

DVI: Disentangling Semantic and Visual Identity for Training-Free Personalized Generation

Guandong Li
iFLYTEK

Yijun Ding
Suning

Abstract

Recent tuning-free identity customization methods have achieved significant success in maintaining facial fidelity by leveraging high-level semantic embeddings from pre-trained face recognition models. However, existing methods often overlook the visual context inherent in reference images—such as lighting distribution, skin texture density, and environmental tone. This limitation frequently leads to “Semantic-Visual Dissonance,” where the generated identity retains accurate facial geometry but loses the unique atmosphere and texture of the input image, resulting in an unnatural “sticker-like” effect. To bridge this gap, we propose DVI (Disentangled Visual-Identity), a novel zero-shot framework that orthogonally disentangles identity customization into a fine-grained semantic stream and a coarse-grained visual stream. Unlike methods relying solely on semantic vectors, DVI exploits the inherent statistical properties of the generative model’s VAE latent space. We elucidate that the first and second-order statistics (mean and variance) of VAE latent variables serve as powerful yet lightweight descriptors for characterizing global visual atmosphere. To fuse these heterogeneous features, we propose a Parameter-Free Feature Modulation mechanism. Instead of training learnable projection layers, this mechanism adaptively modulates the distribution of high-dimensional semantic embeddings using low-dimensional visual statistics, effectively injecting the “visual soul” of the reference image into the generation process. Furthermore, we design a Dynamic Temporal Granularity Scheduler to align with the coarse-to-fine generation characteristic of diffusion models, prioritizing the injection of visual atmosphere in the early denoising stages while refining semantic details in the later stages. Extensive experiments demonstrate that DVI significantly enhances visual consistency and atmospheric fidelity without any parameter fine-tuning, while maintaining robust identity preservation, outperforming existing state-of-the-art methods and achieving excellent results in IBench evaluation.

1. Introduction

Personalized text-to-image generation [4, 11–15, 17, 25] aims to seamlessly integrate specific subject identities into user-provided text descriptions. This technology holds immense application potential in fields such as digital avatar creation, immersive storytelling, and virtual try-on. With the rapid development of diffusion models, the research paradigm in this field has gradually shifted from optimization-based methods (e.g., LoRA [8], DreamBooth [24]) to more efficient tuning-free methods [1, 6, 9, 21]. Tuning-free methods achieve “plug-and-play” zero-shot generation by directly extracting identity features through feed-forward encoders, greatly enhancing user experience.

However, existing tuning-free paradigms face a long-overlooked challenge: *The trade-off between semantic fidelity and visual consistency*. Current mainstream methods mainly rely on pre-trained face recognition models (e.g., ArcFace [3]) or multi-modal encoders (e.g., CLIP [23]) to extract high-level semantic embeddings from reference images [7, 16, 27]. While these embeddings effectively capture the “Who” identity information, their highly compressed nature often discards low-level visual cues of “In what context,” such as the unique lighting distribution, film grain, skin texture density, and overall tonal atmosphere of the reference image.

The absence of visual context frequently leads to Semantic-Visual Dissonance in generated results: while the generated faces retain accurate geometric fidelity to the subject, their lighting, texture, and sharpness often appear generic and over-smoothed, disjointed from the unique visual style of the reference image. This phenomenon is particularly pronounced when processing reference images with strong artistic styles or complex lighting conditions, resulting in a “pasted-on” artifact where the subject lacks organic fusion with the background.

To address this issue, we revisit the composition of identity features. We argue that high-fidelity identity customization requires not only explicit semantic disentanglement but also implicit visual statistics transfer. Based on this insight,

we propose **DVI** (**D**isentangled **V**isual-**I**dentity), a novel training-free identity injection framework.

The core idea of DVI is to orthogonally decompose a single identity feature stream into two complementary streams:

- **Fine-Grained Semantic Stream:** Following existing paradigms, it extracts high-dimensional semantic features to ensure precise reconstruction of facial geometry.
- **Coarse-Grained Visual Stream:** This is our core innovation. Instead of introducing additional heavy encoders, we exploit the potential of the generative model’s own VAE (Variational Autoencoder) latent space. We find that the global first-order (mean) and second-order (variance) statistics of VAE latent feature maps can extremely efficiently characterize the overall visual atmosphere and texture distribution of an image.

To fuse these two heterogeneous feature streams, we discard the linear projectors requiring massive data training in traditional methods and design a Parameter-Free Feature Modulation mechanism. Inspired by the AdaIN concept in style transfer, we use visual statistics as a global distribution bias to directly modulate the distribution of semantic embeddings in the feature space. This approach not only avoids introducing any trainable parameters but also demonstrates strong robustness. Furthermore, considering the temporal characteristics of the generation process, we introduce a Dynamic Temporal Granularity Scheduler to adaptively balance the weights of visual atmosphere laying and semantic detail refinement across different generation stages.

In summary, the contributions of this paper are as follows:

1. **Propose the DVI Framework:** We present a dual-granularity identity disentanglement framework that, for the first time in a tuning-free setting, explicitly distinguishes and synergistically utilizes the semantic and visual attributes of identity, solving the problem where generated images look like the person but lack the correct vibe (resembling in form but not in spirit).
2. **Innovative Use of VAE Statistics:** We demonstrate that without complex attention mechanisms, simply extracting statistical quantities from the VAE latent space effectively captures and transfers the visual atmosphere and texture of the reference image.
3. **Parameter-Free Modulation and Dynamic Scheduling:** We design a training-free feature modulation module and temporal scheduling strategy, significantly improving generation quality and identity consistency in complex lighting and narrative scenes with zero extra training cost.

2. Related Works

2.1. Tuning-Free Identity Customization

With the popularity of diffusion models, personalized generation has evolved from optimization-based methods to more efficient tuning-free methods. These methods aim to achieve identity preservation directly through feed-forward processes without time-consuming test-time fine-tuning for specific subjects. Early attempts like IP-Adapter [28] utilized decoupled cross-attention mechanisms to introduce image prompts, achieving general style and content transfer but with limited fine-grained facial identity preservation. To enhance identity fidelity, recent SOTA methods (e.g., InstantID [25], PhotoMaker [16], FaceClip [18]) typically introduce powerful face recognition models (e.g., ArcFace [3]) as feature extractors. These methods compress reference images into highly abstract semantic embeddings and inject them into the UNet or Transformer of diffusion models by training linear projection layers.

Although these methods have achieved significant results in Identity Recognition, their highly compressed feature representations often lose low-level visual cues from reference images (e.g., skin details, lighting patterns, and environmental tones). This Semantic-Visual Imbalance results in generated images often presenting a homogenized visual effect lacking the original atmosphere. In contrast, DVI is dedicated to completing this missing visual context link without requiring training.

2.2. Visual Context Preservation

Beyond identity customization, preserving the visual attributes (style, texture, layout) of reference images has always been a key topic in generative models. Traditional Style Transfer methods typically use the Gram matrix of Convolutional Neural Networks to capture texture statistical information. In the era of diffusion models, ControlNet [29] and T2I-Adapter [20] introduce extra control branches to preserve spatial structural information, but they focus more on geometric contours than texture atmosphere. Some recent studies have begun to explore using the latent space of Variational Autoencoders (VAE) to preserve visual consistency [2, 19, 21, 26]. Unlike semantic encoders like CLIP, VAE latent variables retain the spatial layout and pixel-level statistical properties of images. Although some works attempt to align VAE features by training complex Reference Attention mechanisms, this introduces high computational and training costs.

DVI proposes a lighter perspective: we believe the “Visual Vibe” of an image can be effectively characterized by the Global Mean and Variance of VAE latent feature maps. This discovery allows us to bypass heavy attention mechanism calculations and directly use statistics as visual descriptors to achieve zero-shot visual atmosphere transfer.

2.3. Feature Modulation and Fusion

Injecting external conditions into diffusion models typically involves two paradigms: Cross-Attention and Concatenation. However, when fusing heterogeneous features (e.g., high-dimensional semantic features and low-dimensional visual statistics), direct concatenation often leads to feature space misalignment. Existing solutions usually rely on training learnable MLP projection layers to map feature dimensions, which not only increases model parameters but may also lead to overfitting or damaging pre-trained model priors. Inspired by Adaptive Instance Normalization (AdaIN [5]) in style generation (StyleGAN [10]), we explore the application of Parameter-Free Feature Modulation in diffusion models. Unlike modules requiring training, DVI uses visual statistics to directly modulate the distribution (shift and scale) of semantic features. This method avoids extra training costs and achieves orthogonal disentanglement and organic fusion of semantic and visual streams in the feature space.

3. Method

The core objective of DVI is to orthogonally disentangle the Fine-Grained Semantic Identity and Coarse-Grained Visual Context from the reference image I_{ref} and synergistically inject them into a pre-trained DiT (Diffusion Transformer) generative model under Zero-Shot and Tuning-Free constraints.

3.1. Coarse-Grained Visual Stream

Existing ID customization methods often employ aggressive image preprocessing, which destroys the compositional proportions and texture density of original images. To capture the global visual atmosphere (e.g., lighting distribution, tone, and film texture) of the reference image, we construct the Coarse-Grained Visual Stream.

Preserve-Aspect-Ratio Latent Encoding. First, we define a texture density preserving preprocessing strategy \mathcal{P} . Given a reference image I_{ref} , we first scale it proportionally based on its shortest edge, followed by a center crop to obtain a fixed-resolution input $x_{vis} \in \mathbb{R}^{3 \times H \times W}$.

$$x_{vis} = \mathcal{P}(I_{ref}) \quad (1)$$

Subsequently, we use the encoder \mathcal{E} of a frozen pre-trained Variational Autoencoder (VAE) to map x_{vis} to the latent space, obtaining the latent feature map Z :

$$Z = \mathcal{E}(x_{vis}) \in \mathbb{R}^{C \times h \times w} \quad (2)$$

where C is the number of latent channels (in our setting $C = 16$), and h, w are the spatial dimensions after down-sampling.

Global Statistics Extraction. We assume that the “visual atmosphere” of an image can be characterized by the statistical distribution of latent features across channel dimensions. Unlike semantic features that focus on specific spatial structures, visual atmosphere is closer to a global style distribution. Therefore, we calculate the first-order moment (mean μ) and second-order moment (standard deviation σ) of Z in the spatial dimension:

$$\mu_c = \frac{1}{h \cdot w} \sum_{i=1}^h \sum_{j=1}^w Z_{c,i,j} \quad (3)$$

$$\sigma_c = \sqrt{\frac{1}{h \cdot w} \sum_{i=1}^h \sum_{j=1}^w (Z_{c,i,j} - \mu_c)^2 + \epsilon} \quad (4)$$

Finally, we concatenate the mean and standard deviation to form a compact visual context descriptor v_{ctx} :

$$v_{ctx} = \text{Concat}(\mu, \sigma) \in \mathbb{R}^{2C} \quad (5)$$

This descriptor (dimension 32) efficiently encodes the brightness, contrast, and texture statistical properties of the reference image, with negligible computational cost.

3.2. Fine-Grained Semantic Stream

To maximize editability while ensuring high-fidelity identity consistency, we do not simply treat identity features as a static vector but construct a hierarchical Semantic Stream Pipeline. Referring to advanced ID customization architectures, we further decouple this stream into three tightly coupled sub-modules.

ID Feature Extraction Module. This module aims to capture multi-level identity information from the reference image I_{ref} . We use a pre-trained Face Recognition Backbone as the encoder \mathcal{S} . To balance overall identity recognition and local facial geometric details, we extract Local Features from the intermediate layers and Global Features from the top layer of the network.

$$F_{raw} = \mathcal{S}(I_{ref}) = \{f_{global}, f_{local}\} \quad (6)$$

These raw features form the “semantic base” of the identity, covering facial topological structures but not yet aligned with the generative model’s text-image space.

ID Feature Fusion Module. To address the heterogeneity between the ID feature space and the diffusion model’s Latent Space, and since direct injection of raw features leads to reduced editability (i.e., “sticker effect”), we design a fusion projection network. This module maps F_{raw} to a semantic space aligned with the text encoder, producing Mapped

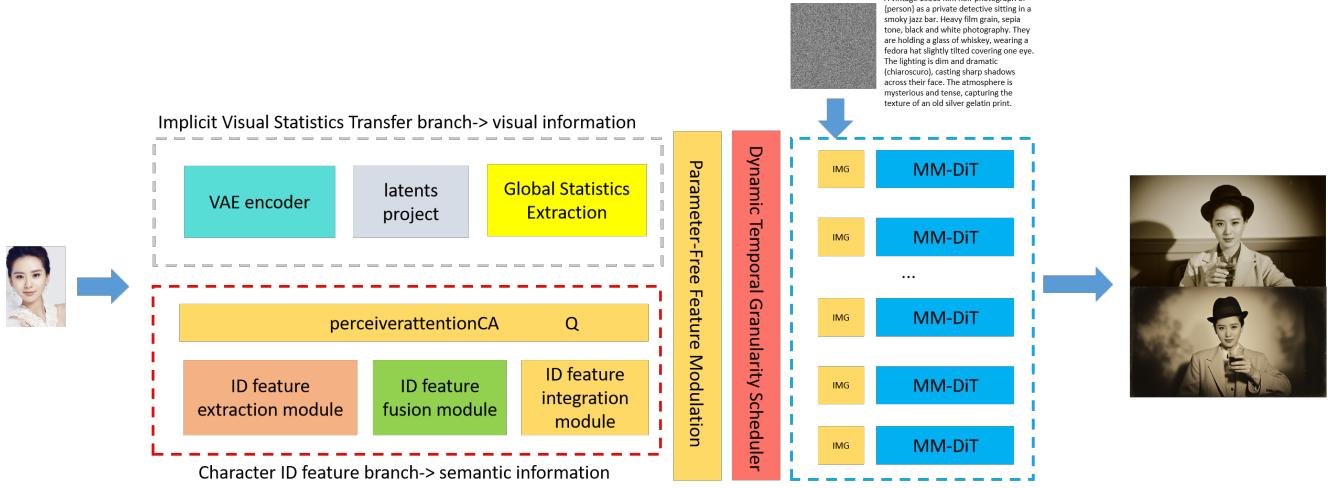


Figure 1. Overview of the DVI Framework. We disentangle identity customization into a person ID branch (bottom) for extracting semantic structure and an implicit visual statistics branch (top) leveraging VAE latent variables to extract atmospheric context. These heterogeneous features are synergistically fused via parameter-free feature modulation and regulated by a dynamic scheduler within the MM-DiT backbone.

Features. In this process, we introduce the concept of Offset Features, utilizing a learnable residual term to fine-tune feature direction, granting flexibility to adapt to different contexts while maintaining core ID attributes.

$$f_{id} = \text{Proj}(F_{raw}) + \delta_{offset} \in \mathbb{R}^{N \times D} \quad (7)$$

where f_{id} is the final aligned high-dimensional semantic embedding (in our setting $D = 2048$). This step is crucial as it ensures the ID features are “editable” rather than rigid pixel copies.

ID Feature Insertion Module. Finally, we inject f_{id} into the generation backbone of the DiT via Cross-Attention. To balance ID preservation and responsiveness to text instructions, we adopt a Dynamic Embedding Strength design. Unlike traditional fixed-weight injection, we dynamically adjust the injection strength α_l at different layers l of the DiT:

$$\text{Attn}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d}}\right)(V + \alpha_l \cdot f_{id}) \quad (8)$$

This design constitutes the “semantic anchor” of DVI. Notably, at this point f_{id} possesses perfect geometric structure and identity semantics but still lacks specific visual atmosphere. This is precisely where we introduce the Coarse-Grained Visual Stream for modulation in the next section.

3.3. Parameter-Free Feature Modulation (PFFM)

This is the core innovation of DVI. The challenge lies in fusing the low-dimensional visual statistics v_{ctx} (32-dim)

into the high-dimensional semantic embeddings f_{id} (2048-dim). Traditional approaches train a Multi-Layer Perceptron (MLP) for dimensional projection, but this violates the Tuning-Free premise.

Inspired by the principle of Adaptive Instance Normalization (AdaIN), we propose a Parameter-Free Feature Modulation mechanism. We treat visual statistics as a Global Distribution Bias to dynamically adjust the distribution of semantic embeddings in the feature space.

Dimension Broadcasting. First, since the dimension of v_{ctx} is much smaller than D , we extend it to the target dimension D via a Repeat/Tile operation, constructing the modulation vector m_{vis} :

$$m_{vis} = \text{Repeat}(v_{ctx}, \lceil D/2C \rceil)[:, D] \quad (9)$$

Distribution Injection. To inject visual atmosphere into semantic features, we perform a normalized modulation operation on f_{id} . Unlike AdaIN which requires learning scale factors γ and offset factors β , we directly use m_{vis} as an additive bias to offset the manifold distribution of semantic features:

$$f_{fused} = \text{Norm}(f_{id}) + \lambda(t) \cdot \Psi(m_{vis}) \quad (10)$$

where $\text{Norm}(\cdot)$ denotes Layer Normalization, Ψ is a simple scaling operator (coefficient set to 0.5 in implementation), and $\lambda(t)$ is a time-variant control coefficient. The physical meaning of this operation is: f_{id} determines the

“center position” of features in the semantic space (i.e., who the identity is), while m_{vis} provides a “directional offset” to this center pointing towards a specific visual style (i.e., in what lighting atmosphere).

3.4. Dynamic Temporal Granularity Scheduler

The diffusion generation process is inherently a Coarse-to-Fine denoising process: early time steps mainly determine global composition, tone, and lighting, while later time steps focus on refining local textures and facial details.

If we inject visual features with constant strength throughout the generation process, it might lead to excessive stylization destroying facial details. Therefore, we design a Dynamic Temporal Granularity Scheduler. We define the visual modulation strength weight $\lambda(t)$ to decay linearly with the denoising time step t (from 1.0 noise state to 0.0 clean state):

$$\lambda(t) = \lambda_{base} \cdot t \quad (11)$$

- **Phase 1 (High Noise, $t \rightarrow 1.0$):** $\lambda(t)$ is large. Visual stream dominates; the model uses VAE statistics to lay down the overall tone and compositional atmosphere.
- **Phase 2 (Low Noise, $t \rightarrow 0.0$):** $\lambda(t)$ approaches 0. Semantic stream dominates; the model focuses on using f_{id} to refine facial details, ensuring identity fidelity.

This dynamic balancing strategy ensures the organic co-existence of visual atmosphere and identity semantics in the temporal dimension, avoiding feature conflicts.

4. Experiments

4.1. Experimental Settings

Implementation Details. Our method is implemented based on the Flux.1 model. For the Fine-Grained Semantic Stream, we employ AntelopeV2 as the face recognition backbone and CLIP ViT-L/14 as the auxiliary image encoder, reusing the feature embedding and feature fusion module weights from PuLID. For the Coarse-Grained Visual Stream, we directly reuse Flux’s native VAE Encoder without loading any additional models. In the inference stage, the default sampling steps are set to $T = 25$, guidance to 4, and the sampler uses Euler. To leave sufficient feature modulation space for the visual stream, we set the ID injection weight to 0.8 (slightly lower than the Baseline’s 1.0). All experiments are completed on 4 NVIDIA A100 (80GB) GPUs. Our evaluation adopts the IBench framework proposed by EditID [13].

4.2. Qualitative Comparison

We compare DVI with PuLID (SDXL version [22]), PuLID (Flux Krea version), DreamO [21], UNO [26],

and UMO [2]. The base model for PuLID (SDXL) is SDXL_base_1.0, while all other models (including DVI) use Flux.1 dev. As shown in Figure 2, we input long text prompts containing strong stylistic descriptions. While maintaining character consistency, DVI achieves superior Visual Atmosphere Integration compared to PuLID; compared to UNO and DreamO, DVI not only avoids severe ID loss but also generates images with more natural texture and light-shadow interaction, strictly adhering to environmental instructions in the prompt.

Figure 2 displays four challenging stylized scenes: vintage film noir detective, classical oil painting portrait, horror suspense corridor, and backlit pastoral memory. In the first column (vintage film), the prompt explicitly requests “Heavy film grain, sepia tone... texture of an old silver gelatin print”. PuLID (SDXL) and DreamO generate faces that are too clean and sharp, presenting a jarring modern digital photo look forcibly desaturated, completely disjointed from the background’s grainy texture. In contrast, DVI successfully “injects” film grain into the facial skin, presenting a natural yellowish aged look, perfectly fitting the “old photo” narrative atmosphere. In the second column (classical oil painting), facing the description “Visible brush strokes, rich textures, cracking paint effect”, UNO generates a painting style but with severe identity drift. DVI, while precisely maintaining ID, renders facial brush stroke textures using the impasto method, appearing as if painted on canvas rather than a pasted photo.

In the third column (horror corridor) and fourth column (backlit pastoral), light-shadow interaction is key. In the third column, DreamO’s facial lighting is flat, failing to reflect the warm light source characteristics of the lantern; in the fourth column, PuLID’s face retains the white light of the original image without producing the “glowing halo effect” under backlight. DVI shows excellent lighting adaptability in both scenes: the face is naturally illuminated by warm candlelight in the horror scene, and shows rim light in the pastoral scene, demonstrating DVI’s precise capture of Visual Context and flexible reorganization of ID features in complex scenes with strong environmental light interference.

We further verified the above visual observations using metrics from the ChineseID + Editable Long Prompts benchmark, as shown in Table 1. DVI surpasses existing SOTA methods on multiple key metrics. On Aesthetic Quality & Visual Atmosphere, DVI achieves the highest scores in Aesthetic (0.700) and Image Quality (0.515) among all compared models. Compared to SDXL-based PuLID (0.675) and the latest PuLID-Krea (0.683) and DreamO (0.678), DVI’s significant improvement strongly proves the effectiveness of the Coarse-Grained Visual Stream. By introducing VAE latent space statistics, DVI successfully eliminates the “Semantic-Visual Dissonance” common in

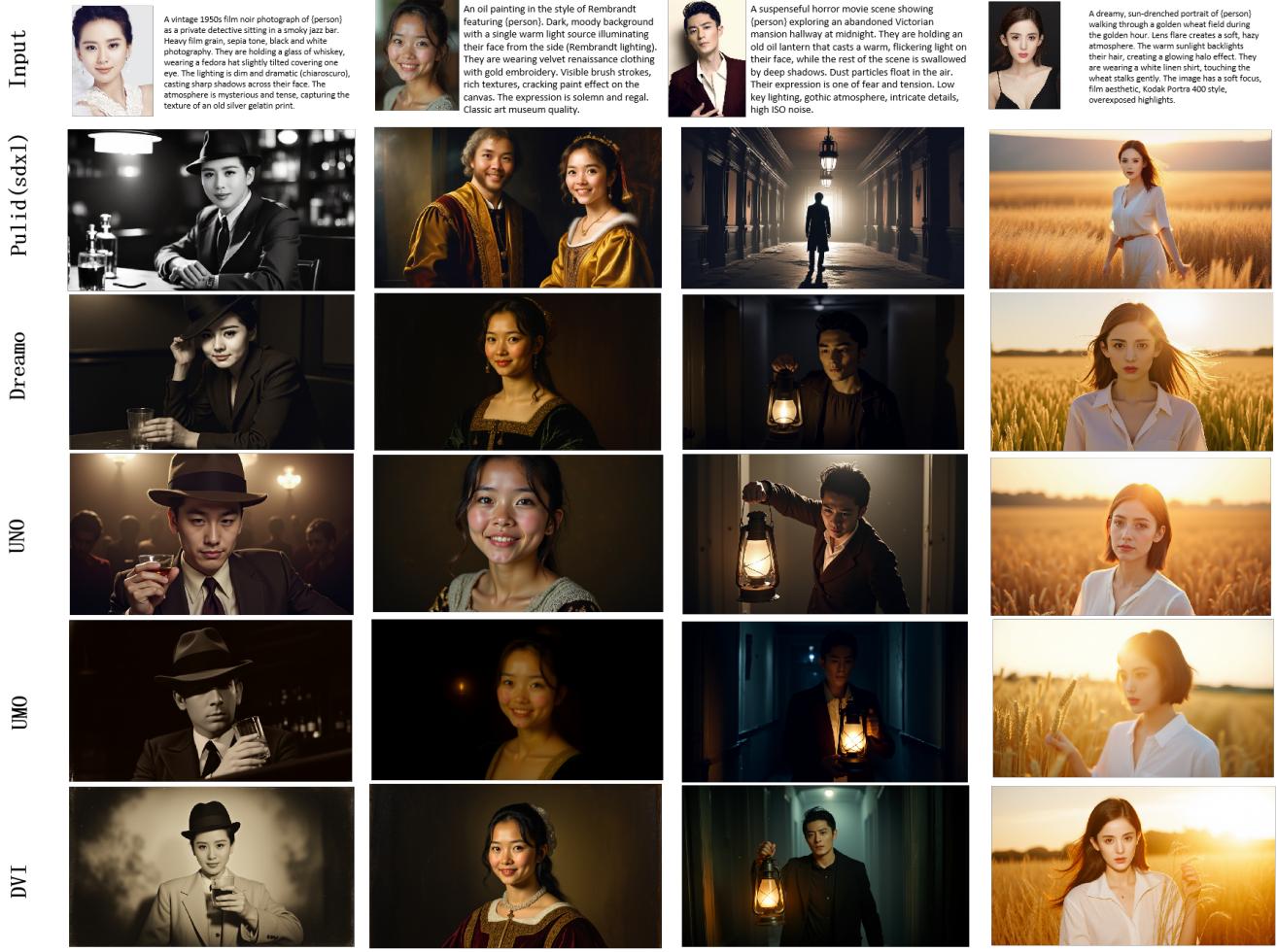


Figure 2. Qualitative Comparison: DVI achieves higher editability while ensuring ID consistency. DVI accurately realizes character consistency and complete presentation of atmospheric visual concepts in complex narrative scenes.

traditional ID injection methods. On the trade-off between Identity Preservation vs. Editability, DVI reaches 0.557 on the FaceSim metric, ranking first and significantly outperforming PuLiD-Krea (0.495) and DreamO (0.398). This indicates DVI’s semantic stream builds the most solid identity anchor. More importantly, while maintaining high ID fidelity, DVI retains extremely high Exprdiv (0.601), on par with UNO (0.614) which has strong style transfer but severe ID loss (FaceSim 0.105). This reflects DVI’s position at a “Golden Balance Point”: it does not sacrifice ID for editability like UNO, nor sacrifice expression flexibility for ID like traditional models. On Visual & Textual Consistency, DVI achieves 0.804 on ClipI, almost tying with the best DreamO (0.805), but DVI’s ID preservation capability (0.557) far exceeds DreamO (0.398). This indicates DVI’s PFFM mechanism successfully injects the “visual soul” of the reference image. Additionally, the high ClipT score (0.269) proves DVI does not sacrifice adherence to

text prompts by introducing the visual stream.

4.3. Ablation Study

4.3.1 Effectiveness of Coarse-Grained Visual Stream

To verify the key role of the Coarse-Grained Visual Stream in introducing style and atmospheric elements, we compared the full DVI model with a version removing the visual stream module.

Foreground-Background Detachment Phenomenon:

When the visual stream is removed (as shown in Figure 3 Left), the model relies solely on the fine-grained semantic stream. Although facial geometric structures are accurately restored, the face presents a “clarity ignoring the environment”. In a Film Noir scene emphasizing low light and grain, the face in the left image retains the high signal-to-noise ratio and uniform lighting of a modern digital photo, clashing with the dim, rough rainy night background.

Injection of Visual Atmosphere Elements: In con-

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

Model	Aesthetic	Image Quality	Exprdiv	Facesim	ClipI	ClipT
PuLID (SDXL)	0.675	0.502	0.593	0.399	0.768	0.248
PuLID (Krea)	0.683	0.505	0.587	0.495	0.793	0.277
DreamO	0.678	0.510	0.601	0.398	0.805	0.266
UNO	0.675	0.465	0.614	0.105	0.797	0.267
UMO	0.669	0.469	0.619	0.397	0.748	0.259
DVI (Ours)	0.700	0.515	0.601	0.557	0.804	0.269



Figure 3. Comparison between removing the visual stream (Left) and the full DVI model (Right). The left image retains ID but the face is too independent and clean, showing obvious Foreground-Background Detachment; the right image successfully injects atmospheric elements like film grain and low-key lighting via the visual stream.

trast, the full scheme introducing the coarse-grained visual stream (Figure 3 Right) successfully “injects” the style elements of the reference image into the identity features. Through VAE statistics modulation, the character’s face is no longer an independent semantic island but is endowed with more environmental atmospheric elements: facial lighting ratios are increased to match the oppressive feel of film noir, skin texture becomes rough to echo film grain, and edge contours soften to blend into the rainy mist.

4.3.2 Comparison of Feature Modulation Designs

We further compared the effects of Parameter-Free Feature Modulation (PFFM) versus simple Concatenation strategies. Under the Tuning-Free setting, simple concatenation often leads to dimensional abruptness in feature space, making it difficult for the pre-trained DiT model to adapt to the forced inclusion of heterogeneous features, often manifesting as “Distribution Mismatch”. This is particularly evident when handling “Double Exposure” art styles requiring high structural fusion. As shown in Figure 4 Left, images generated using the concatenation strategy show rigid structural conflicts at the junction of the portrait silhouette and

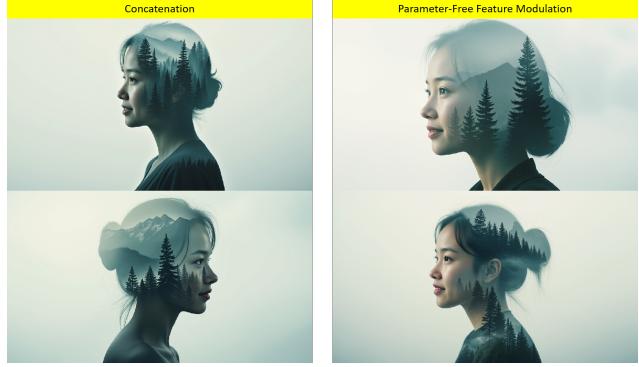


Figure 4. Comparison between simple feature Concatenation (Left) and DVI Parameter-Free Feature Modulation (PFFM, Right) under Double Exposure style. Concatenation leads to muddy tones and rigid edges; PFFM achieves smooth natural fusion.

the landscape (forest). Facial tones appear “Muddy Tones” and covered by shadow, and the transition between hair and trees lacks naturalness. In contrast, DVI’s distribution modulation strategy based on AdaIN demonstrates high robustness. This mechanism uses visual statistics as a global bias to dynamically shift and scale the distribution of semantic features. As shown in Figure 4 Right, PFFM achieves dreamlike “Organic Fusion”.

5. Conclusion

This paper proposes DVI, a zero-shot identity customization framework for text-to-image generation. Addressing the “Semantic-Visual Dissonance” caused by excessive reliance on semantic feature compression in existing Tuning-Free methods, we provide an elegant training-free solution. The core contributions of DVI lie in reformulating the identity customization task as the synergistic injection of dual-granularity features: 1. Constructing a hierarchical semantic stream to provide a solid identity geometric anchor and editability foundation. 2. Mining the inherent potential of the generative model’s latent space, using VAE statistics (mean and variance) as lightweight global visual descriptors. 3. Designing Parameter-Free Feature Modulation (PFFM) and Dynamic Temporal Scheduling to successfully

inject the “visual soul” (lighting, texture, atmosphere) of the reference image into semantic features without extra training costs. Experimental results show that DVI maintains robust identity consistency while significantly enhancing artistic beauty and environmental integration in high-complexity narrative scenes.

References

- [1] Bowen Chen, Mengyi Zhao, Haomiao Sun, Li Chen, Xu Wang, Kang Du, and Xinglong Wu. Xverse: Consistent multi-subject control of identity and semantic attributes via dit modulation. *arXiv preprint arXiv:2506.21416*, 2025. 1
- [2] Yufeng Cheng, Wenzhu Wu, Shaojin Wu, Mengqi Huang, Fei Ding, and Qian He. Umo: Scaling multi-identity consistency for image customization via matching reward. *arXiv preprint arXiv:2509.06818*, 2025. 2, 5
- [3] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019. 1, 2
- [4] 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. *arXiv preprint arXiv:2208.01618*, 2022. 1
- [5] Jawook Gu and Jong Chul Ye. Adain-based tunable cycle-gan for efficient unsupervised low-dose ct denoising. *IEEE Transactions on Computational Imaging*, 7:73–85, 2021. 3
- [6] 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. 1
- [7] Junjie He, Yifeng Geng, and Liefeng Bo. Uniportrait: A unified framework for identity-preserving single-and multi-human image personalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14399–14408, 2025. 1
- [8] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 1
- [9] Liming Jiang, Qing Yan, Yumin Jia, Zichuan Liu, Hao Kang, and Xin Lu. Infiniteyou: Flexible photo refracting while preserving your identity. *arXiv preprint arXiv:2503.16418*, 2025. 1
- [10] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 3
- [11] 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. 1
- [12] Guandong Li. Layout control and semantic guidance with attention loss backward for t2i diffusion model. *arXiv preprint arXiv:2411.06692*, 2024. 1
- [13] Guandong Li and Zhaobin Chu. Editid: Training-free editable id customization for text-to-image generation. *arXiv preprint arXiv:2503.12526*, 2025. 1, 5
- [14] Guandong Li and Zhaobin Chu. Editidv2: Editable id customization with data-lubricated id feature integration for text-to-image generation. *arXiv preprint arXiv:2509.05659*, 2025. 1
- [15] Guandong Li and Xian Yang. Smartbanner: intelligent banner design framework that strikes a balance between creative freedom and design rules. *Multimedia Tools and Applications*, 82(12):18653–18667, 2023. 1
- [16] 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. 1, 2
- [17] Yang Liu, Cheng Yu, Lei Shang, Yongyi He, Ziheng Wu, Xingjun Wang, Chao Xu, Haoyu Xie, Weida Wang, Yuze Zhao, et al. Facechain: A playground for human-centric artificial intelligence generated content. *arXiv preprint arXiv:2308.14256*, 2023. 1
- [18] Zichuan Liu, Liming Jiang, Qing Yan, Yumin Jia, Hao Kang, and Xin Lu. Learning joint id-textual representation for id-preserving image synthesis. *arXiv preprint arXiv:2504.14202*, 2025. 2
- [19] Zhendong Mao, Mengqi Huang, Fei Ding, Mingcong Liu, Qian He, and Yongdong Zhang. Realecustom++: Representing images as real-word for real-time customization. *arXiv preprint arXiv:2408.09744*, 2024. 2
- [20] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 4296–4304, 2024. 2
- [21] Chong Mou, Yanze Wu, Wenzhu Wu, Zinan Guo, Pengze Zhang, Yufeng Cheng, Yiming Luo, Fei Ding, Shiwen Zhang, Xinghui Li, et al. Dreamo: A unified framework for image customization. *arXiv preprint arXiv:2504.16915*, 2025. 1, 2, 5
- [22] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 5
- [23] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1
- [24] 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. 1
- [25] 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. 1, 2
 - [26] Shaojin Wu, Mengqi Huang, Wenzhu Wu, Yufeng Cheng, Fei Ding, and Qian He. Less-to-more generalization: Unlocking more controllability by in-context generation. *arXiv preprint arXiv:2504.02160*, 2025. 2, 5
 - [27] Guangxuan Xiao, Tianwei Yin, William T Freeman, Frédo Durand, and Song Han. Fastcomposer: Tuning-free multi-subject image generation with localized attention. *International Journal of Computer Vision*, 133(3):1175–1194, 2025. 1
 - [28] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023. 2
 - [29] Lvmnin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3836–3847, 2023. 2