-free methods pretrain an ID fusion module on a large portrait dataset, avoiding ID-specific customization during inference. Training-free methods enhance tuning-free approaches, eliminating retraining in both training and inference. Our training-free framework enables ID editability in any ID customization model with a character feature branch.

Current ID customization methods generally prioritize character consistency, thereby overlooking character editability. We define character editability as the ability to generate multidimensional control relative to the input ID in response to changes in text prompts, including variations in facial orientation and limb positioning, as well as the flexible modification of related attributes of the input ID, such as hairstyle, accessories, and even age and gender. Under our definition, current consistency-

focused methods generally lack editability. Taking the state-of-the-art consistency generation method PuLID (Guo et al., 2024) as an example, the model achieves fidelity through ID loss but also introduces semantic and layout losses to control the diversity of the input ID during generation. However, in practical applications, it is nearly impossible to induce significant pose changes in the input ID through prompt words. This prompted us to investigate the reasons behind the loss of editability. We found that the model is essentially performing an ID reconstruction task. During the pre-training of tuning-free methods, the ID and prompt descriptions remain almost identical, and the model leverages feature information from the ID to bias the training parameters toward the distribution of the training set. This implies that the model intends to directly replicate the character's features. However, during inference, when the prompts and ID are inconsistent, excessively strong feature constraints from the character branch result in the output ID lacking the ability to adapt to changes in the prompt. In some cases, this even leads to a "copy-paste" effect between the input ID and output ID in the facial region, as shown in Figure 1. The core contribution of this paper lies in modulating the control strength of the character branch within a training-free framework, achieving a text-to-image model that significantly enhances ID editability while maintaining ID consistency.

The core of our approach lies in enhancing editability while maintaining character consistency, achieving a stable balance between the two. We deconstruct the text-to-image model for ID customization, dividing it into an image generation branch and a character feature branch. The character feature branch is further decoupled into three modules: feature extraction, feature fusion, and feature integration. In feature extraction, we isolate five layers of identity-aware features from the fine-grained local feature extractor EvaCLIP (Sun et al., 2023), which we term "mapping features." In feature fusion, due to the one-to-one correspondence between mapping features and five groups of neural networks, we designate features outside the five-layer mapping features of EvaCLIP as "shift features." We discovered that these two types of features essentially perform semantic compression of local features across network depths, forming an editable feature space. The combination of these features enables predictable editability variations. In feature integration, when the fused features interact with the image generation branch, we introduce a dynamic information fusion mechanism in cross-attention to further enhance editability, ultimately achieving state-of-the-art (SOTA) results in the IBench evaluation framework.

Additionally, most current models are based on the UNet architecture, with SD (Rombach et al., 2022) and SDXL (Podell et al., 2023) serving as foundational models. In practical applications, such as story generation scenarios, there are typically two requirements: 1) the input of long prompts, and 2) higher aesthetic quality for the generated images. Consequently, image generation based on the DiT (Peebles and Xie, 2023) architecture, such as Flux (Flux, 2024), becomes the preferred choice. However, there is scarcely any literature discussing methods to enhance character editability while preserving character consistency in DiT-based image generation. To the best of our knowledge, we are the first to explore improving character editability within the DiT architecture.

Considering the current lack of unified datasets and metrics for evaluating both ID consistency and editability in the field of ID customization, we propose a configurable and modular automated evaluation framework. We designed multiple sets of evaluation character images, accounting for real IDs and generated IDs, and adapted them to various types of prompts, including editability-measuring prompts, short prompts, and more. For the first time, we comprehensively introduce a variety of editability verification metrics to quantitatively assess model performance.

In summary, our contributions are as follows:

1. EditID is the first character customization method to address editability enhancement within the DiT architecture.

2. We propose an editability enhancement approach under a training-free framework, utilizing mapping features, shift features, and dynamic ID integration across three modules. By decoupling the character feature branch, we achieve a highly editable text-to-image model while maintaining ID consistency.

3. We introduce a unified ID consistency evaluation framework and, for the first time, propose multiple editability metrics. EditID demonstrates excellent performance on editability-focused prompts.

2 Related Works

2.1 ID Consistency Methods

Recent personalized T2I diffusion models emphasize ID consistency generation, focusing on robust semantic facial features. Key methods include IP-Adapter-FaceID (IP-Adapter-FaceID, 2024), which uses facial embeddings and decoupled cross-attention for ID consistency; Photomaker (Li et al., 2024), encoding multiple ID images into stacked embeddings; InstantID (Wang et al., 2024), employing IdentityNet to integrate facial features, landmarks, and text via semantic and spatial constraints; and PuLID (Guo et al., 2024), combining Lightning and standard diffusion branches with contrastive alignment and precise ID loss for high ID fidelity with minimal model interference.

2.2 ID Editable Methods

Few T2I works on ID customization prioritize editability, focusing instead on ID consistency. PortraitBooth (Peng et al., 2024) enables text-based expression editing using subject embeddings and emotion-aware cross-attention. ConsistentID (Huang et al., 2024a) integrates multimodal facial prompts and an ID preservation network with facial attention localization for precise detail and consistent identity. Current UNet-based methods (e.g., SD, SDXL) are tuning-free, yield average texture quality, struggle with long prompts, and require pre-training on large facial datasets.

2.3 Training-Free Framework

The training-free framework, distinct from fine-tuning and tuning-free methods, is used in image and video generation. FreeU (Si et al., 2024) uses backbone and skip-connection scaling factors to improve denoising and retain clarity. Freelong (Lu et al., 2024) combines global and local video features for coherent long video generation.To the best of our knowledge, we are the first to introduce the training-free framework to the domain of customized ID editable in text-to-image generation under the DiT architecture.

3 Method

Our approach focuses on enhancing character editability while maintaining consistency. Figure 2 shows the architecture, splitting the ID customization into an image generation main branch and a character feature branch. The feature branch is divided into three modules: feature extraction, fusion, and integration. This framework identifies editability sources as: 1) local ID features from extraction; 2) ID shift features from fusion; and 3) embedding strength for integrating ID information into the DiT backbone. The third aspect affects both editability and consistency. By carefully combining local ID features, we achieve optimal editability.

3.1 Preliminary

3.1.1 DiT Flow Matching

The DiT architecture leverages flow matching to model data generation as an ODE, replacing traditional UNet with Transformer modules to capture global context via self-attention, with details provided in the Appendix (Section A.1). We adopt the Flux framework as the base model for its regularized flow matching and improved noise scheduling.

3.1.2 DiT-Based ID Consistency

ID customization under DiT is rare, mostly using UNet-based SD/SDXL. FluxCustomID (FluxCustomID, 2024) employs ArcFace (Deng et al., 2019) and CLIP (Hafner et al., 2021) with PerceiverResampler (Alayrac et al., 2022), but lacks in fidelity, consistency, and editability. PuLID (Guo et al., 2024), the state-of-the-art, uses a Flux-based approach with a Lightning T2I and standard diffusion branch to address ID embedding interference and fidelity via contrastive semantic comparison and ID loss. Its character feature branch uses ArcFace and EvaCLIP for global and local features. We use PuLID as the baseline, leveraging its ID and alignment losses, and study editability in a training-free framework by segmenting the character feature branch into three modules.

3.2 Feature Mapping

In the character feature extraction module, most methods combine global and local features, with editability found mainly in local features. EditID's extraction module has two branches: Branch 1 uses SCRFD (Deng et al., 2021) for lightweight face detection and ArcFace for global feature extraction of the detected region. Branch 2 employs RetinaFace for detecting five facial keypoints followed by frontal alignment. After facial semantic segmentation, a refined facial region is obtained and EvaCLIP extracts fine-grained local features.From EvaCLIP's 23-layer features, we select five identity-aware layers. Surprisingly, we found that the selec-

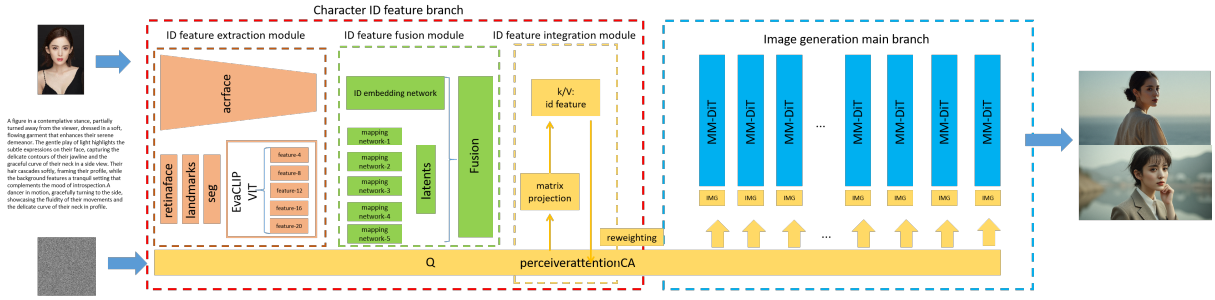

Figure 2: Overview of the EditID Framework. The right half features a DiT-based image generation process. The left half is the ID feature branch, divided into three modules: 1) the ID feature extraction module, which extracts global and local features to generate mapping features; 2) the ID feature fusion module, which fuses mapping features to produce shift features; and 3) the ID feature integration module, which implements a dynamic ID embedding mechanism.

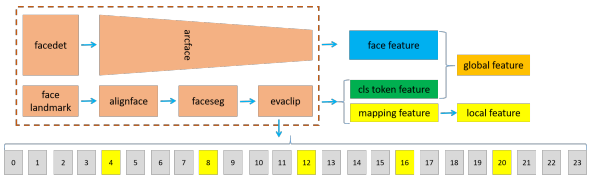

Figure 3: The character feature extraction module combines blue facial features and dark green EvaCLIP CLS token as global features. Yellow marks mapping features from EvaCLIP layers 4, 8, 12, 16, 20. Gray shows unselected features.

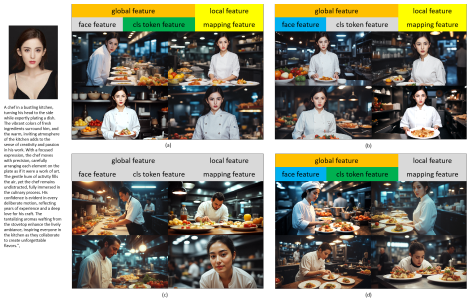

Figure 4: Impact of global and local features on generation editability. Gray indicates unselected features, set to zero.

Figure 5: Relationship between mapping feature variations and editability. White indicates unselected features.

tion of these five identity-aware layers significantly enhances editability. We refer to these five identity-aware features as "mapping features." The global features are composed of the face features from Branch 1 and the CLS token features from Branch 2, while the mapping features directly correspond to local features. Ultimately, both global and local feature sets are output and fed into the feature fusion module. A detailed diagram of the module is shown in Figure 3.

We conducted further analysis on global and local features. When inputting prompts with noticeable action changes, as shown in Figure 4(a) and Figure 4(b), setting the facial features or CLS token features of the global features to zero revealed no significant changes in limb or facial orientation. However, Figure 4(b) exhibited higher character consistency than Figure 4(a), indicating that ArcFace extracts finer-grained facial features compared to CLIP, capturing distinctions more effectively. In Figure 4(c), when all features and local features were set to zero, the character feature branch became ineffective, degenerating into standard Flux image generation. This significantly improved editability but eliminated character consistency. In Figure 4(d), setting local features to zero still preserved good editability, though charac-

ter consistency decreased. This led us to discover that global features predominantly control character consistency, while editability is concealed within local features. Global features tend to encode overall ID information of the face, such as facial structure, exhibiting high coupling and stability. In contrast, local features, through the identity-aware filtering of EvaCLIP, reduce 23 layers of features to 5, essentially achieving semantic compression across network depths. This process decouples features of different facial attributes at a fine-grained level, forming independently operable semantic units. We performed a more granular decomposition of the mapping features in EvaCLIP. As shown in Figure 5, from Figure 5(a) to Figure 5(f), we observed that as the mapping features were filtered, editability changed accordingly, but character consistency was also affected. An increase in editability corresponded to a decrease in character consistency. This prompted us to explore the optimal balance point between consistency and editability.For a more detailed derivation of the formulas, refer to the Appendix (Section A.2.1).

3.3 Feature Shift

In the feature fusion module, global features are input into the ID embedding network, a neural network consisting of three linear layers, while local features are fed into the mapping network. The mapping network shares a similar structure to the ID embedding network, and both facilitate feature transformation. Our method is based on a training-free architecture. Therefore, during the tuning phase, the mapping features and the mapping network establish a one-to-one mapping relationship. When we replace the mapping features with shift features, a feature shift occurs between them and the mapping network. The mapping features consist of only five groups, selected as five identity-aware features from the 23 layers of EvaCLIP. This selection essentially achieves semantic compression across network depths: shallow layers capture compositional structure information, middle layers encode detailed geometric structures, and deep layers associate with high-level semantics. This hierarchical selection constructs an editable semantic space, where different layers correspond to facial editing dimensions of varying granularity. Visualizing the facial features of EvaCLIP provided us with guidance for selecting features. We found that the choice of shift features also significantly impacts editability, and we ultimately selected the

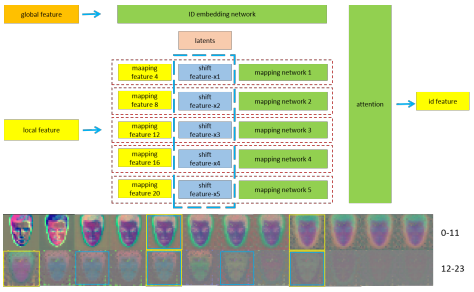

Figure 6: Upper half shows feature fusion with shift features; lower half visualizes EvaCLIP's 23-layer facial features, with yellow modules as mapping features and blue boxes as combined shift and mapping features.

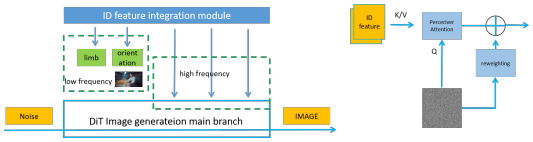

Figure 7: The left figure shows ID interaction between the integration module and image generation, while the right depicts soft ID control within the module.

feature combination of layers 4, 14, 16, 18, and 20.For a more detailed derivation of the formulas, refer to the Appendix (Section A.2.2).

3.4 ID Feature Integration

After the feature fusion module, a single feature set is sent to the ID feature integration module, which interacts with the image generation branch via PerceiverAttention. In Flux's 57 blocks (19 dual-stream + 38 single-stream), 10 are chosen to embed ID information. Early generation focuses on low-frequency features (color, composition), while later stages refine high-frequency details. We modulate ID embedding intensity accordingly during initial denoising (see Figure 7), but overly strong embedding disrupts noise balance, hindering low-frequency decoding. This naive adjustment causes initial generation bias, harming convergence and resulting in darker images with loss of lighting and stability.

Secondly, we adopted a softer approach to ID strength control. In the feature integration module, the generated noise image serves as the Query, while the ID information acts as the Key and Value for weighting. During output, we perform reweighting on the Query to align it with the same dimensional size as the ID feature, followed by information supplementation through a residual connection, using concatenation for fusion. Reweighting can be implemented in various ways. To achieve dimensional transformation without excessively weakening the generated noise, we designed a transfor-

mation matrix. Considering the characteristics of information retention, we explored methods such as randn linear, DCT, and partial Fourier, ultimately adopting the randn linear approach. The editability of image generation primarily stems from the image side. However, text information is embedded within the noise image, which essentially serves as the starting point of the denoising process and contains latent semantic information. By compensating for the noise image, the semantic influence of the text can more smoothly permeate the image generation process that integrates ID embedding information. This is equivalent to introducing additional degrees of freedom in the latent space, enabling text-driven editability to be realized under the constraints of ID information without being overly restricted by the ID embedding.

Ultimately, by combining the mapping features of the ID feature extraction module, the shift features of the ID fusion module, and the soft ID strength control mechanism of the ID integration module, we achieved excellent editability while preserving ID consistency.

4 IBench

To address the lack of robust evaluation metrics for character consistency and editability in personalized character image generation, we propose IBench, a configurable, modular, automated evaluation framework. It implements diverse editability verification metrics to quantify improvements in a training-free architecture.

4.1 Dataset

The evaluation data of IBench consists of two parts: prompts and evaluation images. The evaluation images are divided into three groups: Unsplash, ChineseID, and GenerateID.The prompts in IBench are categorized into three dimensions: short prompts, action prompts for editability (Huang et al., 2024b), and manually collected prompts.for detailed descriptions, see the Appendix (Section A.4.1).

4.2 Evaluation Metrics

We designed the metrics from three dimensions: consistency, editability, and the T2I general evaluation dimension. The T2I general evaluation includes FID, Aesthetic, and Imaging Quality (Ke et al., 2021) metrics. The consistency dimension includes Facesim, ClipT, ClipI, Dino (Zhang et al., 2022), and Fgis metrics. The editability dimension includes Posediv (Yin and Liu, 2017) (Doosti et al., 2020), Landmarkdiff, and Exprdiv metrics. For detailed information, see the Appendix (Section A.4.2).

5 Experiments

5.1 Setting

We use the Flux version of PuLID as the base model. For the Flux model, the sampling steps are set to 20, with a guidance scale of 3.5, a CFG scale of 1, and the Euler sampler is employed. Our final combination of mapping features and shift features consists of five groups [4, 14, 16, 18, 20], and we implement residual dynamic ID information embedding in the ID integration module, using concatenation as the fusion method. All experiments are conducted on four NVIDIA A100 GPUs, with the inference framework being ComfyUI.

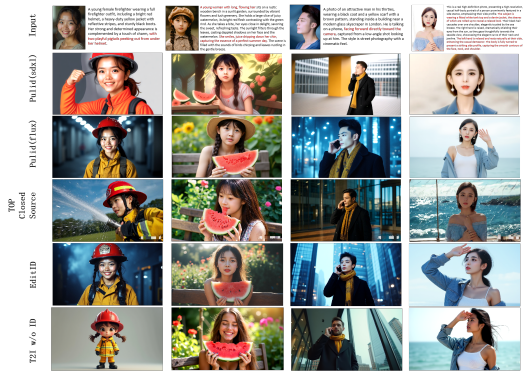

Figure 8: Qualitative Comparison: T2I w/o ID shows Flux T2I output without ID insertion. EditID ensures higher editability and ID consistency, accurately editing hairstyle, accessories, age, face, and limbs.

5.2 Qualitative Comparison

We adopt the SDXL version of InstantID, PuLID, and the Flux version of PuLID as the comparative model group, where the base models for InstantID and PuLID SDXL are `sdxl_base_1.0`. As shown in Figure 8, EditID takes long text prompts as input and, while maintaining model consistency, achieves better editability compared to the nearly synchronous copy-paste face insertion of PuLID Flux. In the first column, when adding "with two playful pigtails peeking out from under her helmet", EditID successfully alters the hairstyle. Compared to Flux T2I, the ID embedding enhances the ability to align the scene synchronously. In the second column, when adding "A young woman with long, flowing hair", EditID enables age changes; however, the expression changes are less pronounced compared to the generation by Flux T2I. In the

Model	FID	Aesthetic	Image Quality	Posediv			Landmarkdiff	Exprdiv
				Yaw	Pitch	Roll		
InstantID	9.780	0.585	0.394	12.62	7.545	5.266	0.044	0.647
PuLID (SDXL)	11.28	0.675	0.502	19.48	6.187	12.69	0.099	0.593
PuLID (Flux)	14.59	0.681	0.431	9.298	5.872	9.473	0.070	0.562
EditID	13.52	0.683	0.454	11.81	6.722	10.60	0.082	0.554
	Facesim	ClipI	ClipT	Dino	Fgis			
InstantID	0.642	0.613	0.185	0.301	0.449			
PuLID (SDXL)	0.399	0.768	0.248	0.129	0.353			
PuLID (Flux)	0.735	0.757	0.243	0.178	0.501			
EditID	0.714	0.769	0.249	0.162	0.459			

Table 1: Evaluation metric results from IBench on ChineseID with editable long prompts

third column, with a side-profile ID image input and the addition of "facing forward directly toward the camera," EditID achieves facial rotation and completes the face when turned from a side profile to a front view. In contrast, PuLID Flux can hardly rotate the face. In the fourth column, with a front-facing ID image input and the addition of "The subject is wearing a fitted white tank top and a denim jacket, the sleeves of which are rolled up to reveal a relaxed look", EditID realizes a front-to-side profile transition while maintaining alignment with Flux T2I. The ID generation quality and detail handling of the PuLID SDXL version are generally poor, with low fidelity, possibly due to the limited generation capability of the SDXL base model. The character consistency is also inferior to the Flux version, and the generated images exhibit a weak non-realistic stylistic attribute. Top closed-source text-to-image models offer good richness in generation, producing scene associations not present in the prompts, but their fidelity and character consistency are both poor.

We primarily focus on the performance of EditID with long prompts. In Table 1, using the evaluation combination of ChineseID with editable long prompts, EditID performs well across the three conventional aesthetic metrics: FID, Aesthetic, and Image Quality. In the Facesim metric, which measures consistency, EditID shows only a slight decrease compared to the Flux version of PuLID. However, as shown in Figure 8, the Flux version of PuLID exhibits excessively strong consistency, even resulting in a copy-paste replication of the input ID face in the output, which significantly limits the applicability of text-image consistency generation. Facial and limb features need to exhibit different variations across various scenes. In the ClipI metric, it is evident that the ID insertion in EditID does not strongly interfere with the original generation capability. The ClipT metric also demonstrates good text-following ability, while Dino and Fgis, which are finer-grained consistency evaluation metrics, show significant improvements. For the most critical editability metrics, EditID achieves a total improvement of 5 points in the three Euler angles of Posediv compared to the Flux version of PuLID, and it also shows a substantial increase in Landmarkdiff. The Facesim metric decreases by only 2 points, indicating that EditID sacrifices only a slight degree of similarity while delivering excellent editability. In fact, the overly strong character consistency constraint in PuLID Flux suggests that releasing excessive consistency in exchange for enhanced editability is a highly prudent choice. Compared to the SDXL version of PuLID, while editability is high, it sacrifices too much character consistency.

5.3 Ablation Study

In our training-free framework, it is necessary to evaluate the impact of changes and combinations of multiple metrics. The following experimental groups primarily focus on similarity and editability metrics.

5.3.1 Combination of Mapping and Shift Features

In Section 3, we qualitatively analyzed the sources of editability variations from mapping features and shift features. The multi-level, fine-grained local semantic features introduced by EvaCLIP are the source of editability. Below, we quantitatively discuss this in two groups. In Table 2, we mainly examine the first two feature selections. In fact, the choice of the first feature is critical. As a shallow feature, it contains rich editable semantic information. In the first four groups, the first feature is the fourth-layer feature of EvaCLIP, and the second feature is the 8th, 12th, and 16th layers, respectively. It is evident that as semantic information weakens, Facesim decreases, but editability significantly increases, showing a strong inverse relationship. Comparing the third to sixth groups, where the first feature is replaced, we observe a decrease in Facesim but a greater increase in editability. This indicates that feature offsets at the same level (shallow layers) can provide substantial editability. Using IBench, we further dissect the combination information at a finer granularity.

In Table 3, after selecting the shallow-layer features, we further examine the mid-layer and deep-layer features. Selecting only shallow-layer features significantly improves editability but results in greater loss of character consistency. When inputting rich long prompts, the generated images

Features	Facesim	ClipT	Posediv			Landmarkdiff	Exprdiv
			Raw	Pitch	Roll		
[4,-,-,-,-]	0.512	0.262	27.46	11.28	13.17	0.115	0.621
[4,8,-,-,-]	0.635	0.256	17.59	8.288	11.46	0.096	0.589
[4,12,-,-,-]	0.617	0.258	19.25	9.033	11.89	0.102	0.591
[4,16,-,-,-]	0.591	0.259	23.10	9.689	12.41	0.108	0.581
[0,12,-,-,-]	0.601	0.261	20.51	9.343	12.18	0.106	0.593
[0,16,-,-,-]	0.574	0.262	24.75	10.41	13.26	0.112	0.585

Table 2: Compare shift features of ChineseID with editable long prompts, where features are 5-layer lists and "-" denotes a layer set to 0.

Features	Facesim	ClipT	Posediv			Landmarkdiff	Exprdiv
			Raw	Pitch	Roll		
[4,12,18,20,-]	0.689	0.253	12.73	6.858	10.42	0.085	0.566
[4,14,18,20,-]	0.684	0.254	13.28	6.973	10.76	0.087	0.565
[4,16,18,20,-]	0.668	0.255	14.46	7.328	11.10	0.090	0.562
[4,12,16,18,20]	0.713	0.247	11.24	6.610	10.26	0.081	0.561
[4,14,16,18,20]	0.714	0.247	11.81	6.722	10.60	0.082	0.554
[4,16,16,18,20]	0.701	0.249	12.77	7.029	10.92	0.086	0.555

Table 3: Quantitatively compare feature combinations of ChineseID with editable long prompts, where features are 5-layer lists and "-" denotes a layer set to 0.

closely resemble ID-free text-to-image (T2I) results. We further screen mid-layer and deep-layer features, as deep-layer features provide richer details, contributing to improved image fidelity. We plotted the differences in Facesim and PoseDiv (raw/pitch/roll) values between our base model PuLID (Flux) and our proposed model in Figure 9. We observed that as the feature groups are adjusted, Facesim and PoseDiv exhibit a linear relationship. We selected the most cost-effective combination from the curve, ensuring high-level consistency while enhancing editability.

5.3.2 Shift Strategy

For the combination of mapping features and shift features, the feature fusion module selects five sets of features to enter the mapping network, ultimately outputting ID features. These features are then integrated into the main image generation branch through the ID dynamic embedding mechanism in the ID integration module. When fewer than five feature sets are available, interpolation can replace zero-padding; when more than five are available, average fusion is an option. Figure 10 compares two strategies for excess features (average and max) and two for insufficient features (padding and interpolate). In Figure 10(a)(b), the max strategy yields sharper images with stronger lighting. In Figure 10(c)(d), interpolation results in lower image quality. Selecting features from 23 layers outperforms feature modification, favoring the mapping and shift feature combination. The average method, based on mean shift, also performs effectively.

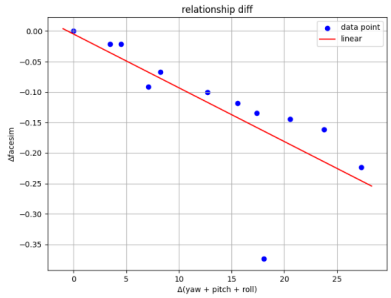

Figure 9: Variation curves of the differences in raw/pitch/roll values from PoseDiv and Facesim compared to the corresponding values of PuLID (Flux).

Figure 10: Feature shift strategies: first row for more than five feature sets, second row for five or fewer.

6 Conclusion

This paper proposes EditID, a training-free ID customization method for text-to-image generation. We are the first to explore enhancing editability within the DiT architecture, achieving state-of-the-art performance with long prompts. Taking the PuLID model as an example, we deconstruct it into a character feature branch and an image generation main branch, further decoupling the character feature branch into three major modules: feature extraction, feature fusion, and ID integration. We analyze the sources of editability from the combination of mapping features and shift features, as well as dynamic ID integration, thereby improving the editability of ID customization. Our approach requires no training, demonstrating its potential for flexible and efficient character-customized image generation. Moreover, this training-free framework can be adapted to enhance any ID customization generation algorithm equipped with a character feature branch. In future work, we will continue to explore and investigate the dynamic ID integration module with the introduction of a training mode. We believe that dynamic ID integration holds great vitality, but it still requires the design of loss functions incorporating richer multi-angle facial information to further achieve simultaneous improvements in both character consistency and editability.

Limitations

As required by ACL Rolling Review, we outline the limitations of our work. EditID, while achieving significant improvements in editability within a training-free framework, has certain constraints. First, the approach relies heavily on the quality of the pre-trained DiT model (e.g., Flux) and the feature extractors (e.g., EvaCLIP, ArcFace). Any limitations in these foundational models, such as biases in the training data or suboptimal feature representations, may affect EditID's performance. Second, the trade-off between editability and ID consistency requires careful tuning of mapping and shift features, which may not generalize perfectly across all types of IDs or prompts. Third, the computational complexity of processing long prompts and integrating ID features may pose challenges for deployment on resource-constrained devices. Finally, the IBench evaluation framework, while comprehensive, may not capture all nuances of editability and consistency, particularly for highly subjective or context-specific scenarios. Future work will address these limitations by exploring more robust feature extractors, automated tuning mechanisms, and expanded evaluation metrics.

Acknowledgments

We thank the iFlyTek team for their support in providing computational resources and feedback during the development of EditID. This work was supported by internal funding from iFlyTek.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, and 1 others. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736.

Jiankang Deng, Jia Guo, Xiang An, Zheng Zhu, and Stefanos Zafeiriou. 2021. Masked face recognition challenge: The insightface track report. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1437–1444.

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

Bardia Doosti, Shujon Naha, Majid Mirbagheri, and David J Crandall. 2020. Hope-net: A graph-based model for hand-object pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6608–6617.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, and 1 others. 2024. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*.

Flux. 2024. Flux: High-resolution image synthesis with transformers. Https://github.com/black-forest-labs/flux.

FluxCustomID. 2024. Fluxcustomid: Customizing id for flux-based image generation. Https://github.com/damo-cv/FLUX-customID.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. 2022. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*.

Zinan Guo, Yanze Wu, Chen Zhuowei, Peng Zhang, Qian He, and 1 others. 2024. Pulid: Pure and lightning id customization via contrastive alignment. *Advances in Neural Information Processing Systems*, 37:36777–36804.

Markus Hafner, Maria Katsantoni, Tino Köster, James Marks, Joyita Mukherjee, Dorothee Staiger, Jernej Ule, and Mihaela Zavolan. 2021. Clip and complementary methods. *Nature Reviews Methods Primers*, 1(1):20.

Jiehui Huang, Xiao Dong, Wenhui Song, Zheng Chong, Zhenchao Tang, Jun Zhou, Yuhao Cheng, Long Chen, Hanhui Li, Yiqiang Yan, and 1 others. 2024a. Consistentid: Portrait generation with multimodal fine-grained identity preserving. *arXiv preprint arXiv:2404.16771*.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, and 1 others. 2024b. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21807–21818.

IP-Adapter-FaceID. 2024. Ip-adapter-faceid: Text compatible image prompt adapter for text-to-image diffusion models. Https://huggingface.co/h94/IP-Adapter-FaceID.

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. 2021. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5148–5157.

Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. 2023. Multi-concept

customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1931–1941.

Zhen Li, Mingdeng Cao, Xintao Wang, Zhongang Qi, Ming-Ming Cheng, and Ying Shan. 2024. 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.

Yang Liu, Cheng Yu, Lei Shang, Yongyi He, Ziheng Wu, Xingjun Wang, Chao Xu, Haoyu Xie, Weida Wang, Yuze Zhao, and 1 others. 2023. Facechain: a playground for human-centric artificial intelligence generated content. *arXiv preprint arXiv:2308.14256*.

Yu Lu, Yuanzhi Liang, Linchao Zhu, and Yi Yang. 2024. Freelong: Training-free long video generation with spectralblend temporal attention. *arXiv preprint arXiv:2407.19918*.

William Peebles and Saining Xie. 2023. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205.

Xu Peng, Junwei Zhu, Boyuan Jiang, Ying Tai, Donghao Luo, Jiangning Zhang, Wei Lin, Taisong Jin, Chengjie Wang, and Rongrong Ji. 2024. Portraitbooth: A versatile portrait model for fast identity-preserved personalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27080–27090.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 2023. 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.

Chenyang Si, Ziqi Huang, Yuming Jiang, and Ziwei Liu. 2024. Freeu: Free lunch in diffusion u-net. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4733–4743.

Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. 2023. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*.

Yoad Tewel, Omri Kaduri, Rinon Gal, Yoni Kasten, Lior Wolf, Gal Chechik, and Yuval Atzmon. 2024. Training-free consistent text-to-image generation. *ACM Transactions on Graphics (TOG)*, 43(4):1–18.

Qixun Wang, Xu Bai, Haofan Wang, Zekui Qin, Anthony Chen, Huaxia Li, Xu Tang, and Yao Hu. 2024. Instantid: Zero-shot identity-preserving generation in seconds. *arXiv preprint arXiv:2401.07519*.

Guangxuan Xiao, Tianwei Yin, William T Freeman, Frédo Durand, and Song Han. 2024. Fastcomposer: Tuning-free multi-subject image generation with localized attention. *International Journal of Computer Vision*, pages 1–20.

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. 2023. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*.

Xi Yin and Xiaoming Liu. 2017. Multi-task convolutional neural network for pose-invariant face recognition. *IEEE Transactions on Image Processing*, 27(2):964–975.

Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. 2022. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*.

Shilong Zhang, Lianghua Huang, Xi Chen, Yifei Zhang, Zhi-Fan Wu, Yutong Feng, Wei Wang, Yujun Shen, Yu Liu, and Ping Luo. 2024. Flashface: Human image personalization with high-fidelity identity preservation. *arXiv preprint arXiv:2403.17008*.

A Example Appendix

The structure of the supplementary material is as follows:

A.1 DiT Flow Matching

The core of diffusion models lies in achieving data generation through a progressive denoising process. Traditional diffusion models define the forward diffusion process as

$$q(x_t|x_0) = \mathcal{N}(x_t; \alpha_t x_0, \sigma_t^2 I) \tag{1}$$

where α_t and σ_t are noise scheduling coefficients, and $t \in [0, T]$ represents continuous time steps. Based on flow matching theory (Esser et al., 2024), the generation process can be modeled as an ordinary differential equation (ODE):

$$dx = v_\theta(x_t, t, c)dt \tag{2}$$

where v_θ is the vector field to be learned, and c is the conditional input (e.g., text prompts). Compared to traditional diffusion models, which rely on noise prediction targets, flow matching directly learns the transport mapping from the data distribution to the noise distribution. Its training objective can be expressed as:

$$\mathcal{L}_{FM}(\theta) = \mathbb{E}_{t,q(x_0),p(x_1)}\left[\|v_\theta(x_t, t, c) - (x_1 - x_0)\|^2\right] \tag{3}$$

where $x_t = (1-t)x_0 + tx_1$ represents a linear interpolation path, and $t \in [0, 1]$.

In the DiT architecture, we replace the traditional UNet with Transformer modules, leveraging the self-attention mechanism to model global context. Given a conditional embedding sequence c, the vector field predictor in DiT can be decomposed as:

$$v_\theta = \text{Proj}(\text{Attn}(Q, K, V)) \tag{4}$$

where $Q = x_t W_Q$, $K = cW_K$, and $V = cW_V$ are the query, key, and value matrices, respectively. W denotes learnable projection matrices, and Attn represents the multi-head attention mechanism. This architecture is particularly well-suited for long-text conditional generation, as its self-attention mechanism effectively captures long-range dependencies between prompts.

Currently, image generation models based on the DiT architecture and flow matching are primarily represented by SD3 (Esser et al., 2024) and Flux (Flux, 2024). This paper selects Flux as the foundational framework. The Flux framework builds upon DiT by introducing a regularized flow matching strategy, employing an improved noise scheduling function:

$$\alpha_t = \cos^2(\pi t/2), \quad \sigma_t = \sin^2(\pi t/2) \tag{5}$$

This ensures a smooth transition from data to noise while maintaining numerical stability during the flow matching process.

A.2 Method Formula

A.2.1 Feature Mapping Formula

Feature mapping module consists of two parts: global features and cross-layer local features (mapping features). Its core objective is to decouple the global features that control ID consistency from the local features that carry the freedom of editability. As shown in Branch 1, global features are jointly extracted by two-dimensional encoders:

$$F_{\text{global}} = [\Psi_{\text{arcface}}(I); \Phi_{\text{CLS}}(I)] \in \mathbb{R}^{d_g} \tag{6}$$

where $\Psi_{\text{arcface}} : \mathbb{R}^{H\times W\times 3} \to \mathbb{R}^{d_a}$ represents the ArcFace-based dense facial encoder, responsible for extracting deep semantic features related to facial ID; $\Phi_{\text{CLS}} : \mathbb{R}^{H\times W\times 3} \to \mathbb{R}^{d_c}$ denotes the CLS token feature from the EvaCLIP image encoder, capturing the overall visual context information of the character. The two are fused through a concatenation operation $[;]$ into a global feature vector of dimension $d_g = d_a + d_c$.

As shown in Branch 2, in the cross-layer local features, editable local features are obtained through hierarchical semantic compression:

$$F_{\text{local}} = \{\phi_l(I)\}_{l\in L_{\text{map}}} \in \mathbb{R}^{5\times d_l} \tag{7}$$

where $L_{\text{map}} = \{l_1, l_2, \ldots, l_5\}$ represents the five feature layers selected from the 23 layers of EvaCLIP, and $\phi_l(I)$ denotes the output feature map of the l-th layer. Through a cross-layer filtering mechanism, feature responses strongly correlated with facial attributes are isolated, forming independently manipulable semantic units.

A.2.2 Feature Shift Formula

Feature shift module fuses global and local features, introducing richer editability through mapping features and shift features. The formula is as follows:

$$F_{\text{edit}} = \text{Attn}\big(\theta_{\text{ID}}(I), M_{\text{Map}}(F'_{\text{local}})\big) \in \mathbb{R}^{d_{\text{id}}}, \tag{8}$$

where $F'_{\text{local}} = [F_{\text{local}}(l); F_{\text{shift}}(l)]$ and F_{edit} is the output combined feature, $\theta_{\text{ID}}(I)$ is derived from

an ID embedding network composed of three linear layers, and M_{Map} represents the mapping network. Here, F'_{local} is the combination of mapping features and shift features, F_{local} denotes the local features composed of mapping features, and F_{shift} represents the shift features. The total number of mapping features and shift features satisfies:

$$|L_{\text{map}}| + |L_{\text{shift}}| = 5, \tag{9}$$

where L_{map} and L_{shift} denote the sets of mapping features and shift features, respectively.

A.3 More Details about Ablation Study

A.3.1 Editability in ID Integration Module

The dynamic ID integration design in the ID integration module is also a significant source of editability. In this module, we primarily considered two dimensions: reweighting and feature fusion methods. Reweighting ensures dimensional consistency with the ID embedding features without compromising the noise features, while the fusion method appropriately compensates the features after ID integration back into the sampled noise features, enhancing editability in the text dimension. We conducted a qualitative analysis in Figure 11 to examine the impact of the randn linear approach in reweighting and various fusion methods. The fusion methods include: Weight: Assigning different fusion weights to the two feature sets; Dropout: Randomly masking features after reweighting to reduce information redundancy; Concat: Concatenating the two feature sets and then computing their mean for fusion; Sum: Directly summing the two feature sets; Multiply: Multiplying the two feature sets; Max: Taking the maximum of the two feature sets. We observed that in Figure 11(a)(b)(c), the ID still exhibited a strong binding effect, with no significant changes in facial orientation, though the image fidelity decreased considerably. In Figure 11(d)(e)(f), ID consistency declined, but editability gradually increased. Ultimately, we selected the concat fusion method from Figure 11(d), combining it with mapping features and shift features to achieve a high level of consistency.

A.4 More Details about IBench

A.4.1 Dataset

The evaluation data of IBench consists of two parts: prompts and evaluation images. The evaluation images are divided into three groups: Unsplash, ChineseID, and GenerateID. The Unsplash group includes 49 images, covering a variety of skin tones, significant variations in character angles, and instances of facial occlusion. ChineseID comprises 100 images of Chinese individuals collected from the internet, including well-known figures from fields such as film and sports, representing diversity in gender, age, and multiple angles. GenerateID consists of 100 ID images generated by a text-to-image model, featuring refined facial features, diverse poses, accessories, and hairstyles, as well as characters under light and shadow rendering, ensuring both aesthetic quality and prominent ID characteristics.We present partial data from ChineseID, GenerateID and Unsplash in Figure 12 Figure 13 Figure 14.

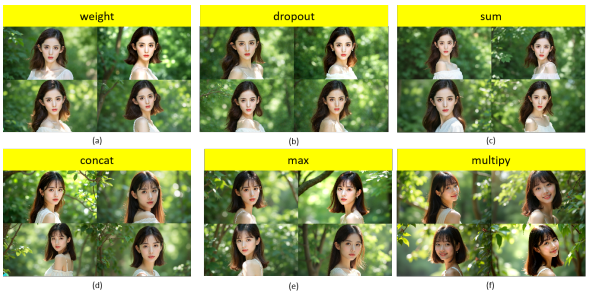

Figure 11: Effect diagram of different feature fusion methods after reweighting.

The prompts in IBench are categorized into three dimensions: short prompts, action prompts for editability, and manually collected prompts. Short prompts, designed to be compatible with the evaluation of the UNet architecture, are largely sourced from mainstream evaluation reports. Editable long prompts are sourced from the augmented prompts in VBench (Huang et al., 2024b), with 41 groups selected, including long descriptions of character actions and scene stories, serving as a key focus of this evaluation. Manually collected prompts consist of 80 groups of prompts collected from text-to-image users, manually curated to include rich action descriptions with story elements.We display the prompt words for short prompt, editable long prompt, and manually collected prompt in the Table 6 Table 7 Table 8 below.

In IBench, we pair Unsplash with short prompts as one group, ChineseID with editable long prompts as another group, and GenerateID with manually collected prompts as a third group. However, in practice, these images and prompts can be cross-validated.

A.4.2 Evaluation Metrics

We designed the metrics from three dimensions: consistency, editability, and the T2I general evaluation dimension. In the T2I general evaluation, the focus is primarily on the aesthetic attributes of the images. In the consistency dimension, the evaluation centers on the similarity of character ID generation and prompt-following consistency. In the editability dimension, we propose multiple innovative measurement indicators to assess the editability of character IDs.

T2I General Evaluation Dimension **FID**: By comparing the distribution differences between `imagewithid` (generated images with ID information) and `imagewithoutid` (generated images without ID information) in the feature space of a pre-trained InceptionV3 model, we quantify the similarity between the two distributions.

Aesthetic: The LAION aesthetic predictor is used to evaluate the aesthetic quality score of `imagewithid` images. It reflects dimensions such as the harmony and richness of layout and color, as well as the realism and naturalness of the images.

Imaging Quality: This assesses distortions (e.g., overexposure, noise, blur) present in `imagewithid` images, using the MUSIQ (Ke et al., 2021) image quality predictor trained on the SPAQ dataset for evaluation.

Consistency Dimension **Facesim**: This calculates the facial similarity between the ID image and `imagewithid`. We use SCRFD from InsightFace to detect facial regions and ArcFace to extract facial feature vectors, then compute the cosine similarity to measure the similarity of the generated facial regions.

ClipT: This computes the cosine similarity between the CLIP text encoding of the input prompt and the CLIP image encoding features of `imagewithid`. It is used to evaluate the ability of the generated images to follow changes in the prompt.

ClipI: This calculates the cosine similarity between the CLIP image encodings of `imagewithid` and `imagewithoutid`. It measures the similarity between the two images before and after ID insertion. A higher ClipI score indicates that the modifications to image elements after ID insertion cause less interference compared to the original model's generation.

Dino (Zhang et al., 2022): This computes the cosine similarity between the DINO image encodings of the ID image and `imagewithid`. DINO features are more fine-grained and can be used to measure the changes in the generated image relative to the ID image.

Fgis: Corresponding to the DINO metric, this calculates the cosine similarity of the DINO image encodings for the facial regions of the ID image and `imagewithid`.

Model	FID	Aesthetic	Image Quality	Posediv			Landmarkdiff	Exprdiv
				Yaw	Pitch	Roll		
InstantID	12.87	0.613	0.403	13.31	7.994	5.977	0.046	0.662
PuLID (SDXL)	8.514	0.663	0.530	18.61	7.093	10.13	0.087	0.541
PuLID (Flux)	20.31	0.684	0.452	7.620	5.647	8.445	0.064	0.482
EditID	19.28	0.682	0.464	9.795	6.881	9.683	0.075	0.485
	Facesim	ClipI	ClipT	Dino	Fgis			
InstantID	0.646	0.607	0.178	0.301	0.375			
PuLID (SDXL)	0.384	0.768	0.252	0.189	0.390			
PuLID (Flux)	0.724	0.729	0.231	0.286	0.579			
EditID	0.701	0.739	0.238	0.262	0.475			

Table 4: Evaluation metric results from IBench on GenerateID with manually collected prompts

Model	FID	Aesthetic	Image Quality	Posediv			Landmarkdiff	Exprdiv
				Yaw	Pitch	Roll		
Instantid	61.13	0.568	0.422	24.23	13.94	12.60	0.125	0.541
PuLID(sdxl)	31.24	0.659	0.490	22.41	11.58	12.74	0.107	0.669
PuLID(flux)	43.64	0.697	0.461	20.29	12.14	12.19	0.099	0.574
EditID	16.84	0.696	0.486	21.07	12.88	13.16	0.104	0.612
	Facesim	ClipI	ClipT	Dino	Fgis			
Instantid	0.184	0.699	0.219	0.071	0.099			
PuLID(sdxl)	0.372	0.832	0.251	0.113	0.206			
PuLID(flux)	0.393	0.803	0.238	0.128	0.211			
EditID	0.380	0.813	0.240	0.091	0.131			

Table 5: Evaluation metric results from IBench on Unsplash with short prompts

Editability Dimension **Posediv**: This calculates the differences in Euler angles (yaw, pitch, and roll) of the facial regions between the ID image and `imagewithid`. Facial detection is performed using MTCNN (Yin and Liu, 2017), and Euler angles are extracted using Hopenet (Doosti et al., 2020). This metric is used to assess the editability of the facial regions.

Landmarkdiff: This computes the difference in the average Euclidean distance of five normalized key points between the facial regions of the ID image and `imagewithid`.

Exprdiv: This calculates the proportion of expression changes in the facial regions between the ID image and `imagewithid`.

A.5 More Results

A.5.1 ChineseID Experimental Results

We observed that the performance on ChineseID with editable long prompts and GenerateID with manually collected prompts is similar. Therefore, we will focus on the performance of ChineseID editable long prompts in subsequent analysis.

No.	Prompt
1	Clothing portrait, a person wearing a spacesuit
2	Portrait, a person wearing a surgical mask
3	Background portrait, with a beautiful purple sunset at the beach in the background
4	Portrait, pencil drawing
5	Portrait, latte art in a cup
6	Portrait, side view, in papercraft style
7	Portrait, Madhubani, wearing a mask
8	Portrait, anime artwork
9	Portrait, energetic brushwork, bold colors, abstract forms, expressive, emotional
10	Portrait, a person wearing a doctoral cap

Table 6: Examples of partial prompts from short prompt

A.5.2 Unsplash Experimental Results

This paper focuses more on long prompts, but IBench itself also includes an evaluation for short prompts. When compared with the ChineseID editable long prompt and GenerateID Typemovie prompt, our evaluation metrics on the Unsplash short prompt typically show lower facesim (face similarity) and higher posediv (pose diversity), which aligns with general variation patterns. However, as shown in Figure 14 and Table 6, the large variations in portrait angles in Unsplash, combined with overly short prompts, lead to generally poor model performance. Due to the coarse text control of short prompts, we incorporated a [man/woman] design in the prompts. When the model correctly identifies the gender of the image, it replaces "person" in the prompt accordingly.

A.6 Supplementary Experimental Materials

Prompts from Figure 10 and Figure 11:

- A figure in a contemplative stance, partially turned away from the viewer, dressed in a composed outfit that enhances his serene demeanor. The gentle play of light highlights the subtle expressions on his face, capturing the delicate contours of his jawline and the graceful curve of his neck in a side view. His profile is framed by the tranquil setting in the background, which complements the mood of introspection.

- This is a young Asian individual with fair skin, whose straight, deep brown hair falls just below the shoulders. Thick bangs partially cover the forehead, accentuating a sharply defined side profile and large, gentle eyes that exude a hint of tenderness. She is slightly turning, with her shoulder gently tilted, displaying an elegantly natural pose as if she is slowly turning to admire the lush green scenery around her. Her high, refined nose stands out in the soft sunlight, and her slightly parted lips reveal a serene yet enchanting smile. Sunlight filters through the dense branches and leaves, casting dappled shadows on her profile and infusing the entire scene with a gentle, peaceful ambiance.

A.7 Show case of Some EditID Results

Figure 15 shows some results of editid.As shown in the image, for the input ID, not only can it display facial and body movements, but it can also alter attributes such as age and hairstyle through prompts, demonstrating exceptional editability. Moreover, the quality of the generated images is extremely high.

Figure 12: Examples of partial character ID images from ChineseID

Figure 13: Examples of partial character ID images from GenerateID

Figure 14: Examples of partial character ID images from Unsplash

No.	Prompt
1	A solitary man, clad in a long, dark trench coat and a wide-brimmed hat, walks through a dimly lit alleyway, the only illumination coming from flickering street lamps casting elongated shadows. His footsteps echo softly against the cobblestones, creating a rhythmic pattern in the stillness of the night. The air is thick with mist, swirling around his silhouette, adding an air of mystery to his journey. Occasionally, he pauses, glancing over his shoulder, as if sensing an unseen presence. The distant sound of a train whistle punctuates the silence, enhancing the eerie, atmospheric setting of his solitary walk.
2	A young boy with tousled hair and a curious expression kneels in a sunlit garden, surrounded by vibrant blooms. He carefully places a delicate glass dome over a single, exquisite red rose, its petals glistening with morning dew. The sunlight filters through the glass, casting a kaleidoscope of colors onto the grass. His small hands gently adjust the dome, ensuring the rose is perfectly encased. The scene captures a moment of wonder and protection, as the boy admires the rose's beauty, the garden's lush greenery and colorful flowers providing a serene, enchanting backdrop.
3	A joyful child, wearing a bright yellow raincoat and red rubber boots, splashes gleefully in a series of puddles on a rainy day. The scene captures the child's infectious laughter as they jump, sending droplets flying in all directions. The overcast sky and gentle rain create a soothing backdrop, while the child's playful antics bring warmth and energy to the scene. As they stomp through the water, their reflection shimmers in the puddles, adding a magical touch. The child's carefree spirit and the rhythmic sound of raindrops create a heartwarming and lively atmosphere.
4	A young athlete, clad in a sleek black swimsuit and swim cap, stands at the edge of an Olympic-sized pool, the water shimmering under bright overhead lights. With a focused gaze, she adjusts her goggles, preparing for her training session. She dives gracefully into the water, her form streamlined and powerful, creating minimal splash. As she glides through the water, her strokes are precise and rhythmic, showcasing her dedication and skill. The camera captures her underwater, bubbles trailing behind her as she propels forward with determination. Finally, she emerges at the pool's edge, breathing deeply, her expression a mix of exhaustion and triumph.
5	A bearded man with a thoughtful expression stands in a cozy, dimly lit room filled with vintage decor, wearing a plaid shirt and jeans. He carefully selects a vinyl record from a wooden shelf lined with albums, the warm glow of a nearby lamp casting soft shadows. As he gently places the record onto the turntable, his fingers move with precision and care, reflecting his appreciation for music. The room is filled with the soft crackle of the needle touching the vinyl, and he closes his eyes momentarily, savoring the nostalgic sound. The ambiance is intimate, with the gentle hum of the record player and the soft lighting creating a serene atmosphere.
6	A passionate teacher stands at the front of a bright, modern classroom, holding a vibrant red marker in her hand, gesturing animatedly as she explains a complex concept to her attentive students. Her expression is one of enthusiasm and engagement, with her eyes sparkling with the joy of teaching. The whiteboard behind her is filled with colorful diagrams and notes, illustrating the topic at hand. Sunlight streams through large windows, casting a warm glow over the room, while students, seated at sleek desks, listen intently, some taking notes, others nodding in understanding, creating an atmosphere of dynamic learning and interaction.
7	A fierce woman stands confidently in her intricately detailed cosplay costume, embodying a warrior from a fantasy realm. Her armor, crafted from shimmering silver and deep blue materials, glistens under the ambient light, highlighting the ornate designs etched into the metal. Her long, flowing cape billows behind her as she strikes a powerful pose, her eyes focused and determined. The costume includes a helmet adorned with intricate patterns and a pair of gauntlets that suggest strength and agility. She holds a beautifully crafted sword, its blade reflecting the light, ready for battle. The background is a mystical landscape, with towering mountains and a sky painted in hues of twilight, enhancing the epic atmosphere of her warrior persona.

Table 7: Examples of partial prompts from editable long prompt

No.	Prompt
1	Person B, surprised, flying car, backyard, daytime, casual outfit, excitement, medium shot, A young Black woman, dressed in simple, comfortable home attire, steps out of her house into the bright daylight. She gazes in awe at a futuristic flying car gracefully descending in her backyard, close to the landing point. Her expression transforms from astonishment to sheer excitement, capturing the moment perfectly. The scene is set in a modern American neighborhood, with elements of advanced technology like sleek flying vehicles and high-tech devices enhancing the atmosphere. The composition is a medium shot, focusing on her joyful reaction amidst the backdrop of her home.
2	Young female singer, microphone, concert stage, retro neon lights, posters, evening, shiny silver jacket, skinny jeans, high heels, curly hair, dynamic atmosphere, mid-shot, eye level, A vibrant young female singer stands confidently in front of a microphone, prepared to captivate the audience with her performance. She is on a lively concert stage adorned with dazzling retro neon lights and colorful posters, evoking a nostalgic ambiance. It is a bustling evening, and she is dressed in a striking shiny silver jacket that glimmers under the lights, complemented by tight-fitting skinny jeans and stylish high heels. Her hair is voluminous and curly, adding to her energetic presence. The stage lights pulse and flicker in sync with the upbeat music, enhancing the dynamic atmosphere of the scene. The composition is framed as a mid-shot from an eye-level perspective, inviting viewers into the exhilarating moment of the performance.
3	Young woman, self-service check-in system, modern office, morning, professional attire, tablet, busy staff, large screen, real-time data, A focused young Chinese woman with long hair, dressed in professional attire, is operating a sleek tablet at a self-service check-in kiosk. The scene is set in a contemporary office filled with busy staff moving around her. It's 9 AM in the morning, with natural light streaming through large windows. In the background, a large screen displays real-time data, adding a high-tech feel to the atmosphere. The composition is a medium shot at eye level, capturing the dynamic environment and the woman's concentration on her task.
4	Animated character, cashier interaction, shopping process, afternoon, friendly cashier, crowd of customers, A cheerful animated character with large, expressive eyes and a stylish outfit, handing selected items to a friendly cashier at a brightly lit checkout counter. The character is smiling while making a payment, and the cashier is engaging in friendly conversation. In the background, a diverse crowd of customers, including Asian individuals, patiently waits in line. The scene takes place in the afternoon, with warm sunlight streaming through the store's windows, creating a lively and inviting atmosphere.
5	Li Bai, courtyard, well railing, poetry, candlelight, night, warm light, glowing words, close-up, eye level, A young Chinese poet, Li Bai, dressed in a soft moon-white robe, stands in a serene traditional courtyard at night, his face warmly illuminated by the gentle flicker of candlelight. He gazes thoughtfully at the well railing, deeply immersed in reciting poetic verses. Each word he utters releases a faint glow that softly drifts into the air. The ancient courtyard is adorned with lush green pine trees and a bright, shining moon above, adding to the tranquil and poetic atmosphere. The scene is captured in a close-up, eye-level perspective, evoking a sense of intimacy and reflection.
6	Young man, park bench, Shanghai, evening, dark coat, jeans, lost expression, skyline, wilted trees, A young Chinese man wearing a dark coat and jeans, sitting alone on a weathered park bench in a quiet park in Shanghai during the evening. He has slightly disheveled hair and a lost expression in his eyes, reflecting deep thoughts. The background features a few wilted trees and a distant city skyline, creating a serene yet melancholic atmosphere. The scene is captured in a mid-range view, at eye level, emphasizing his solitude and contemplation as the soft evening light casts gentle shadows around him.

Table 8: Examples of partial prompts from manually collected prompt

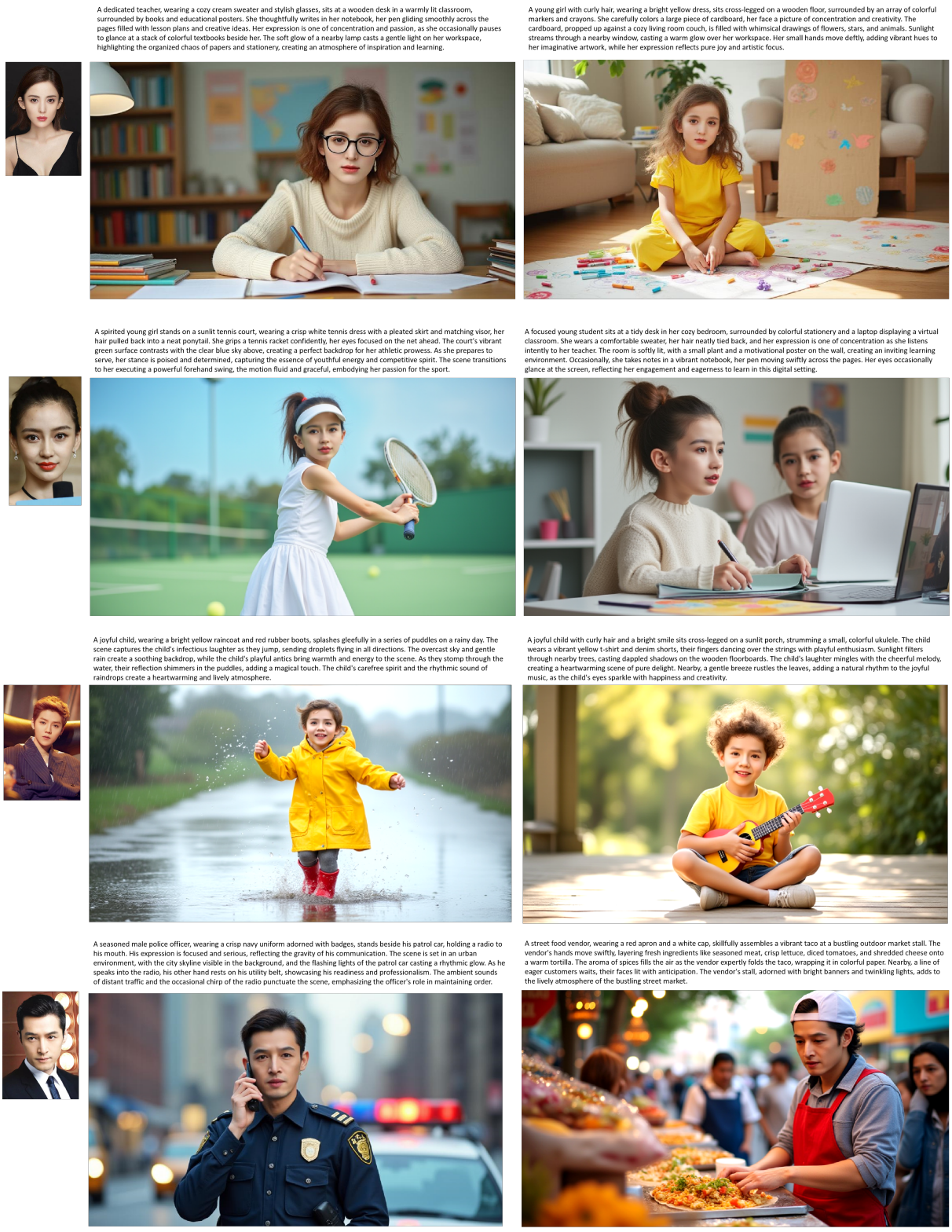

Figure 15: Some results of EditID