PIXELSMILE: TOWARD FINE-GRAINED FACIAL EXPRESSION EDITING

Anonymous authors
Paper under double-blind review

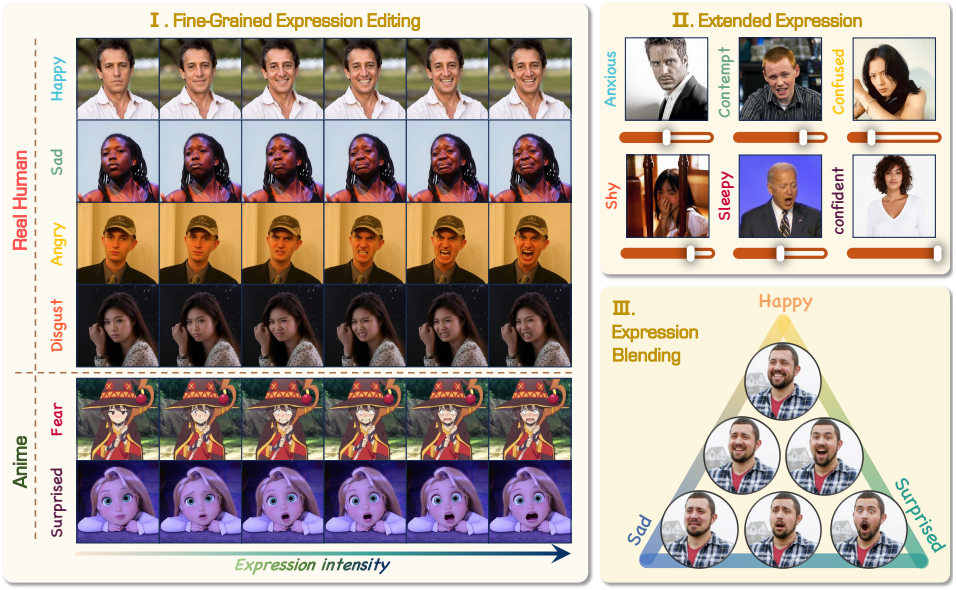


Figure 1: **Overview of PixelSmile**. It enables 1) continuous and precise control of facial expression intensity across real-world and anime domains, 2) editing across 12 distinct expression categories, and 3) seamless blending of multiple expressions.

ABSTRACT

Fine-grained facial expression editing is limited by intrinsic semantic overlap. To address this, we construct the **Flex Facial Expression** (FFE) dataset with continuous affective annotations and establish **FFE-Bench** to evaluate structural confusion, editing accuracy, linear controllability, and the trade-off between expression editing and identity preservation. We propose **PixelSmile**, a diffusion framework that disentangles expression semantics via fully symmetric joint training. PixelSmile combines intensity supervision with contrastive learning to produce stronger and more distinguishable expressions, achieving precise and stable linear expression control through textual latent interpolation. Extensive experiments demonstrate that PixelSmile achieves superior disentanglement and identity preservation, enabling fine-grained continuous expression editing and smooth expression blending. Our project page is available here.

1 INTRODUCTION

Recent advances in diffusion-based image editing and identity-consistent generation (Xu et al., 2025a; Guo et al., 2024; Jiang et al., 2025) have greatly improved natural-language portrait manipulation. Despite this progress, fine-grained facial expression editing remains challenging. Current models can distinguish expressions such as happy and sad, but struggle with semantically overlapping pairs such as fear and surprise or anger and disgust. Most methods rely on discrete categories, forcing continuous human expressions into rigid class boundaries. Consequently, they blur expression boundaries, leading to structured cross-category confusion, unstable intensity control, and identity degradation.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

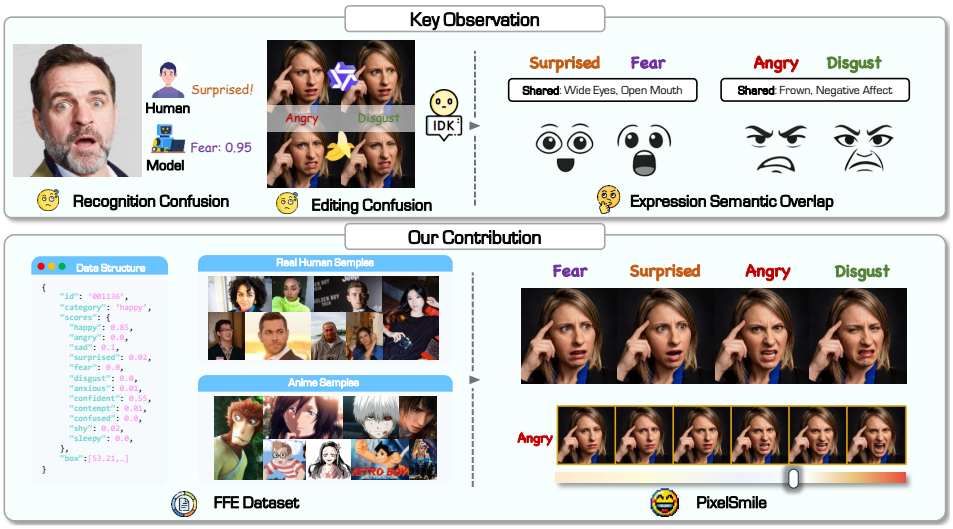


Figure 2: **Observation of Expression Semantic Overlap.** Inherent expression overlap causes systematic confusion across human annotators, recognition models, and generative models (top). We resolve this via the FFE dataset (bottom left) and PixelSmile framework (bottom right), utilizing continuous supervision and symmetric training for disentangled editing.

To understand this limitation, we examine the semantic structure of facial expressions. As illustrated in Fig. 2, facial expressions lie on a continuous semantic manifold where adjacent emotions naturally overlap. This overlap causes systematic confusion across human annotators, classifiers, and generative models, which often fail to distinguish adjacent expressions such as fear and surprise or anger and disgust. Training generative models on ambiguous samples with discrete, potentially conflicting labels leads to entangled latent representations. This entanglement hinders precise control and causes expression leakage, where editing one emotion inadvertently introduces characteristics of another or degrades identity consistency.

Addressing this challenge requires moving beyond rigid one-hot supervision, which fails to capture the nuanced structure of human affect and propagates semantic entanglement into generative models. We therefore construct the **Flex Facial Expression** (FFE) dataset with continuous 12-dimensional affective score vectors. We further establish **FFE-Bench** to evaluate structural confusion, editing accuracy, linear controllability, and the editing–identity trade-off. Across real and anime domains, FFE provides diverse same-identity expressions with continuous annotations, enabling models to learn fine-grained expression boundaries and supporting systematic evaluation of controllable editing.

Building upon this data-centric foundation, we propose **PixelSmile**, a diffusion-based framework for disentangling expression semantics. It introduces Fully Symmetric Joint Training to contrast the confusing expression pairs identified above. Combined with Textual Latent Interpolation, it enables precise, linear intensity control at inference without reference images. Together, continuous affective supervision and symmetric learning enable robust, identity-preserving editing.

In summary, our contributions are threefold:

- **Systematic Analysis of Semantic Overlap.** We reveal and formalize structured semantic overlap among facial expressions, showing that it, rather than mere classification error, is a key source of failure in both recognition and generative editing.
- **Dataset and Benchmark.** We construct FFE, a 60k cross-domain dataset covering 12 expression categories with continuous affective annotations, and establish FFE-Bench to evaluate structural confusion, accuracy, linearity, and the editing–identity trade-off.
- **PixelSmile Framework.** We propose a diffusion-based framework combining Fully Symmetric Joint Training and Textual Latent Interpolation. It disentangles expression semantics and enables fine-grained, identity-preserving editing with linear control.

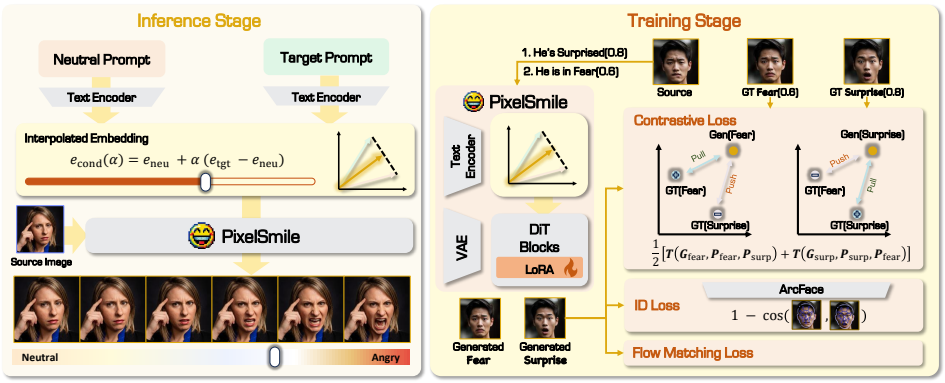


Figure 3: **Framework Overview.** (1) **Inference Stage**. We interpolate neutral and target textual embeddings with coefficient α for continuous expression intensity control. (2) **Training Stage**. Fully Symmetric Joint Training exchanges positive and negative roles within confusing expression pairs, combining Flow Matching, contrastive learning, and identity supervision for intensity alignment, expression separation, and identity preservation.

2 Related Work

Facial Expression Editing. Early facial expression editing methods used conditional GANs (Goodfellow et al., 2014) for multi-domain translation (Choi et al., 2018; Pumarola et al., 2018; Liu et al., 2019; Ding et al., 2018) or manipulated StyleGAN latent representations (Karras et al., 2019; 2020; Shen et al., 2020; Härkönen et al., 2020; Shen & Zhou, 2021; Yüksel et al., 2021). Other approaches use facial priors, such as Action Units (AUs) or 3DMM parameters, for interpretable control (Wei et al., 2025; Pumarola et al., 2018; Ding et al., 2023; Jang et al., 2025; Feng et al., 2021; Daněček et al., 2022). Their reliance on human facial structure limits transfer to stylized characters and the representation of high-level affective semantics. More recently, diffusion models (Ho et al., 2020) and multimodal pretraining have enabled text-driven editing, allowing users to specify target expressions in natural language (Hertz et al., 2022; Brooks et al., 2023). Models such as GPT-Image (OpenAI, 2025), Nano Banana Pro (Raisinghani, 2025), Qwen-Image (Wu et al., 2025), and LongCat-Image (Team et al., 2025) offer strong editing and identity preservation (ByteDance, 2025; Labs, 2025; Liu et al., 2025b), but lack explicit mechanisms to distinguish semantically overlapping expressions. Our approach explicitly models expression semantics for fine-grained editing across realistic and stylized portraits.

Continuously Controlled Generation. Prior works achieve continuous editing through interpolatable generative subspaces: ConceptSlider (Gandikota et al., 2024) interpolates LoRA weights, while other methods (Baumann et al., 2025; Garibi et al., 2025; Zhong et al., 2025; Kamenetsky et al., 2025; Jain et al., 2025; Ye et al., 2025; Dalva et al., 2024; Gandikota et al., 2025) manipulate text embeddings or modulation features. Recent methods, including SliderEdit (Zarei et al., 2025), Kontinuous-Kontext (Parihar et al., 2025), and concurrent works (Wolf et al., 2026; Xu et al., 2025b; Yin et al., 2025), extend this capability to FLUX.1 Kontext (Labs et al., 2025). Although reduced-strength editing and pixel interpolation may smooth transitions, entangled latent spaces still cause semantic ambiguity and identity drift under strong control. In contrast, our formulation models expression semantics for fine-grained linear control while preserving identity across strengths.

Facial Expression Datasets and Benchmarks. Existing expression datasets do not fully meet the data requirements of fine-grained, continuously controllable editing: in-the-wild datasets (Mollahosseini et al., 2017; Li et al., 2017) lack same-identity expression pairs, video datasets (Qiu et al., 2025; Zhang et al., 2025) emphasize temporal or cross-modal dynamics, and MEAD (Wang et al., 2020) provides only three discrete intensity levels. Our FFE combines diverse same-identity expression pairs across realistic and anime domains with continuous expression annotations, supporting fine-grained continuous control. Benchmarks such as F-Bench (Liu et al., 2025a) and SEED (Zhu et al., 2025), and common metrics such as CLIP, SSIM, and LPIPS do not jointly characterize expression confusion, control linearity, and the editing–identity trade-off. Our FFE-Bench jointly evaluates structural confusion, editing accuracy, linear control, and the editing–identity trade-off.

3 DATASET AND BENCHMARK

To support fine-grained, linearly controllable facial expression editing, we introduce FFE and the dedicated FFE-Bench. The dataset combines diverse same-identity expression variations with continuous affective annotations, while the benchmark evaluates semantic disentanglement, control linearity, and the balance between expression editing and identity preservation.

3.1 THE FFE DATASET

Facial-Attribute-Guided Generation. FFE contains 60K facial images (30K each in real and anime domains), covering 12 expressions with continuous annotations. Starting from diverse identities curated from public sources (snmahsa, 2021; aisegmentcn, 2020), we generate same-identity expression variations using *Nano Banana Pro*. We observe that expression names alone can yield ambiguous edits, whereas descriptions of local facial attributes provide more precise guidance. We therefore build a prompt library organized by expression and intensity, combining category names with descriptions of mouth shape, eyebrow movement, and eye openness.

Continuous Annotation and Validation. *Gemini 3 Pro* then assigns each generated image a continuous expression vector $\mathbf{v} \in [0,1]^{12}$. Each dimension records the intensity of one expression, allowing overlapping expressions to be represented beyond a single discrete label. The generation target serves as a consistency reference rather than being directly assigned as the annotation. We cross-check the predicted expression against the generation target: samples with large discrepancies undergo manual review, while consistent samples are randomly spot-checked to reduce generation and annotation noise. Generation targets, image-based predictions, and human judgments thus provide complementary evidence for sample selection. These checks screen for inconsistent edits and annotations; agreement with the target alone does not establish intensity-score reliability. We therefore additionally compare the continuous scores with human judgments on a sampled subset spanning all expression categories and varying intensities, obtaining Pearson and Spearman correlations of 0.79 and 0.81, respectively. Construction and validation details are provided in Appendix F.

3.2 THE FFE-BENCH

To capture structured cross-category confusion caused by semantic entanglement among facial expressions, we evaluate facial expression editing from four complementary aspects: structural confusion, editing accuracy, control linearity, and the expression–identity trade-off. All expression classifications and intensity scores are predicted by Gemini 3 Pro.

Mean Structural Confusion Rate (mSCR). $C_{i,j}$ measures directed confusion from expression i to j, where N_i is the number of samples edited toward class i, and $\hat{y}_k^{(i)}$ is the predicted dominant expression. The mSCR averages the bidirectional confusion rate, $\mathrm{BCR}(i,j)$, over predefined confusing pairs. Lower mSCR indicates less cross-category confusion and better semantic disentanglement.

Expression Editing Accuracy (Acc). Acc is the proportion of generated images whose predicted dominant expression matches the target instruction. Acc measures target correctness, complementing the pairwise confusion captured by mSCR.

Control Linearity Score (CLS). To evaluate continuous controllability, we use coefficients α at inference and compute their Pearson correlation with VLM-predicted intensity scores. Higher CLS indicates more linear control.

Harmonic Editing Score (HES). HES balances expression strength and identity preservation through their harmonic mean in Eq. 1, where S_E is the target expression score and S_{ID} is the mean source–edited face cosine similarity. The identity evaluation models are specified in Appendix B. HES is high only when both expression strength and identity fidelity are preserved.

The confusion rates and harmonic editing score are defined as

$$C_{i,j} = \frac{1}{N_i} \sum_{k=1}^{N_i} \mathbf{1}(\hat{y}_k^{(i)} = j), \quad \mathrm{BCR}(i,j) = \frac{C_{i,j} + C_{j,i}}{2}, \quad \mathrm{HES} = \frac{2 S_E S_{\mathrm{ID}}}{S_E + S_{\mathrm{ID}}}. \tag{1}$$

Table 1: **Quantitative Evaluation of General Editing Models**. Top three results are highlighted in , , and , respectively.

Method	mSCR ↓	Acc-6 ↑	Acc-12 ↑	ID Sim ↑
Seedream (ByteDance, 2025)	0.3725	0.5294	0.3737	0.7221
NBP (Raisinghani, 2025)	0.1754	0.8431	0.6200	0.7107
GPT (OpenAI, 2025)	0.1107	0.8039	0.6300	0.5056
FLUX (Labs, 2025)	0.2850	0.4510	0.3310	0.4146
LongCat (Team et al., 2025)	0.1754	0.6275	0.4100	0.6036
Qwen (Wu et al., 2025)	0.2625	0.4510	0.2900	0.6938
Ours (w/o training)	0.2400	0.5294	0.3500	0.6769
Ours	0.0550	0.8627	0.6000	0.6522

Table 2: **Quantitative Evaluation of Linear Control Models**. Top three results are highlighted in , , and , respectively.

Method	CLS-6 ↑	CLS-12 ↑	ID Sim ↑	**HES**↑
SAEdit (Kamenetsky et al., 2025)	-0.0183	0.0007	-	-
ConceptSlider (Gandikota et al., 2024)	0.3161*	-	0.6250*	0.3656*
AttributeControl (Baumann et al., 2025)	0.2856*	-	0.3609*	0.2712*
K-Slider (Parihar et al., 2025)	-0.0459	-0.0634	0.7974	0.3272
SliderEdit (Zarei et al., 2025)	0.5599	0.5217	0.7414	0.3441
Ours (w/o training)	0.6892	0.5217	0.6769	0.4086
Ours	0.8078	0.7305	0.6522	0.4723

* Results use only two expression classes: happy and surprised.

4 METHOD

As illustrated in Fig. 3, PixelSmile is a fine-grained facial expression editing framework built on a pretrained Multi-Modal Diffusion Transformer (MMDiT) (Esser et al., 2024) with LoRA adaptation (Hu et al., 2022). To disentangle expression semantics and enable continuous intensity control, we introduce two components: (1) a textual latent interpolation mechanism for smooth intensity control; and (2) a fully symmetric joint training framework using symmetric contrastive learning to reduce cross-category confusion while preserving identity consistency.

4.1 TEXTUAL LATENT INTERPOLATION

Given a neutral prompt P_{neu} and a target expression prompt P_{tgt}, the frozen MMDiT text encoder maps them to embeddings e_{neu} and e_{tgt}, respectively. We define the residual direction and continuous conditioning embedding as

$$e_{\text{cond}}(\alpha) = e_{\text{neu}} + \alpha \cdot \Delta e, \quad \Delta e = e_{\text{tgt}} - e_{\text{neu}}, \tag{2}$$

where Δe captures the semantic shift from neutral to the target expression. At $\alpha = 0$, the conditioning is neutral, while $\alpha = 1$ recovers the full target expression. Intermediate values smoothly vary expression intensity, and extrapolation with $\alpha > 1$ strengthens the expression.

Intensity-Aware Flow Matching. To align textual interpolation with visual intensity, we incorporate continuous expression intensity supervision into Flow Matching (Lipman et al., 2023) training. Each training image has a ground-truth intensity coefficient $\alpha_{\text{gt}} \in [0, 1]$ derived from its continuous expression annotations. During LoRA fine-tuning, we set $\alpha = \alpha_{\text{gt}}$ and feed $e_{\text{cond}}(\alpha)$ into the dual-stream attention blocks. The score-supervised velocity loss is

$$\mathcal{L}_{\text{FM}} = \mathbb{E}_{t,x_0,x_1}\left[\left\|v_\theta(x_t, t, e_{\text{cond}}(\alpha)) - (x_1 - x_0)\right\|_2^2\right], \tag{3}$$

where x_0 and x_1 denote the source and edited target latents, respectively. This objective aligns the interpolation coefficient with visual intensity, enabling reference-free continuous control by varying α at inference.

4.2 FULLY SYMMETRIC JOINT TRAINING

As stated in Sec. 1 and illustrated in Fig. 2, facial expressions form a continuous, overlapping semantic manifold. For example, *Surprise* and *Fear* share similar facial cues, causing structural confusion near class boundaries under discrete supervision. We introduce Fully Symmetric Joint Training with symmetric feature-space contrastive learning.

Symmetric Construction. Given a semantically overlapping expression pair (E_a, E_b), the model generates G_a and G_b from the same input image, conditioned on E_a and E_b, respectively. For the G_a branch, the ground-truth images P_a and P_b, corresponding to E_a and E_b, serve as the positive and hard negative, respectively; their roles are reversed for the G_b branch. This symmetric design avoids directional bias and consistently separates confusing expressions.

Symmetric Contrastive Loss. A frozen CLIP image encoder extracts expression-semantic features from all images. The symmetric contrastive loss is defined as

$$\mathcal{L}_{\text{SC}} = \frac{1}{2}\left[\mathcal{T}(G_a, P_a, P_b) + \mathcal{T}(G_b, P_b, P_a)\right], \tag{4}$$

where $\mathcal{T}$ pulls each generated sample toward its target and away from the confusing expression. We consider three formulations of $\mathcal{T}$: hinge-based (Schroff et al., 2015), log-ratio (Oh Song et al., 2016), and InfoNCE-style (Oord et al., 2018). We adopt the InfoNCE-style objective for its optimization stability. Detailed formulations and ablations are provided in Appendices A and D, respectively.

Identity Preservation. Strong extrapolation ($\alpha > 1$) or contrastive learning may degrade identity consistency. To preserve biometric features, we introduce an identity loss using ArcFace (Deng et al., 2019) as the frozen encoder $\Phi_{\text{arc}}(\cdot)$. For generated images G_a, G_b and corresponding targets P_a, P_b, the loss is

$$\mathcal{L}_{\text{ID}} = \frac{1}{2} \sum_{i \in \{a,b\}} \left[1 - \cos\big(\Phi_{\text{arc}}(G_i), \Phi_{\text{arc}}(P_i)\big)\right]. \tag{5}$$

4.3 Overall Training Objective

We fine-tune the MMDiT LoRA parameters using symmetric dual-branch training, jointly optimizing a confusing expression pair (a, b). The overall objective is

$$\mathcal{L}_{\text{total}} = \frac{1}{2}\big(\mathcal{L}^a_{\text{FM}} + \mathcal{L}^b_{\text{FM}}\big) + \lambda_{\text{sc}}\mathcal{L}_{\text{SC}} + \lambda_{\text{id}}\mathcal{L}_{\text{ID}}, \tag{6}$$

where λ_{sc} and λ_{id} balance expression disentanglement and identity preservation. This objective jointly promotes continuous intensity control, expression separation, and identity consistency.

5 Experiments

5.1 Experimental Setup

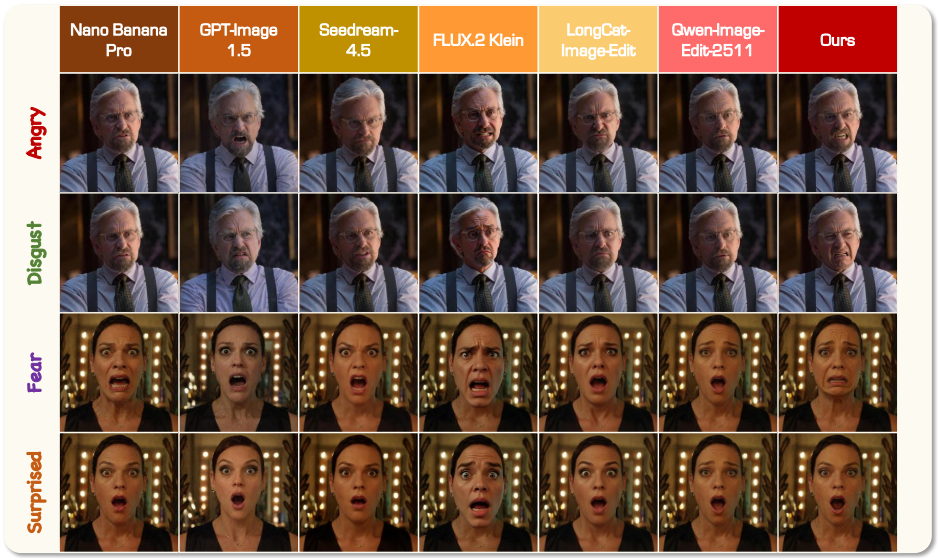
We implement PixelSmile based on Qwen-Image-Edit-2511. We train separate LoRA adapters for the real-world and anime domains. Domain-specific encoders are used for contrastive supervision and identity preservation. Specifically, contrastive supervision uses CLIP-ViT-L/14 (Radford et al., 2021) for real-world images and DanbooruCLIP (OysterQAQ, 2023) for anime. ArcFace identity supervision is applied only to the real-world adapter; the anime adapter is trained without the identity loss. Additional implementation details are provided in Appendix B.

Baselines. We divide the baselines into two groups: general editing models for overall editing quality and linear control models for continuous, predictable intensity control. **Group 1: General Editing Models.** This group includes strong general-purpose text-guided editing systems: three closed-source models, Nano Banana Pro (NBP), GPT-Image-1.5 (GPT), and Seedream-4.5 (Seedream); and three open-source models, Qwen-Image-Edit-2511 (Qwen), FLUX.2 Klein (FLUX), and LongCat-Image-Edit (LongCat). We use the abbreviations in parentheses hereafter. **Group 2: Linear Control Models.** We compare with recent control-oriented methods, including Kontinuous-Kontext (K-Slider), SliderEdit, and SAEdit, as well as earlier latent-control methods, ConceptSlider and AttributeControl. Earlier methods are limited by predefined attribute categories and inversion-induced information loss. We therefore treat them as reference baselines rather than primary competitors in multi-category evaluation. We also evaluate a training-free variant, denoted as Ours (w/o training), which applies textual latent interpolation directly to the pretrained Qwen-Image-Edit-2511 backbone without LoRA fine-tuning.

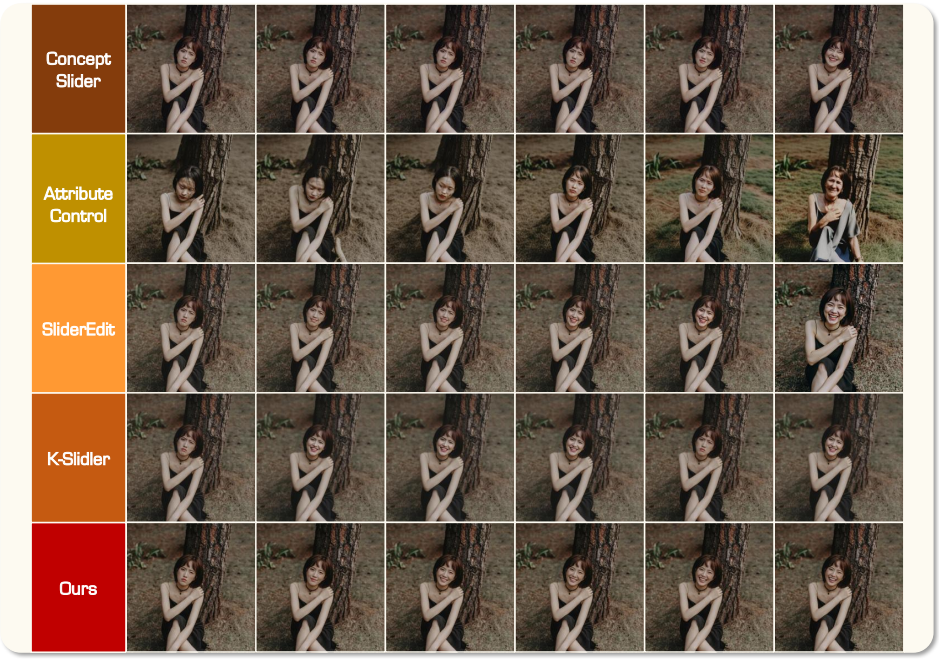
Metrics. Following Sec. 3.2, Group 1 is evaluated on editing accuracy and expression disentanglement using Acc-6, Acc-12, and mSCR, while Group 2 is evaluated on linear intensity control and identity fidelity using CLS-6, CLS-12, and HES. We also report ID Sim, computed by averaging source–edited face cosine similarities across ArcFace (Deng et al., 2019), AdaFace (Kim et al., 2022), and FaceNet (Schroff et al., 2015). This average serves as S_{ID} in HES, whereas training uses only ArcFace for identity supervision.

5.2 Quantitative Evaluation

Tables 1 and 2 quantitatively compare PixelSmile with Group 1 and Group 2 models, respectively.



(a) Comparison with general editing models



(b) Comparison with linear control models

Figure 4: **Qualitative Comparisons.** Compared with general editing models (top), PixelSmile achieves clearer expression changes with improved identity consistency; compared with linear control models (bottom), it produces smooth and monotonic expression transitions while better preserving facial identity. Full qualitative comparisons are provided in Appendix C.

Evaluation with Group 1. For the six basic expressions, PixelSmile achieves the highest Acc-6 of 0.86, outperforming NBP (0.84) and GPT (0.80), and remains competitive on Acc-12. More importantly, PixelSmile obtains the lowest mSCR of 0.055, outperforming GPT (0.111) and NBP (0.175), while most other models exceed 0.20. Values near 0.5 indicate collapse of a confusing pair into one expression. For ID Sim, prior observations (Xu et al., 2025a) place realistic facial editing around 0.6–0.7; scores above 0.8 may indicate rigid "copy-paste" behavior, while scores below 0.5 suggest severe distortion. Among the baselines, Seedream retains high ID Sim but shows substantial structural confusion due to weak edits, whereas FLUX falls below 0.5 and severely degrades identity.

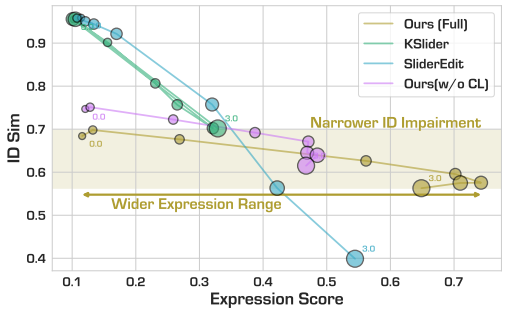


Figure 5: **Quantitative Evaluation of Linear Control Methods**. Comparison of the trade-off between ID similarity and expression score across different models. PixelSmile achieves an optimal balance, providing a wider expression range while preserving identity fidelity.

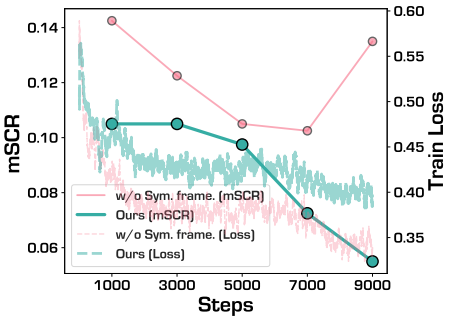


Figure 6: **Ablation on Symmetric Training Dynamics.** The asymmetric variant reduces loss faster initially but leads to higher structural confusion, while the symmetric framework achieves lower and more stable mSCR.

PixelSmile produces strong expressions within the natural ID Sim range, better balancing expression strength and identity preservation.

Evaluation with Group 2. ConceptSlider and AttributeControl support only two expression attributes, happy and surprised, and produce weak edits; we therefore report them as reference baselines using partial metrics such as CLS-2. As a text-to-image method without explicit identity preservation, SAEdit is included only as a quantitative reference. Table 2 and Fig. 5 show that zero-shot textual latent interpolation with Qwen outperforms existing control baselines, achieving CLS-6 of 0.69 and HES of 0.41. Fully Symmetric Joint Training further improves CLS-6 to 0.81, CLS-12 to 0.73, and HES to 0.47, highlighting the importance of explicit expression-semantic modeling. The high average ID similarity of K-Slider and SliderEdit largely results from weak edits, with values near 1.0 at low intensities. K-Slider exhibits negative CLS and irregular intensity responses below ~ 0.3, indicating failed linear control. SliderEdit increases expression intensity, but its ID similarity drops rapidly to ~ 0.4 as the expression score approaches 0.5, revealing a strong strength–identity trade-off. In contrast, PixelSmile increases expression intensity monotonically up to a peak score approaching 0.8 while maintaining ID similarity within the natural 0.6–0.7 range, demonstrating stable control and a better balance between expression strength and identity preservation.

5.3 Qualitative Comparison

Fig. 4 shows representative qualitative comparisons; full results appear in Appendix Figs. 7 and 8.

Comparison with Group 1. As shown in Fig. 4(a), existing models struggle to balance clear expression editing and identity preservation. NBP, Qwen, Seedream, and LongCat preserve identity but produce weak, often barely noticeable expression changes. GPT produces clearer expression changes but introduces moderate identity drift. FLUX performs poorly in both aspects, with weak edits and severe identity degradation. In contrast, PixelSmile produces clear, recognizable expressions while preserving facial identity, achieving a better editing–identity balance.

Comparison with Group 2. Fig. 4(b) compares continuous control for the representative *Happy* expression. Ideal control should increase expression intensity monotonically while preserving identity. ConceptSlider and AttributeControl show limited linear responses and increasing identity degradation at higher strengths. SliderEdit exhibits step-like control, with little change at most values followed by an abrupt increase and substantial identity degradation. K-Slider is unstable, with expression changes weakly correlated with the control parameter. For the more challenging *Surprised* expression, these baselines show further degraded linearity and identity preservation, as shown in Appendix Fig. 8. In contrast, PixelSmile maintains a stable monotonic increase in expression intensity while preserving identity across the entire control range. Even for more difficult expressions such as *Disgust*, our method produces clear and controllable expression changes.

Table 3: **Ablation Study**. Top three results are highlighted in , , and , respectively.

	Ablation	mSCR ↓	Acc-6 ↑	Acc-12 ↑	CLS-6 ↑	CLS-12 ↑	HES ↑	ID Sim ↑
Loss	w/o Contrastive Loss	0.2725	0.6471	0.4500	0.6978	0.5889	0.4500	0.7018
	w/o ID Loss	0.0550	0.8824	0.6500	0.8215	0.6874	0.4451	0.5749
	w/o Sym. Frame.	0.1350	0.7843	0.4700	0.7939	0.6488	0.4253	0.6402
Constraint	w/ Log-Ratio Constraint	0.1750	0.8039	0.5300	0.7917	0.6546	0.4933	0.6943
	w/ Hinge Constraint	0.0950	0.8824	0.6600	0.7997	0.7228	0.4758	0.6280
Data	MEAD	0.2125	0.7647	0.4700	0.7047	0.6187	0.4235	0.5735
Ours	Full Setting	0.0550	0.8627	0.6000	0.8078	0.7305	0.4723	0.6522

5.4 Ablation Study

Table 3 presents ablation results for the main components of PixelSmile.

Loss Components. Removing the identity loss improves editing-related metrics but substantially degrades identity preservation, particularly at large editing intensities. Conversely, removing the contrastive loss yields high identity similarity but weak expression edits and severe structural confusion: the model tends to reconstruct the source rather than separate overlapping expressions. The losses thus complement each other (Appendix Figs. 12 and 11).

Symmetric Training. We compare the full framework with an asymmetric variant that applies contrastive supervision to only one branch. As shown in Fig. 6, the asymmetric variant reduces training loss faster initially but converges to lower editing accuracy and higher structural confusion. The symmetric design converges more gradually, while bidirectional supervision stabilizes optimization and improves expression disentanglement.

Contrastive Objectives. Log-ratio favors identity preservation but weakens expression editing, whereas Hinge favors editing strength at the cost of identity consistency. InfoNCE achieves the best overall balance and is adopted as the default objective. Details are provided in Appendix A.

Training Dataset. The same architecture trained on MEAD (Wang et al., 2020) under identical optimization settings underperforms its FFE-trained counterpart across all metrics, demonstrating the importance of diverse identities and continuous expression annotations. MEAD preprocessing and additional ablation details are provided in Appendix D.

5.5 Further Analysis

User Study. We conduct a user study on 2,400 images with 10 trained annotators. Annotators rank the three methods by expression continuity and identity consistency over complete editing sequences; these rankings are converted into scores. The mean continuity and identity scores are PixelSmile (4.48, 3.80), K-Slider (1.36, 4.06), and SliderEdit (3.16, 1.14), showing that PixelSmile achieves the highest continuity while maintaining strong identity consistency (Appendix E).

Expression Blending. To examine compositionality in the learned expression space, we interpolate six basic expressions pairwise, producing 15 zero-shot combinations. Nine yield plausible compound expressions, demonstrating that the learned representation supports continuous composition. Semantically overlapping pairs such as Fear–Surprise collapse into a single dominant expression, whereas conflicting pairs such as Angry–Happy produce unstable results. These outcomes reveal compositional structure constrained by semantic overlap and conflicting facial cues (Appendix C).

6 Conclusion

We presented PixelSmile, a framework for fine-grained facial expression editing that disentangles overlapping expression semantics through continuous affective supervision and Fully Symmetric Joint Training. We also introduced FFE, a cross-domain dataset with continuous expression annotations, and FFE-Bench for evaluating structural confusion, editing accuracy, linear controllability, and identity preservation. Experiments show that PixelSmile reduces semantic confusion, improves intensity control, and better balances expression strength with identity fidelity, supporting continuous and semantically grounded facial affect manipulation.

AI USE STATEMENT

We used generative AI tools to assist with translation, drafting, and refinement of the manuscript. For dataset construction, Nano Banana Pro generated expression-edited images, Gemini 3 Pro provided continuous expression annotations, and Qwen3-VL-235B-A22B extracted attributes for dataset characterization. Gemini 3 Pro also provided expression predictions for benchmark evaluation. These uses are described in Section 3 and Appendix F. Generated samples and annotations underwent consistency checks, manual review, and sampled human validation as described therein. The authors have reviewed AI-assisted manuscript content and take responsibility for the final text, analyses, and results.

ETHICS STATEMENT

Data Sources and Usage. All data in FFE are collected from publicly available sources and used in compliance with their respective licenses and terms of use. The real-world subset is derived from the Human Images Dataset (snmahsa, 2021) and the Matting Human Dataset (aisegmentcn, 2020), both distributed under the MIT License. We follow the corresponding license and attribution requirements. The anime subset consists of stylized fictional characters from publicly available anime productions. We do not claim ownership of the underlying source content, which remains with its respective rights holders. The data are used for non-commercial academic research, and we do not collect or use private or login-restricted data.

Privacy and Data Processing. For the real-world subset, associated personal metadata is removed during processing. Each identity is represented by an internal numeric index, without retaining names, account information, or other associated personal identifiers. The dataset is intended for facial expression generation and editing rather than identification, profiling, or re-identification.

Potential Risks and Mitigation. Facial expression editing is a dual-use technology that may be misused to create misleading facial content, impersonate individuals, or manipulate perceived emotions. Our work focuses on expression manipulation for non-commercial academic research and does not aim to alter identity or enable deceptive applications. To mitigate potential risks, the dataset is curated to exclude offensive content, and associated personal metadata is not retained. We encourage downstream users to comply with applicable laws, licenses, platform policies, and institutional ethics guidelines, obtain appropriate consent when editing identifiable individuals, and disclose synthetic or edited media when it may be mistaken for authentic content. Appropriate labeling, watermarking, or provenance mechanisms are encouraged in public-facing applications.

REPRODUCIBILITY STATEMENT

The method and training objectives are described in Section 4, with detailed contrastive formulations in Appendix A. Appendix B provides the model configurations, supervision encoders, LoRA settings, and training hyperparameters. The evaluation metrics are defined in Section 3.2, with additional ablation and user-study protocols in Appendices D and E. Appendix F documents the data sources, generation and annotation procedures, consistency filtering, and human validation, together with the annotation prompt templates.

REFERENCES

aisegmentcn. Matting human datasets. https://github.com/aisegmentcn/matting_human_datasets, 2020. Accessed: 2026-03-03.

Stefan Andreas Baumann, Felix Krause, Michael Neumayr, Nick Stracke, Melvin Sevi, Vincent Tao Hu, and Björn Ommer. Continuous, subject-specific attribute control in T2I models by identifying semantic directions. In *CVPR*, 2025.

Tim Brooks, Aleksander Holynski, and Alexei A Efros. InstructPix2Pix: Learning to follow image editing instructions. In *CVPR*, 2023.

ByteDance. Seedream 4.5: Advanced AI image generation model. https://seed.bytedance.com/en/seedream4_5, 2025. Accessed: 2026-03.

Jingjing Chang, Yixiao Fang, Peng Xing, Shuhan Wu, Wei Cheng, Rui Wang, Xianfang Zeng, Gang Yu, and Hai-Bao Chen. OneIG-Bench: Omni-dimensional nuanced evaluation for image generation. *arXiv preprint arXiv:2506.07977*, 2025.

Wei Cheng, Su Xu, Jingtan Piao, Chen Qian, Wayne Wu, Kwan-Yee Lin, and Hongsheng Li. Generalizable neural performer: Learning robust radiance fields for human novel view synthesis. *arXiv preprint arXiv:2204.11798*, 2022.

Wei Cheng, Ruixiang Chen, Siming Fan, Wanqi Yin, Keyu Chen, Zhongang Cai, Jingbo Wang, Yang Gao, Zhengming Yu, Zhengyu Lin, et al. DNA-Rendering: A diverse neural actor repository for high-fidelity human-centric rendering. In *ICCV*, 2023.

Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation. In *CVPR*, 2018.

Yusuf Dalva, Kavana Venkatesh, and Pinar Yanardag. FluxSpace: Disentangled semantic editing in rectified flow transformers. *arXiv preprint arXiv:2412.09611*, 2024.

Radek Daněček, Michael J Black, and Timo Bolkart. EMOCA: Emotion driven monocular face capture and animation. In *CVPR*, 2022.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. ArcFace: Additive angular margin loss for deep face recognition. In *CVPR*, 2019.

Hui Ding, Kumar Sricharan, and Rama Chellappa. ExprGAN: Facial expression editing with controllable expression intensity. In *AAAI*, 2018.

Zheng Ding, Xuaner Zhang, Zhihao Xia, Lars Jebe, Zhuowen Tu, and Xiuming Zhang. DiffusionRig: Learning personalized priors for facial appearance editing. In *CVPR*, 2023.

Shichuan Du and Aleix M Martinez. Compound facial expressions of emotion: from basic research to clinical applications. *Dialogues in clinical neuroscience*, 2015.

Shichuan Du, Yong Tao, and Aleix M Martinez. Compound facial expressions of emotion. *PNAS*, 2014.

Paul Ekman. An argument for basic emotions. *Cognition & emotion*, 1992.

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

Yao Feng, Haiwen Feng, Michael J Black, and Timo Bolkart. Learning an animatable detailed 3D face model from in-the-wild images. *ACM Transactions on Graphics*, 40(4):1–13, 2021.

Rohit Gandikota, Joanna Materzyńska, Tingrui Zhou, Antonio Torralba, and David Bau. Concept Sliders: LoRA adaptors for precise control in diffusion models. In *ECCV*, 2024.

Rohit Gandikota, Zongze Wu, Richard Zhang, David Bau, Eli Shechtman, and Nick Kolkin. SliderSpace: Decomposing the visual capabilities of diffusion models. In *ICCV*, 2025.

Daniel Garibi, Shahar Yadin, Roni Paiss, Omer Tov, Shiran Zada, Ariel Ephrat, Tomer Michaeli, Inbar Mosseri, and Tali Dekel. TokenVerse: Versatile multi-concept personalization in token modulation space. *TOG*, 2025.

Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, 2014.

Zinan Guo, Yanze Wu, Zhuowei Chen, Lang Chen, Peng Zhang, and Qian He. PuLID: Pure and lightning ID customization via contrastive alignment. In *Advances in neural information processing systems*, 2024.

Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. GANSpace: Discovering interpretable GAN controls. In *NeurIPS*, 2020.

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

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *ICLR*, 2022.

Rahul Jain, Amit Goel, Koichiro Niinuma, and Aakar Gupta. AdaptiveSliders: User-aligned semantic slider-based editing of text-to-image model output. In *CHI*, 2025.

Wooseok Jang, Youngjun Hong, Geonho Cha, and Seungryong Kim. ControlFace: Harnessing facial parametric control for face rigging. In *CVPR*, 2025.

Liming Jiang, Qing Yan, Yumin Jia, Zichuan Liu, Hao Kang, and Xin Lu. InfiniteYou: Flexible photo recrafting while preserving your identity. In *ICCV*, 2025.

Ronen Kamenetsky, Sara Dorfman, Daniel Garibi, Roni Paiss, Or Patashnik, and Daniel Cohen-Or. SAEdit: Token-level control for continuous image editing via sparse autoencoder. *arXiv preprint arXiv:2510.05081*, 2025.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *CVPR*, 2020.

Minchul Kim, Anil K Jain, and Xiaoming Liu. AdaFace: Quality adaptive margin for face recognition. In *CVPR*, 2022.

Black Forest Labs. FLUX.2: Frontier Visual Intelligence. https://bfl.ai/blog/flux-2, 2025.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, et al. FLUX.1 Kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742*, 2025.

Shan Li, Weihong Deng, and JunPing Du. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *CVPR*, 2017.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

Lu Liu, Huiyu Duan, Qiang Hu, Liu Yang, Chunlei Cai, Tianxiao Ye, Huayu Liu, Xiaoyun Zhang, and Guangtao Zhai. F-Bench: Rethinking human preference evaluation metrics for benchmarking face generation, customization, and restoration. In *ICCV*, 2025a.

Ming Liu, Yukang Ding, Min Xia, Xiao Liu, Errui Ding, Wangmeng Zuo, and Shilei Wen. STGAN: A unified selective transfer network for arbitrary image attribute editing. In *CVPR*, 2019.

Shiyu Liu, Yucheng Han, Peng Xing, Fukun Yin, Rui Wang, Wei Cheng, Jiaqi Liao, Yingming Wang, Honghao Fu, Chunrui Han, et al. Step1X-Edit: A practical framework for general image editing. *arXiv preprint arXiv:2504.17761*, 2025b.

Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor. AffectNet: A database for facial expression, valence, and arousal computing in the wild. *IEEE transactions on affective computing*, 2017.

Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. Deep metric learning via lifted structured feature embedding. In *CVPR*, 2016.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

OpenAI. Introducing GPT-Image-1.5. https://openai.com/index/new-chatgpt-images-is-here/, 2025. Accessed: 2026-03.

OysterQAQ. DanbooruCLIP. https://huggingface.co/OysterQAQ/DanbooruCLIP, 2023. Accessed: 2023-05-18.

Dongwei Pan, Long Zhuo, Jingtan Piao, Huiwen Luo, Wei Cheng, Yuxin Wang, Siming Fan, Shengqi Liu, Lei Yang, Bo Dai, et al. RenderMe-360: A large digital asset library and benchmarks towards high-fidelity head avatars. In *NeurIPS*, 2023.

Rishubh Parihar, Or Patashnik, Daniil Ostashev, R Venkatesh Babu, Daniel Cohen-Or, and Kuan-Chieh Wang. Kontinuous Kontext: Continuous strength control for instruction-based image editing. *arXiv preprint arXiv:2510.08532*, 2025.

Albert Pumarola, Antonio Agudo, Aleix M Martinez, Alberto Sanfeliu, and Francesc Moreno-Noguer. GANimation: Anatomically-aware facial animation from a single image. In *ECCV*, 2018.

Zongyang Qiu, Bingyuan Wang, Xingbei Chen, Yingqing He, and Zeyu Wang. EmoVid: A multimodal emotion video dataset for emotion-centric video understanding and generation. *arXiv preprint arXiv:2511.11002*, 2025.

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*, pp. 8748–8763. PMLR, 2021.

Naina Raisinghani. Introducing Nano Banana Pro. https://blog.google/innovation-and-ai/products/nano-banana-pro/, 2025. Accessed: 2026-09-25.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. FaceNet: A unified embedding for face recognition and clustering. In *CVPR*, 2015.

Yujun Shen and Bolei Zhou. Closed-form factorization of latent semantics in GANs. In *CVPR*, 2021.

Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of GANs for semantic face editing. In *CVPR*, 2020.

snmahsa. Human images dataset (men and women). https://www.kaggle.com/datasets/snmahsa/human-images-dataset-men-and-women, 2021. Accessed: 2026-03-03.

Meituan LongCat Team, Hanghang Ma, Haoxian Tan, Jiale Huang, Junqiang Wu, Jun-Yan He, Lishuai Gao, Songlin Xiao, Xiaoming Wei, Xiaoqi Ma, Xunliang Cai, Yayong Guan, and Jie Hu. LongCat-Image technical report. *arXiv preprint arXiv:2512.07584*, 2025.

Kaisiyuan Wang, Qianyi Wu, Linsen Song, Zhuoqian Yang, Wayne Wu, Chen Qian, Ran He, Yu Qiao, and Chen Change Loy. MEAD: A large-scale audio-visual dataset for emotional talking-face generation. In *ECCV*, 2020.

Mengting Wei, Tuomas Varanka, Xingxun Jiang, Huai-Qian Khor, and Guoying Zhao. MagicFace: High-fidelity facial expression editing with action-unit control. *arXiv preprint arXiv:2501.02260*, 2025.

Alon Wolf, Chen Katzir, Kfir Aberman, and Or Patashnik. Continuous control of editing models via adaptive-origin guidance. *arXiv preprint arXiv:2602.03826*, 2026.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-Image technical report. *arXiv preprint arXiv:2508.02324*, 2025.

Hengyuan Xu, Wei Cheng, Peng Xing, Yixiao Fang, Shuhan Wu, Rui Wang, Xianfang Zeng, Daxin Jiang, Gang Yu, Xingjun Ma, et al. WithAnyone: Towards controllable and ID consistent image generation. *arXiv preprint arXiv:2510.14975*, 2025a.

Zhenyu Xu, Xiaoqi Shen, Haotian Nan, and Xinyu Zhang. NumeriKontrol: Adding numeric control to diffusion transformers for instruction-based image editing. *arXiv preprint arXiv:2511.23105*, 2025b.

Weixin Ye, Hongguang Zhu, Wei Wang, Yahui Liu, and Mengyu Wang. All-in-one slider for attribute manipulation in diffusion models. *arXiv preprint arXiv:2508.19195*, 2025.

Xingxi Yin, Jingfeng Zhang, Yue Deng, Zhi Li, Yicheng Li, and Yin Zhang. InstructAttribute: Fine-grained object attributes editing with instruction. *arXiv preprint arXiv:2505.00751*, 2025.

Oğuz Kaan Yüksel, Enis Simsar, Ezgi Gülperi Er, and Pinar Yanardag. LatentCLR: A contrastive learning approach for unsupervised discovery of interpretable directions. In *ICCV*, 2021.

Arman Zarei, Samyadeep Basu, Mobina Pournemat, Sayan Nag, Ryan Rossi, and Soheil Feizi. SliderEdit: Continuous image editing with fine-grained instruction control. *arXiv preprint arXiv:2511.09715*, 2025.

Zhicheng Zhang, Weicheng Wang, Yongjie Zhu, Wenyu Qin, Pengfei Wan, Di Zhang, and Jufeng Yang. VidEmo: Affective-tree reasoning for emotion-centric video foundation models. *arXiv preprint arXiv:2511.02712*, 2025.

Weizhi Zhong, Huan Yang, Zheng Liu, Huiguo He, Zijian He, Xuesong Niu, Di Zhang, and Guanbin Li. Mod-Adapter: Tuning-free and versatile multi-concept personalization via modulation adapter. *arXiv preprint arXiv:2505.18612*, 2025.

Hao Zhu, Wayne Wu, Wentao Zhu, Liming Jiang, Siwei Tang, Li Zhang, Ziwei Liu, and Chen Change Loy. CelebV-HQ: A large-scale video facial attributes dataset. In *ECCV*, 2022.

Yule Zhu, Ping Liu, Zhedong Zheng, and Wei Liu. SEED: A benchmark dataset for sequential facial attribute editing with diffusion models. *arXiv preprint arXiv:2506.00562*, 2025.

Cailin Zhuang, Ailin Huang, Yaoqi Hu, Jingwei Wu, Wei Cheng, Jiaqi Liao, Hongyuan Wang, Xinyao Liao, Weiwei Cai, Hengyuan Xu, et al. ViStoryBench: Comprehensive benchmark suite for story visualization. *arXiv preprint arXiv:2505.24862*, 2025.

Appendix

A Details of the Symmetric Contrastive Loss

A.1 Triplet Constraint Formulations

In this section, we provide detailed formulations of the triplet constraint function $\mathcal{T}(G, P, N)$ used in Sec. 4. All features are extracted using a frozen CLIP image encoder to represent expression semantics and are ℓ_2-normalized before distance computation. For brevity, we denote $d_{G,P} = d(G, P)$ and $d_{G,N} = d(G, N)$ as cosine distances, and $s_{G,P} = \mathrm{sim}(G, P)$ and $s_{G,N} = \mathrm{sim}(G, N)$ as cosine similarities.

Hinge-based Formulation. The margin-based objective is

$$\mathcal{T}_{\mathrm{hinge}}(G, P, N) = \max\big(0,\ d_{G,P} - d_{G,N} + m\big), \tag{7}$$

where m is a fixed margin.

Log-Ratio Formulation. We adopt a smooth distance-ratio objective:

$$\mathcal{T}_{\mathrm{ratio}}(G, P, N) = \log\left(\frac{d_{G,P} + \epsilon}{d_{G,N} + \epsilon}\right), \tag{8}$$

where ϵ is a small constant for numerical stability.

InfoNCE-style Formulation. The probabilistic contrastive objective is

$$\mathcal{T}_{\mathrm{nce}}(G, P, N) = -\log \frac{\exp(s_{G,P}/\tau)}{\sum_{x\in\{P,N\}} \exp(s_{G,x}/\tau)}, \tag{9}$$

where τ is a temperature parameter.

A.2 Implementation Details

Unless otherwise specified, we use the InfoNCE-style formulation with temperature $\tau = 0.07$. For the hinge-based variant, the margin is set to $m = 0.2$. For the log-ratio formulation, we set $\epsilon = 10^{-6}$ for numerical stability. All variants are evaluated under identical training schedules.

B Additional Experimental Details

This section provides additional implementation details for PixelSmile, including the model configuration, supervision encoders, LoRA settings, and training hyperparameters. All experiments are conducted on four NVIDIA H200 GPUs.

Model Configuration. We implement PixelSmile based on Qwen-Image-Edit-2511. To account for the distinct visual distributions of real-world and anime images, we train two independent LoRA adapters, one for each domain. Following prior work (Zhuang et al., 2025; Chang et al., 2025), we adopt domain-specific image encoders for contrastive supervision. Specifically, CLIP-ViT-L/14 (Radford et al., 2021) is used for the real-world domain, while DanbooruCLIP (OysterQAQ, 2023) is used for the anime domain. For real-world images, identity preservation is supervised using the pretrained ArcFace antelopev2 model (Deng et al., 2019). Since reliable face recognition models are primarily designed for real-world faces, the identity preservation loss is not applied to the anime-domain adapter.

LoRA Configuration. We apply LoRA to the major attention and MLP modules of the diffusion transformer. The LoRA rank is set to 64, the scaling factor is set to 128, and the dropout rate is set



Figure 7: **Qualitative Comparison with General Editing Models.** PixelSmile produces clearer expression changes while preserving facial identity, whereas existing editing models either weaken expression editing or degrade identity consistency.

to 0. All parameters of the pretrained backbone remain frozen during training, and only the inserted LoRA parameters are optimized.

Training Hyperparameters. We train each domain-specific adapter for 100 epochs using the AdamW optimizer. The optimizer parameters are set to $\beta_1 = 0.9$, $\beta_2 = 0.999$, weight decay 0.001, and $\epsilon = 10^{-8}$. The initial learning rate is 10^{-4}, followed by a cosine learning-rate schedule with 500 warmup steps. Training is performed using bf16 mixed precision. The batch size is 4 per GPU, and the number of gradient accumulation steps is set to 1. For the overall training objective, we set the symmetric contrastive loss weight to $\lambda_{\text{sc}} = 1.0$ and the identity preservation loss weight to $\lambda_{\text{id}} = 0.1$. The symmetric contrastive loss adopts the InfoNCE-style formulation described in Appendix A.

Identity Evaluation. For evaluation, S_{ID} averages source–edited face cosine similarities from three face recognition models: ArcFace (Deng et al., 2019), AdaFace (Kim et al., 2022), and FaceNet (Schroff et al., 2015). This ensemble score is used in HES and is distinct from the ArcFace-based identity supervision used during training.

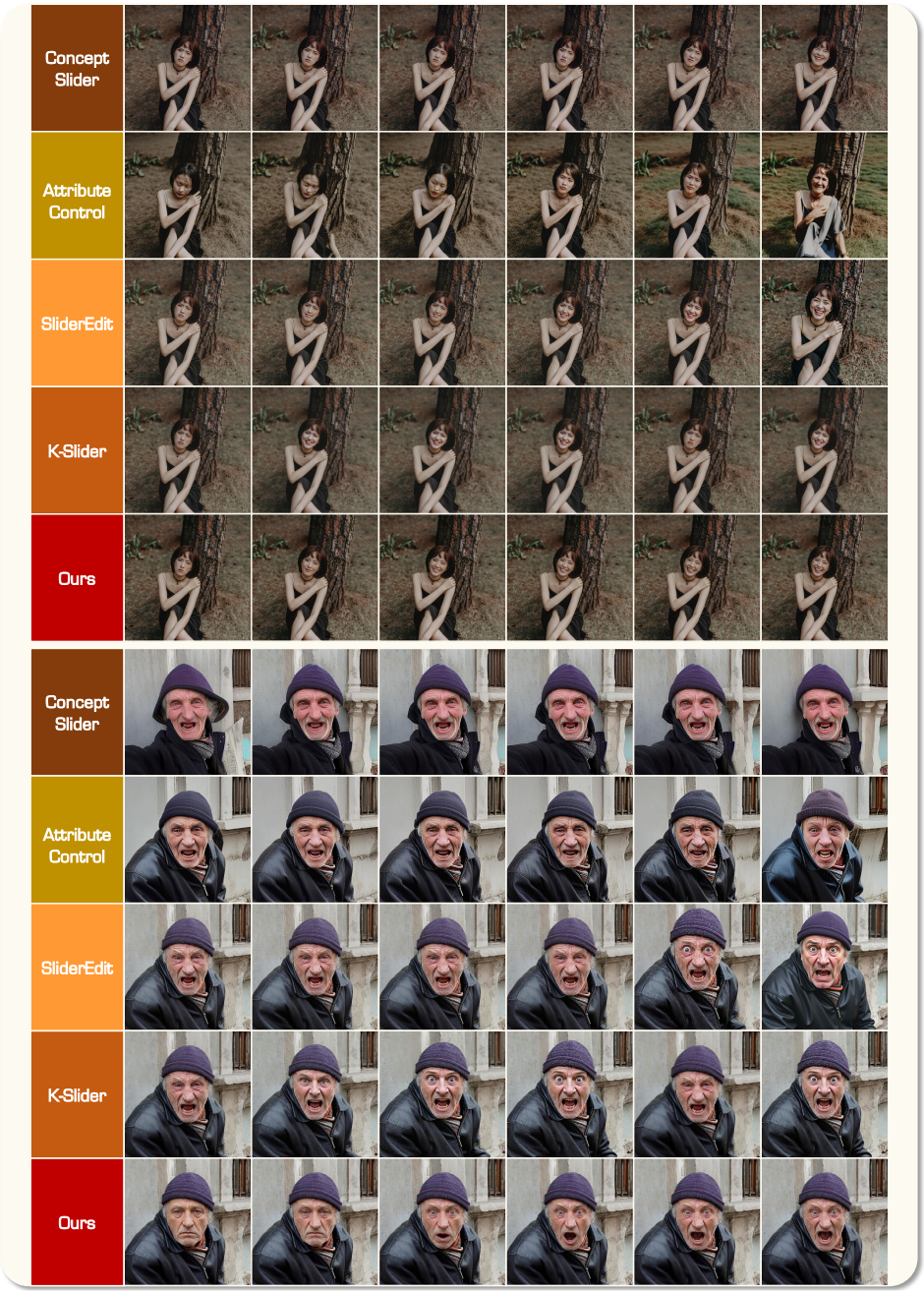


Figure 8: **Qualitative Comparison with Linear Control Models.** PixelSmile achieves smooth and monotonic expression transitions while preserving facial identity, whereas existing control methods either produce unstable responses or sacrifice identity consistency. The figure illustrates two representative expressions: happy (upper half) and surprised (lower half).

C Additional Qualitative Results

This section provides additional qualitative results for PixelSmile. We present more expression editing results across multiple expression categories, intensity levels, and real-world and anime domains, together with full qualitative comparisons and additional expression blending results.

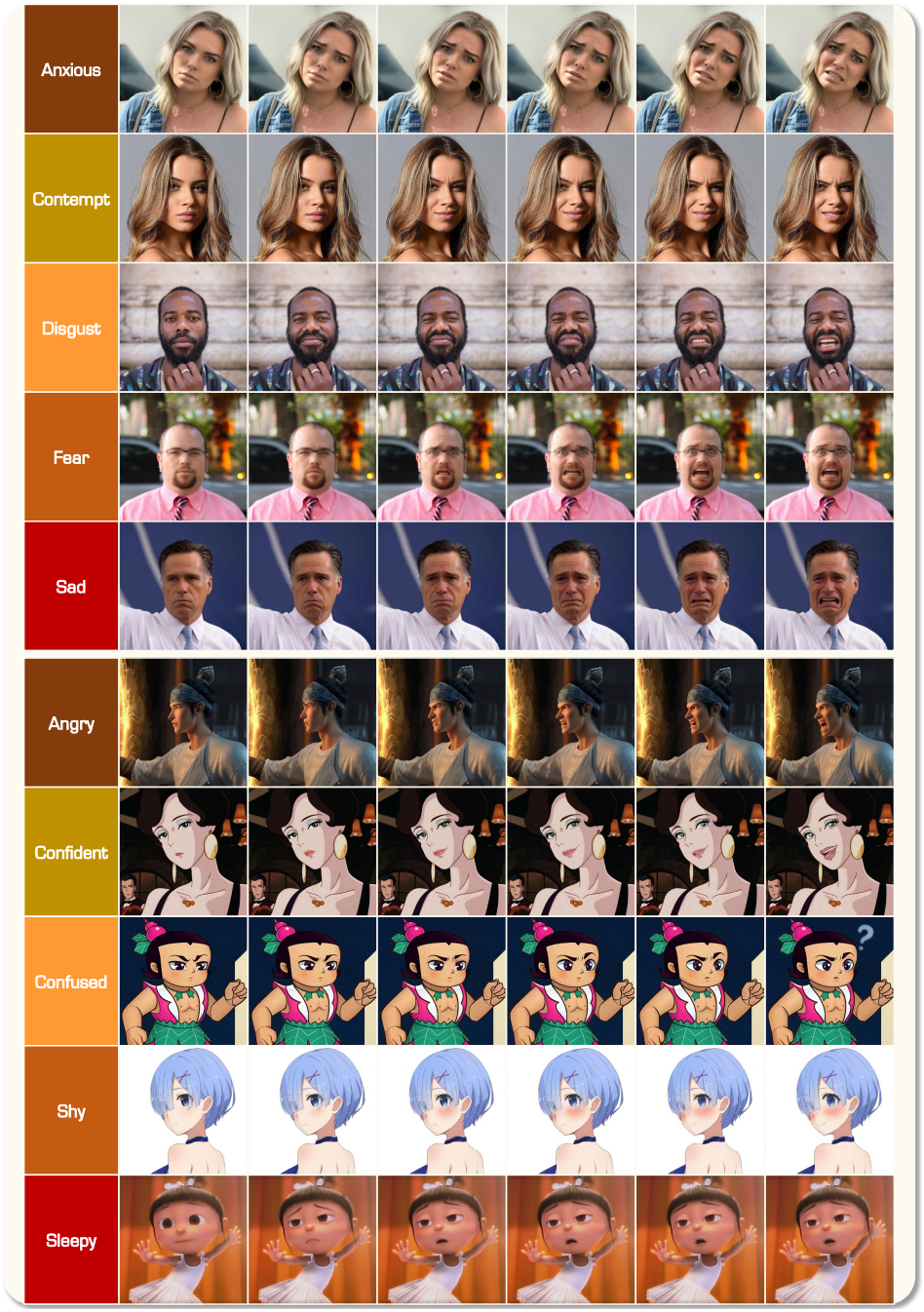


Figure 9: **Additional Expression Editing Results.** We show the remaining ten expressions across both real and anime domains. The upper and lower halves show results on real and anime images, respectively. Expression intensity increases from left to right for each expression.

C.1 Full Qualitative Comparisons

Figures 7 and 8 present the full qualitative comparisons with general editing models and linear control models, respectively. These figures extend the representative results shown in the main paper with additional expressions and examples.

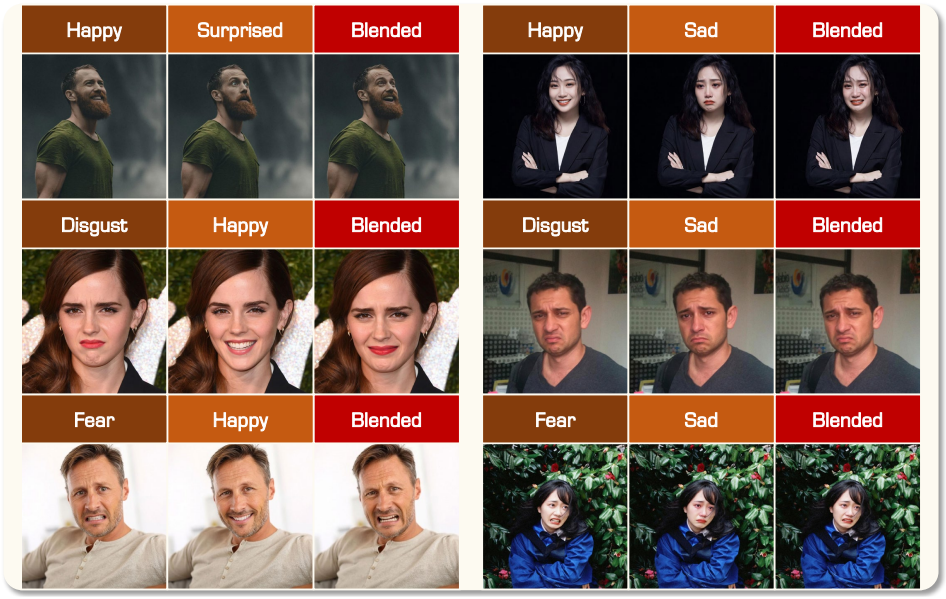


Figure 10: **Expression Blending Results.** Visualizing compositional facial expressions generated by smoothly blending multiple emotional categories in PixelSmile.

C.2 Additional Expression Editing Results

Figure 9 presents additional editing results for the remaining ten expressions across both real-world and anime domains. As the control parameter increases, the expression intensity changes smoothly while the facial identity remains consistent.

C.3 Expression Blending Results

Human facial expressions are often compositional rather than purely categorical (Du et al., 2014; Du & Martinez, 2015; Pan et al., 2023). To examine whether such compositionality emerges in the learned expression space, we perform pairwise interpolation among the six basic expressions, producing 15 zero-shot expression combinations.

Figure 10 presents representative qualitative results. Nine of the fifteen combinations produce perceptually plausible compound expressions, indicating that the learned representation supports continuous linear composition. For example, interpolating expressions with compatible facial cues yields outputs that preserve recognizable characteristics of both endpoints.

Not all expression pairs can be combined successfully. Semantically overlapping pairs, such as Fear and Surprise, may collapse toward a single dominant expression, while pairs with conflicting facial movements, such as Angry and Happy, can produce unstable or less natural results. These failures suggest that linear composition is constrained by the semantic and physiological compatibility of the component expressions. Overall, the results demonstrate meaningful compositional structure in the learned expression space while also revealing its limitations for incompatible expression pairs.

D Additional Ablation Details

This section provides supplementary analyses for the ablations reported in Table 3.

D.1 Ablation on Loss Components

Removing the identity loss improves editing accuracy and expression disentanglement but substantially degrades identity consistency. Without explicit identity supervision, the model may alter identity-related attributes, such as hairstyle, skin texture, and facial structure, particularly at large

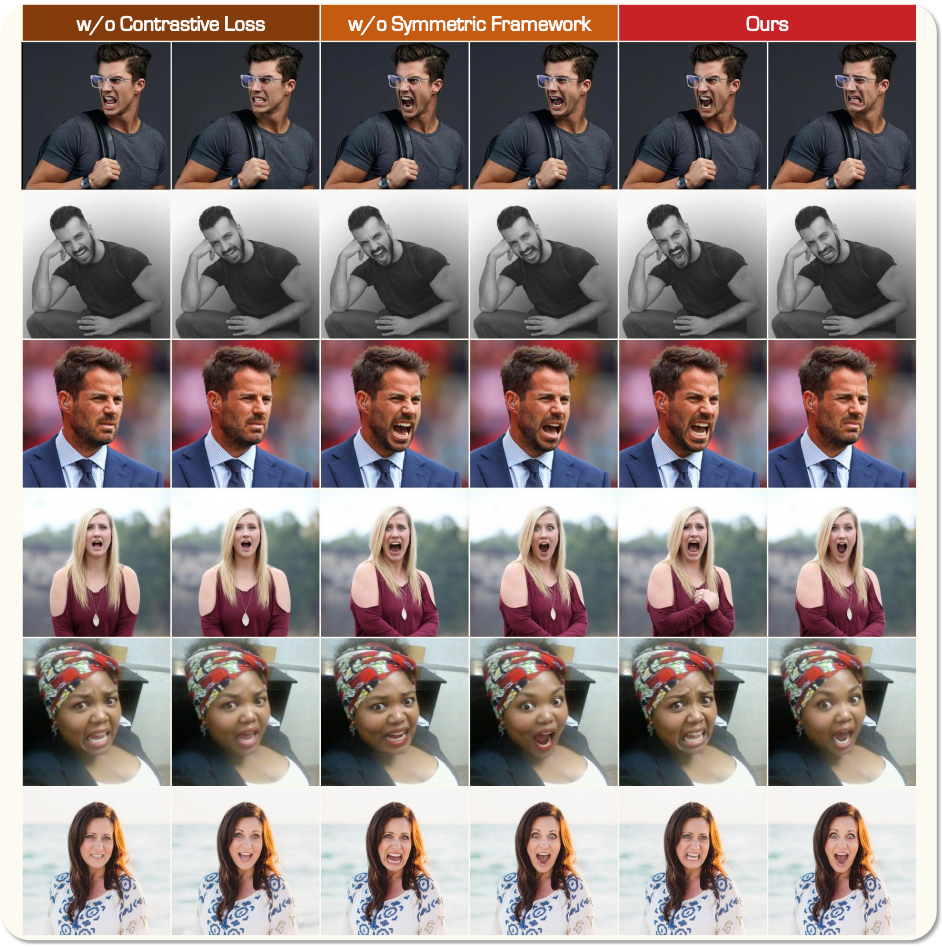


Figure 11: **Ablation on Contrastive Loss and Symmetric Training.** Both variants (w/o Contrastive Loss and w/o Symmetric Framework) suffer from expression confusion, while our full method achieves precise expression disentanglement. The upper three rows show angry and disgust, and the lower three rows show fear and surprised.

editing intensities. As shown in Fig. 12, the full model preserves a consistent facial identity, whereas the variant without $\mathcal{L}_{\text{ID}}$ exhibits noticeable identity drift.

Removing the symmetric contrastive loss produces high identity similarity but weak expression editing and severe structural confusion. The model tends to remain close to the source image and fails to distinguish expressions with overlapping facial cues. Fig. 11 shows that the full model generates more clearly separated expressions than the variant without $\mathcal{L}_{\text{SC}}$. These results confirm the complementary roles of the two losses: $\mathcal{L}_{\text{ID}}$ preserves identity, while $\mathcal{L}_{\text{SC}}$ promotes expression strength and semantic separation.

D.2 Ablation on Symmetric Training

We compare Fully Symmetric Joint Training with an asymmetric variant that applies contrastive supervision to only one branch. As shown in Fig. 6, the asymmetric variant reduces the training loss faster at the beginning but converges to lower editing accuracy and higher structural confusion. The symmetric model converges more gradually but achieves better final performance, indicating that bidirectional supervision stabilizes training and improves expression disentanglement.

D.3 Ablation on Contrastive Objectives

We compare Hinge, Log-ratio, and InfoNCE formulations of the triplet objective $\mathcal{T}$. Hinge produces stronger expression edits but causes greater identity degradation, whereas Log-ratio better preserves identity but weakens expression editing. InfoNCE achieves the best balance among editing accuracy, structural confusion, and identity preservation, and is therefore used as the default formulation.



Figure 12: **Ablation on ID loss.** Without the ID loss, high expression intensities cause identity drift in hairstyle and skin texture. Our full method preserves identity consistently.

D.4 Ablation on Training Dataset

Among existing human-centric datasets (Cheng et al., 2023; Pan et al., 2023; Cheng et al., 2022; Wang et al., 2020; Zhu et al., 2022), we select MEAD (Wang et al., 2020) to evaluate the contribution of FFE. MEAD contains seven discrete facial expressions recorded from multi-view videos at three intensity levels: low, medium, and high. We use its front-view subset and map these levels to 0.5, 0.75, and 1.0, respectively, to match the conditioning range of PixelSmile.

We uniformly sample video frames and construct same-identity triplets $(I_{\text{orig}}, P_a, P_b)$, where I_{orig} is the neutral source frame and P_a, P_b contain two distinct target expressions of the same subject. The model is trained using the same architecture and optimization settings as the full method. As reported in Table 3, the MEAD-trained model underperforms the model trained on FFE across all metrics. This gap suggests that MEAD's limited identity diversity and three discrete intensity levels are insufficient for fine-grained intensity modeling and expression disentanglement, whereas FFE provides richer identities and continuous expression annotations.

E Details of the User Study

We conduct a user study to evaluate the perceptual continuity and identity consistency of continuous expression editing. The study includes 2,400 generated images evaluated by 10 trained annotators.

For each evaluation sample, annotators are shown the same source image and the editing results produced by PixelSmile, K-Slider, and SliderEdit for the same target expression. Each method generates an image sequence using increasing control strengths. The three sequences are presented together, allowing annotators to assess the complete editing trajectory rather than individual outputs.

Annotators rank the three methods independently according to two criteria:

- **Expression continuity:** whether expression intensity changes smoothly and monotonically across the control levels;
- **Identity consistency:** whether the subject's facial identity remains stable throughout the editing sequence.

The rankings are converted into scores and averaged across annotators and evaluation samples. As shown in Fig. 13, the resulting mean continuity and identity scores are PixelSmile (4.48, 3.80), K-Slider (1.36, 4.06), and SliderEdit (3.16, 1.14). PixelSmile receives the highest continuity score while maintaining strong identity consistency. In comparison, K-Slider preserves identity but produces weak and irregular expression changes, whereas SliderEdit achieves stronger changes at the cost of substantial identity degradation. These human judgments are consistent with the quantitative CLS and HES results reported in the main paper.

F Additional Dataset Details

This section details the source images, facial-attribute-guided generation, continuous annotation and consistency filtering, human validation, and statistics of FFE.

F.1 Data Sources and Base Image Selection

We curate high-quality base images from real-world and anime domains. For the real-world domain, approximately 6K portraits are collected from the publicly available Human Images Dataset (snmahsa, 2021) and Matting Human Dataset (aisegmentcn, 2020), covering diverse demographics, scenes, and compositions. For the anime domain, we collect stylized portraits from 207 anime productions covering 629 fictional characters, retaining approximately 6K images after quality filtering and face detection. In both domains, automated face detection followed by manual verification removes images with unclear identities, severe occlusion, or low visual quality.

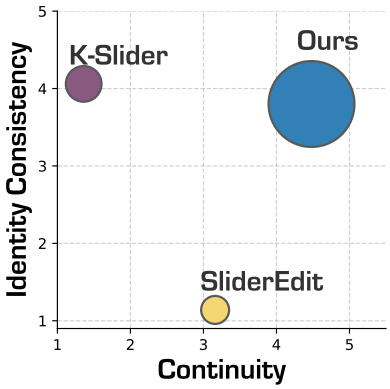


Figure 13: **User study results.** The axes show human-rated identity consistency and editing continuity. Bubble size represents the automatic HES from FFE-Bench.

F.2 Facial-Attribute-Guided Expression Generation

We observe that directly prompting *Nano Banana Pro* with expression names can produce ambiguous or inaccurate edits, particularly for semantically overlapping expressions. Descriptions of local facial attributes provide more explicit guidance for editing individual facial features. We therefore construct a structured prompt library organized by expression category and intensity, covering six basic emotions (Ekman, 1992) and six extended emotions: Confused, Contempt, Confident, Shy, Sleepy, and Anxious.

Each expression is described through localized attributes, including mouth shape, eyebrow movement, and eye openness. Descriptions at different intensities specify corresponding changes in these features. Candidate attribute combinations are generated and filtered by a vision-language model to remove anatomically inconsistent or semantically conflicting descriptions. Each generation instruction combines a global expression category with the corresponding local facial descriptions, providing guidance beyond the category name alone.

For each base identity, we use these structured instructions to synthesize multiple expressions at varying intensities with *Nano Banana Pro*. The resulting dataset contains approximately 60K images, with 30K per domain. The intended expression guides generation but is not directly assigned as the output image's annotation; generated results undergo image-based scoring and consistency filtering as described below.

F.3 Continuous Annotation and Consistency Filtering

Each generated image receives a 12-dimensional continuous expression vector $\mathbf{v} \in [0,1]^{12}$, with each dimension representing the intensity of one expression category. This representation accommodates semantic overlap rather than assigning a single discrete label. We use *Gemini 3 Pro* for scoring after comparing it with GPT-4o and Qwen3-VL-235B-A22B through qualitative inspection and consistency evaluation, in which Gemini 3 Pro showed the most stable expression-intensity estimates. Table 6 provides the scoring prompt.

We compare the predicted expression with the intended generation target to identify potentially unsuccessful edits or unreliable annotations. Samples with large discrepancies undergo manual review and filtering, while consistent samples receive random manual spot checks. Ambiguous or low-confidence samples are removed through this process. Generation targets, image-based model predictions, and human judgments thus provide complementary evidence for sample selection, rather than relying on either the target instruction or a single model prediction alone. Agreement with the target serves as a screening criterion; the reliability of the continuous scores is additionally assessed against human judgments below.

F.4 Human Validation of Continuous Annotations

To assess annotation reliability, we compare Gemini-based scores with human judgments on 360 images: 30 images per expression across the 12 categories, spanning varying intensities. The scores achieve Pearson and Spearman correlations of 0.79 and 0.81, respectively. This evaluation measures agreement with human intensity judgments, complementing the target-consistency checks used during dataset construction.

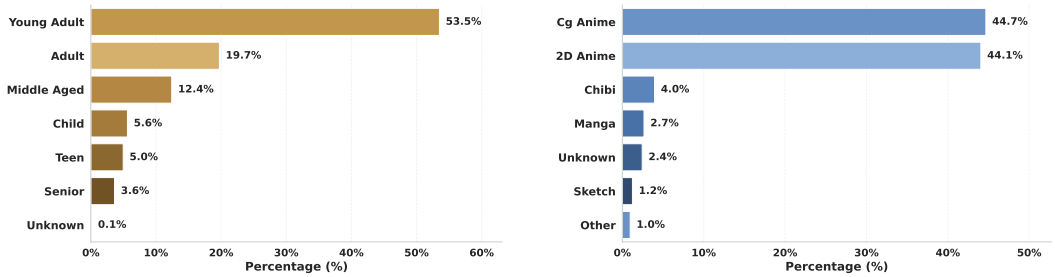
F.5 Dataset Statistics and Annotation Templates

Tables 4 and 5 provide the prompts used to extract descriptive and categorical attributes for statistical analysis of the human and anime subsets. These prompts use *Qwen3-VL-235B-A22B*, rather than the Gemini model used for continuous expression scoring. The two annotation tasks serve different purposes: dataset characterization and expression-intensity supervision, respectively.

We present the statistical analysis of FFE in Fig. 14 and Fig. 15, covering categorical distributions and textual description patterns across the real-world and anime domains.

From Fig. 14, the real-world subset is diverse but imbalanced, dominated by young adults (53.5%), with children, teens, and seniors forming smaller proportions. Similar trends are observed in other attributes, where female samples are more frequent and light-to-medium skin tones constitute the majority, indicating that the dataset inherits non-uniform demographic characteristics. This bias reflects common patterns in portrait-centric internet images and introduces challenges for expression modeling. In contrast, the anime subset exhibits broader stylistic diversity, with CG and 2D anime each accounting for about 44%, along with additional styles such as chibi, manga, and sketch. Compared with the real-world subset, the anime subset also shows a flatter age distribution. However, it contains more unknown labels in attributes like gender and age, suggesting that stylized characters are inherently more ambiguous under real-world categorization schemes, which increases the difficulty of consistent annotation and evaluation.

Fig. 15 further reveals domain-specific textual patterns. The real-world subset emphasizes natural appearance cues such as clothing, hairstyle, and facial details, while the anime subset contains more stylized and visually distinctive descriptions. This difference highlights that expression editing in FFE involves both visual transformation and domain-dependent semantic interpretation, requiring models to generalize across heterogeneous distributions. Overall, these statistics indicate that FFE combines substantial diversity with realistic biases, making it a challenging and representative benchmark for fine-grained, diverse, and real-world facial expression editing.



(a) Age distribution in the real-world domain

(b) Style distribution in the anime domain

Figure 14: **Statistical distributions of annotated data in FFE.** The results provide insights into the underlying data characteristics across real-world and anime domains.

Table 4: **Human Dataset Annotation Prompt Template.**

Human Dataset Annotation Prompt

System Context

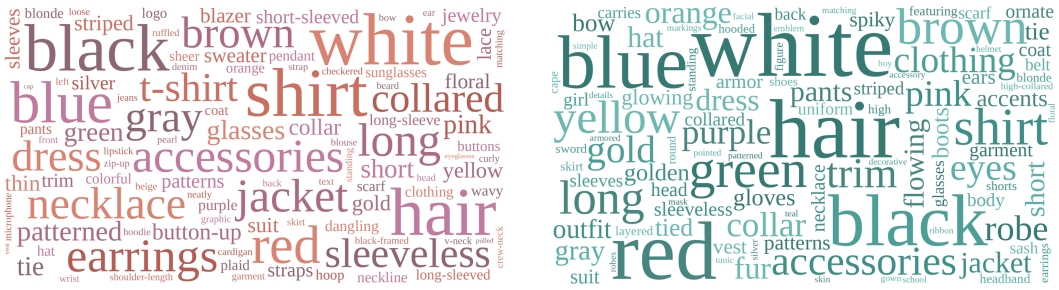
You are an image annotation assistant. For a single person image, output strict JSON only.

Requirements

- Describe only visible facts; do not infer identity, story, or intent.
- `categorical` values must be selected from the provided enums.
- All three fields in `descriptions` must be present.
- Write all description sentences in English, each 8–25 words.

JSON Schema

```
{
  "categorical": {
    "gender": "male/female/androgynous/unknown",
    "age_group": "child/teen/young_adult/adult/middle_aged/senior/
        unknown",
    "skin_tone": "very_light/light/medium/dark/very_dark/unknown",
    "expression": "neutral/happy/sad/angry/surprised/fear/disgust/
        other/unknown"
  },
  "descriptions": {
    "appearance_sentence": "One sentence about clothing and
        accessories.",
    "action_sentence": "One sentence about action or pose.",
    "background_sentence": "One sentence about scene/background."
  }
}
```



(a) Real-world appearance descriptions

(b) Anime-style appearance descriptions

Figure 15: **Visualization of appearance-related textual descriptions in FFE.** The visualizations highlight the distribution and diversity of annotations across real-world and anime domains.

Table 5: **Anime Dataset Annotation Prompt Template.**

Anime Dataset Annotation Prompt

System Context

You are an anime image annotation assistant. For a single character image, output strict JSON only.

Requirements

- Describe only visible facts; do not infer identity, story, or intent.
- `categorical` values must be selected from the provided enums.
- All three fields in `descriptions` must be present.
- Write all description sentences in English, each 8–25 words.

JSON Schema

```
{
  "categorical": {
    "gender": "male/female/androgynous/unknown",
    "age_group": "child/teen/young_adult/adult/middle_aged/senior/
        unknown",
    "expression": "neutral/happy/sad/angry/surprised/fear/disgust/
        other/unknown",
    "anime_style": "2d_anime/chibi/manga/sketch/cg_anime/other/
        unknown"
  },
  "descriptions": {
    "appearance_sentence": "One sentence about clothing and
        accessories.",
    "action_sentence": "One sentence about action or pose.",
    "background_sentence": "One sentence about scene/background."
  }
}
```

Table 6: **Facial Expression Scoring Prompt Template.** The same prompt is applied to both human and anime domains, highlighting the domain-agnostic nature of our scoring pipeline.

Facial Expression Scoring Prompt

Role and Task Definition
You are an expert AI specialized in analyzing facial expressions in both photorealistic and anime-style images. Your task is to analyze the input image and determine the intensity score for specific expression categories.

Target Emotion Definitions
Analyze the image based on the following visual definitions. Crucially, pay close attention to the distinctions between similar expressions (e.g., Fear vs. Surprise, Anger vs. Disgust).

- Happiness: Faintly smiling, cheerful, ecstatic; corners of the mouth raised or laughing.
- Sadness: Somber, sorrowful, devastated; frowning with downcast eyes or tears.
- Anger: Annoyed, angry, furious; furrowed brows and glaring eyes.
- Fear: Wary, frightened, terrified; knitted brows and pupil constriction.
- Surprise: Taken aback, amazed, stunned; eyes wide open or jaw dropped.
- Disgust: Distasteful, repulsed, gagging; wrinkling the nose or raising the upper lip.
- Embarrassment: A shy smile, blushing cheeks, avoiding eye contact, looking down, or covering the face with hands.
- Confidence: Self-assured smile, corners of the mouth raised, sharp and firm gaze, slightly smug.
- Confusion: Knitted brows, looking aside or rolling the eyes, head tilted in deep thought.
- Drowsiness: Heavy drooping eyelids, yawning, lethargic or low-energy look.
- Contempt: Asymmetrical smirk (one corner raised), sneering, looking down on someone.
- Nervousness: Visible sweat drops, tense facial muscles, uneasy or restless gaze.

Output Format Requirements

- Return format: Output a standard JSON object.
- Keys: Use the exact emotion labels provided above.
- Values: The intensity score (a float between 0.00 and 1.00).
- Constraint: Do not explain your reasoning. Output only the JSON object. Do not include Markdown formatting such as ```` ```json ... ``` ````.

Output Example

```
{
  "Happiness": 0.90,
  "Surprise": 0.75,
  "Anger": 0.05,
  ...
}
```