Barycentric Guidance: Turning Foundation Image Editors into Continuous Affective Controllers

Harvey Mannering
University of Southampton
hgm1g14@soton.ac.uk

Zhiwu Huang
University of Southampton
Zhiwu.Huang@soton.ac.uk

Adam Prügel-Bennett
University of Southampton
apb@ecs.soton.ac.uk

Abstract

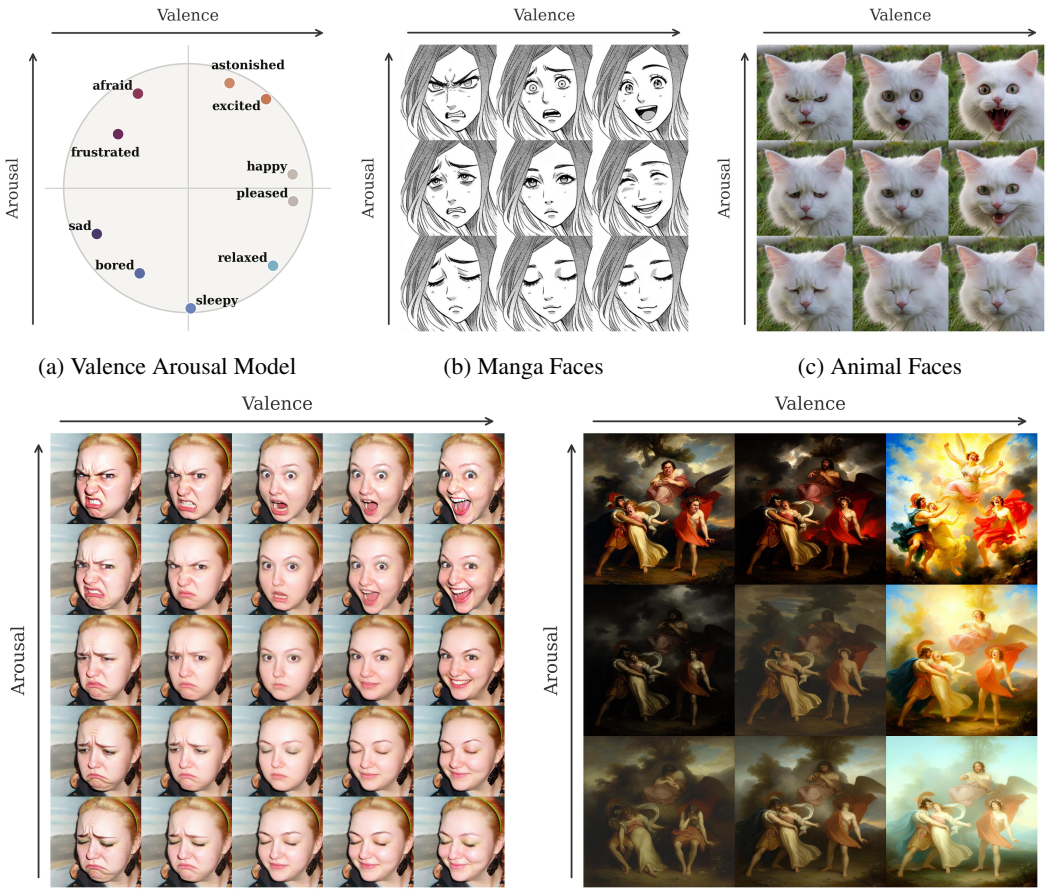
Emotions are expressed on a continuum and are naturally described in the valence-arousal (VA) space. Existing VA-based image editors are typically training-based and domain-specific, limiting them to VA regions seen during training and making them brittle across datasets and domains. Meanwhile, despite producing high-quality edits, state-of-the-art foundation image editors are not continuous affective controllers: non-linear emotion representations in text encoders makes prompt interpolation poorly calibrated, so accurate and continuous VA control remains unreliable. To fill this gap, we introduce Barycentric Guidance, a training-free framework that turns pretrained foundation image editors into continuous affective controllers over VA, requiring no finetuning, extra data, or model changes. Our method maps any target VA point to simplex-based guidance weights, enabling VA controllability in three aspects: (i) accurate targeting of VA states, (ii) broad coverage across the VA plane, and (iii) continuous control. Across three face datasets, we improve VA target accuracy by 25–29%, expand VA-plane coverage by 28–50%, and increase expression diversity by 29–49%, over the strongest baseline. Unlike prior domain-specific methods, our training-free approach exploits the foundation model's inherent versatility, enabling generalization beyond human faces to animals, artworks, and complex scenes within a single framework. Code is available at https://github.com/harveymannering/BarycentricGuidance.

1 Introduction

Emotions are usually expressed in degrees, not only as discrete categories such as happy or sad. The valence–arousal (VA) model [1] describes this by representing affect as a continuous two-dimensional space: valence ranges from unpleasant to pleasant, and arousal from calm to excited (see Figure 1a). This dimensional view provides an intuitive "affective control" for visual emotion editing: changing VA should yield smooth, calibrated changes in facial expression (or visual mood) while keeping non-affective content, such as identity, hair, and background, unchanged. In this paper, we cast VA-based editing as a control problem: given a VA target, and optionally a source image, an editor should track that target reliably, behaving as a continuous affective controller over the VA plane.

Despite this appeal, existing VA-based editing methods are typically training-based and tailored to specific tasks or domains, most commonly human faces or scenes [6–9]. These approaches often struggle to cover the full VA space: due to dataset bias toward near-neutral and socially common expressions, rare regions (e.g., high-arousal negative or low-arousal states) are underrepresented, leading to weak generalisation to unseen VA combinations and reduced controllability at the extremes

40th Conference on Neural Information Processing Systems (NeurIPS 2026).



(a) Valence Arousal Model (b) Manga Faces (c) Animal Faces

(d) Human Faces (e) Artworks

Figure 1: Valence-Arousal-based image editing results from our approach. **(a)** The VA model of emotion from Russell [1] **(b)** - **(e)** VA-based editing across a range of datasets including Manga Facial Expression [2], AFHQ [3], CelebA-HQ [4], and EmoArt [5]. VA ranges [-1,1] with uniform steps.

[10]. Moreover, models trained for one domain often generalise poorly across domains e.g., between face datasets, from human faces to animal/cartoon/3D faces, or between faces and non-face imagery.

At the same time, state-of-the-art foundation image editing diffusion models [11–13] can perform a wide range of high-quality edits, including discrete emotional manipulation. However, they are not continuous affective controllers by default: non-linear text embedding spaces make prompt interpolation poorly calibrated, so accurate and continuous traversal of VA states remains unreliable.

We focus on VA because our aim is continuous affective control rather than a comprehensive model of emotion. Its two interpretable dimensions are naturally expressible in language and widely used in prior continuous affective editing work [6–9]. This exposes a gap between VA-based methods, which provide continuous control but are typically task-specific, and foundation image editors, which offer broad semantic and identity-aware editing but lack reliable continuous affect control. Alternative representations such as valence-arousal-dominance (VAD) and appraisal models capture richer affective factors, but are less standardised as smooth geometric control spaces for image editing.

To address this gap, we introduce Barycentric Guidance, a training-free framework that turns pre-trained foundation editors into continuous affective controllers over VA, requiring no additional data, finetuning, or model modification. Our method maps any target VA point to simplex-based guidance weights, enabling (i) accurate targeting of VA states, (ii) broad coverage across the VA plane, and (iii) continuous control with principled intensity scaling from a neutral centre. When needed, we further

improve perceptual continuity with lightweight refinement (e.g., partial re-diffusion and interpolation such as morphing or optical flow).

We present our main quantitative evaluation on facial expression editing (see Figure 1d), where VA control is well studied and benchmark datasets are available. Compared to prior VA-based facial editing methods [6–9], our approach produces more diverse expressions, more accurate VA alignment, and stronger performance across a wider range of face datasets. Importantly, because our method is entirely training-free, it also generalises beyond human faces: Figure 1b, 1c, & 1e illustrate results on cartoon faces, animal faces, and artworks, respectively. In summary, our contributions are:

- We propose a training-free framework that turns pretrained foundation image editors into continuous affective controllers in VA space, focusing on three aspects of VA controllability: accuracy, coverage, and continuity.
- We evaluate our approach on facial expression editing, finding that we achieve state of the art alignment with the valence-arousal model.
- We demonstrate that our approach can be applied beyond human faces, enabling training-free emotion editing of non-human faces, scenes, and artworks.

2 Related Works

Deep Generative Models Generative models learn data distributions and sample from them to produce new data. Early deep learning approaches include VAEs [14] and GANs [15], while recent work focuses on autoregressive [16, 17] and diffusion models [18, 19]. Diffusion and its flow-based variants [20, 21] have been particularly impactful due to stable training, high-quality, and diverse outputs, enabling text-to-image systems such as Stable Diffusion [22] and FLUX [23]. However, because text input alone makes fine-grained control difficult, a series of finetuning approaches such as ControlNet [24] and IP-Adapter [25] have been developed. These finetuning methods, in turn, face their own challenges due to the large size of text-to-image models themselves. Recently, a spate of image editing diffusion models have been released. Popular open source examples include FLUX.1 Kontext [11], Qwen Image Edit [12], and FLUX.2 [13]. These models excel at discrete editing, but struggle to edit the properties of images in a continuous fashion [26], making their application to VA-based editing non-trivial. While finetuning-based approaches enable continuous editing [26, 27], they incur high computational cost and limit generalization. In contrast, our method is training-free, enabling broad applicability and easy adaptation to new models.

Continuous Emotion Editing Early work on facial expression editing used GANs [28, 29, 7, 6] with various control signals, including discrete emotions [30], Action Units [31], the VA model [1], and learned continuous emotion spaces. However, these methods cannot leverage the generalization ability of text-to-image models, making them brittle when applied to out-of-distribution face images. For diffusion models, a common way to control facial expressions is by conditioning on a 3D face model [32–34], Action Units [35], or landmarks [36, 37]. While these methods enable fine-grained control, they lack an intuitive and easily interpretable model of emotion, such as that provided by the valence–arousal model. Paskaleva et al. [38] unify Action Units and VA via an MLP that augments Stable Diffusion text conditioning. While limited by artifacts and training overhead, their emotion model is an interesting unification of existing approaches. FaceCrafter [9] finetunes Arc2Face [39] for ID, expression, and pose conditioned face generation. Concurrent work, PixelSmile [27], finetunes Qwen Image Edit with a LoRA [40] using a contrastive loss allowing for continuous control of several discrete emotions, we on the other hand focus on emotion editing based on the VA model.

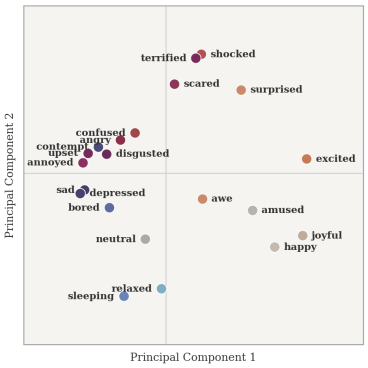


Figure 2: PCA Performed on CLIP Text Embedddings.

Beyond faces, EmotiCrafter [8] trains a transformer to maps text and VA inputs to Stable Diffusion text embeddings for image generation of scenes. Emotional image editing [41–43], generation [44],

and stylization [45] methods have also been applied to control the emotion evoked by an image. While these approaches achieve impressive results, they typically require supervised finetuning or rely on discrete textual prompts. Together, the diffusion-based methods described achieve remarkable performance in emotion generation and editing across their respective domains. However, finetuning on specific datasets requires large compute resources and can restrict the capabilities of a foundation models. We instead focus on how best to leverage foundation models without any finetuning, thus inheriting the multi-domain strengths of the original text-to-image models.

3 Proposed Approach

3.1 Motivation & Background

First, we explore the ability of text encoders to control valence and arousal in facial expressions. Specifically, we examine the CLIP [46] embeddings used in the Stable Diffusion [22] and FLUX.1 [23] models. We use CLIP to embed prompts with the template "a portrait photo of a <emotion> face", where "<emotion>" is replaced with an emotion word (e.g. "happy", "sad", etc). We then do dimensionality reduction on the CLIP embeddings using PCA. Interestingly, this largely recovers the valence–arousal model on the two principal components, as shown in Figure 2, suggesting CLIP implicitly encodes VA structure. This holds across different CLIP models and prompt templates, but not for encoders such as T5 [47] and QwenVL [48] (see Section B for details).

While this linear subspace enables VA control in CLIP-based models, newer systems rely less on CLIP. For example, FLUX.1 Kontext uses CLIP only as a secondary encoder, with T5 as primary, which lacks such structure. Since the text encoders for Qwen Image Edit and FLUX.2 also show no evidence of having learned a valence-arousal subspace, a different approach must be taken to control facial expressions based on valence and arousal. Moreover, prompt interpolation often yields abrupt edits (see Sections 3.2 and 4.1). To bypass these problems with the text encoder embedding spaces, we therefore work in the diffusion models latent space instead.

3.2 Barycentric Guidance

Image editing models excel at discrete editing, but struggle with continuous control. We can leverage this advantage toward the goal of continuous editing by taking linear combinations of discrete outputs. To this end, as shown in Figure 3, we approximate the VA model by constructing a triangle mesh in 2D of discrete emotions. The vertices of this mesh are defined by discrete emotion anchors relevant to the target domain; for instance, in facial expression editing, points are based on the average locations of the valence arousal points in the AffectNet dataset [49], with additional expressions excited, sleepy, and serene used to fill out the remainder of the VA space.

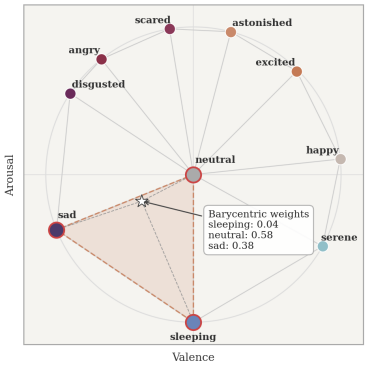
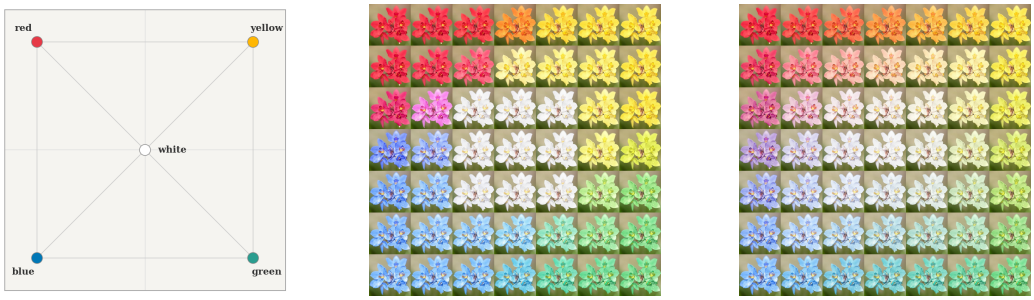


Figure 3: Triangle Mesh created from VA Points.

For a given valence arousal point (v, a), we first need to identify the triangle within which the point lies. If the point is outside all triangles, we choose the triangle that is closest to (v, a). From this we can get the three discrete emotions that are closest to our valence arousal point. Barycentric interpolation can be used to determine the weights (w_1, w_2, w_3) to assign to each discrete emotion by solving the system of equations

$$\begin{bmatrix} A_x & B_x & C_x \\ A_y & B_y & C_y \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ w_3 \end{bmatrix} = \begin{bmatrix} v \\ a \\ 1 \end{bmatrix}, \tag{1}$$

where A, B, and C are the three corners of the triangle, and $\sum_{i=1}^{3} w_i = 1$. These weights can then be used to interpolate between different prompt embeddings, however, we find that this prompt interpolation results in sudden and uneven changes when making edits (see Figure 4b). Instead of interpolating the text embedding, we therefore reformulate classifier-free guidance [50] to be



(a) Simplex mesh of colours. (b) Prompt Interpolation. (c) Barycentric Guidance.

Figure 4: Colour editing comparison between prompt interpolation and barycentric guidance. Prompts: "Make this flower <colour>, keep everything else the same colour."; model: FLUX.2 Klein 9B. Source image from EmoSet [51].

compatible with our weights (w_1, w_2, w_3). Given a diffusion model D_θ, we can reformulate the noise prediction as

$$\begin{aligned} D^{edit}_{\text{bary}}(x_t) =& w_1 D_\theta(x_t, y_1) + w_2 D_\theta(x_t, y_2) + w_3 D_\theta(x_t, y_3), \\ \approx& \nabla_{x_t} \log p_t(x_t) + \sum_{i=1}^{3} w_i \nabla_{x_t} \log p_t(y_i \mid x_t), \end{aligned} \tag{2}$$

where we have conditions y_i for each discrete emotion. Conditions take the form of text prompts, for example, for facial expression editing we use a prompt template such as "Change this persons facial expression to be <emotion> while maintaining the same facial features, composition, lighting, and camera angle". Unlike prompt interpolation, this formulation lets us directly manipulate the directions that increase or decrease the likelihood of specific emotions in y_i according to the implicit classifier $p(y_i|x_t)$, thus allowing for smoother interpolations.

To show this visually, consider the example for continuous colour control in Figure 4. By defining the triangle mesh of different colours shown in Figure 4a, we can control the colour of our edit in two dimensions. However, using weights derived from our mesh to simply combine text embeddings, as shown in Figure 4b, results in large homogeneous patches of colour and sudden changes in colour. Barycentric guidance by comparison, shown in Figure 4c, gives smooth transition colours.

If we want to increase the likelihoods of y_i further we can combine Barycentric guidance with classifier-free guidance (CFG), using the guidance scale w_{cfg}, as

$$(1 - w_{cfg}) D_\theta(x_t) + w_{cfg} \sum_{i=1}^{3} w_i D_\theta(x_t, y_i). \tag{3}$$

We also find that the guidance scale can serve as a useful proxy for emotional intensity. To leverage this we propose adjusting the guidance scale like

$$w^*_{cfg} = 1 + (w_{cfg} - 1)\sqrt{v^2 + a^2}, \tag{4}$$

meaning the effective guidance strength is weak towards the center of the VA model and grows stronger towards more extreme values.

To apply Barycentric Guidance to image generation rather than editing, we can also introduce a base prompt y_{base} (e.g. y_{base}="a man is running fast") that specifies the scene semantics. If we also apply CFG on top we then have

$$\begin{aligned} D^{gen}_{bary}(x_t) =& (w_{cfg} - s) D_\theta(x_t, y_{base}) + (1 - w_{cfg}) D_\theta(x_t) + s \left(\sum_{i=1}^{3} w_i D_\theta(x_t, y_i) \right), \\ \approx& \nabla_{x_t} \log p_t(x_t) + (w_{cfg} - s) \nabla_{x_t} \log p_t(y_{base} \mid x_t) + s \sum_{i=1}^{3} w_i \nabla_{x_t} \log p_t(y_i \mid x_t), \end{aligned} \tag{5}$$

where $s < w_{cfg}$ is a parameter which trades-off the strength of the base prompt against the strength of the emotion prompts.

3.3 Geometry-Aware Latent Interpolation

While Barycentric Guidance provides a robust framework for semantic interpolation, guidance can occasionally struggle with fine grained geometric consistency in highly structured domains, such as human faces. To bridge this gap, we propose an optional structural refinement module. This module complements Barycentric Guidance by explicitly interpolating latent geometries before the final diffusion pass.

Specifically for human faces, we first generate edits for three discrete emotions corresponding to three emotions found in our triangle mesh. We then combine them using face morphing, weighting each image and set of face landmarks according to our barycentric weights. An inbetween image allows us to smoothly interpolate facial geometric, but also creates noticeable ghosting artifacts. To rectify this, we can simply restart the reverse diffusion process part way through using this morphed image in the style of SDEdit [52], effectively creating a structural prior. We find that restarting the diffusion process one-third of the way through the total steps ensures that morphed geometry is preserved without ghosting, as shown in Section A.3. For non-human faces, landmarks cannot be used, however, we find that optical flow based image in betweening works well as a substitute.

4 Experiments

In this section, we evaluate VA-based face and non-face image editing and generation. Unless otherwise stated, we use FLUX.2-dev, though some experiments employ FLUX.2-klein, FLUX.1 Kontext, or Qwen Image Edit 2511. See Section A.9 for hyperparameters and further experimental details.

Table 1: Impact of facial expression diversity (VA range) on FID. Images are taken from AffectNet.

VA Subset	FID ↓
Neutral Only	8.36
$\lvert VA\rvert \leq 0.25$	8.56
$\lvert VA\rvert \leq 0.5$	10.05
$\lvert VA\rvert \leq 0.75$	11.64
Full VA Range	13.40

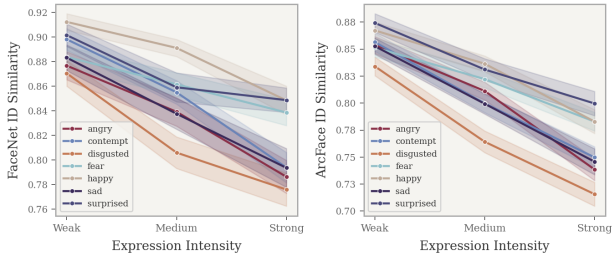


Figure 5: As facial expression diversity increases, FaceNet/ArcFace similarity decrease. Comparing neutral to target emotions across intensity levels, ID similarity drops as expression intensity increases. Images from MEAD dataset (examples shown in Figure 12).

Valence-Arousal Accuracy We measure the accuracy of VA editing and generation using RMSE between predicted and target VA values. To cover the entire VA space we generate 100 images using a 10 by 10 grid of images with uniformly spaced target VA values in the range of $[-1, 1]$. These 10 by 10 grids are also used when calculating smoothness and diversity metrics.

Valence-Arousal Diversity Following EmoStyle [7], we report the standard deviation of the predicted VA values as a measure of facial expression diversity. However, we note that this metric is imperfect, since a model producing only extreme facial expressions registers as more diverse than one producing a mix of extreme and mild expressions. We therefore propose measuring the entropy of predicted VA points using Kernel Density Estimation as a better diversity measure.

For face generation specifically, we note one important caveat when looking at diversity, which is that facial expression diversity trades-off against ID similarity as measured by facial recognition networks like FaceNet [53] or ArcFace [54]. To validate this on real data, we used the MEAD dataset [55], which contains videos of 60 actors displaying basic emotions at three intensity levels (weak, medium, strong), along with neutral expressions. We extract mid-frame images for 48 actors (split 0) and compute cosine similarity between neutral and emotional expressions across intensities. Although identity should ideally remain constant, we observed a consistent decrease in ID similarity with increasing expression strength (Figure 5), confirming a strong identity-expression trade-off.

Table 2: Facial Expression Editing comparison with other VA editing models. We treat Entropy, VA Std, RMSE, and S^*_{2D} as the primary controllability metrics. †: The method does not control VA directly; instead, it approximates VA editing using polar coordinates to steer emotion.

Dataset	Model	VA Controllability				Image Quality/ Preservation			
		RMSE↓	VA Entropy↑	VA Std↑	S^*_{2D}↓	FID↓	IS↑	ID Sim.↑	LPIPS↓
AffectNet	GANmut†	-	0.616	0.47 / 0.38	-	36.89	3.90	**0.76**	0.128
	EmoStyle	0.610	0.423	0.46 / 0.31	**0.147**	36.72	2.81	0.31	0.276
	Ours (Qwen)	**0.431**	0.813	0.55 / 0.49	0.278	**10.67**	**4.71**	0.72	**0.062**
	Ours	0.433	**0.916**	**0.64 / 0.57**	0.303	11.02	4.42	0.55	0.096
FFHQ	GANmut†	-	0.687	0.50 / 0.36	-	73.20	4.49	**0.71**	0.282
	EmoStyle	0.568	0.457	0.48 / 0.32	**0.148**	25.50	4.12	0.67	0.172
	Ours	**0.418**	**0.890**	**0.64 / 0.51**	0.307	**13.85**	**5.61**	0.55	**0.097**
CelebA-HQ	GANmut†	-	0.651	0.48 / 0.37	-	59.59	3.24	**0.70**	0.267
	EmoStyle	0.550	0.513	0.47 / 0.34	**0.149**	17.44	2.94	0.65	0.154
	Ours	**0.412**	**0.877**	**0.65 / 0.54**	0.293	**11.80**	**3.53**	0.54	**0.125**

Valence-Arousal Smoothness We evaluate smoothness over the full VA grid using a Laplacian-based curvature measure. Consider a function $z(x, y)$ which maps from a point on a square lattice (x, y) to a valence arousal point (v, a). We estimate local curvature via a discrete Laplacian, approximated using finite differences, and then measure smoothness as

$$S^*_{\text{2D}} = \frac{1}{(W-2)(H-2)} \sum_{i=2}^{H-1} \sum_{j=2}^{W-1} \sqrt{\left(\frac{\Delta z^v(x,y)}{r_v}\right)^2 + \left(\frac{\Delta z^a(x,y)}{r_a}\right)^2}, \tag{6}$$

$$\Delta z(x,y) = z(x+1,y) + z(x-1,y) + z(x,y+1) + z(x,y-1) - 4z(x,y), \tag{7}$$

where normalizing by the ranges of each dimension r_v and r_a ensures that this measurement is scale invariant. That is, we only consider the smoothness relative to the overall change in expression, which avoids methods that produce little to no change in the emotion being registered as smoother.

Image Quality Finally, we report Fréchet Inception Distance (FID) [56] and Inception Score (IS) [57] to measure image quality, calculated using 3.5k images. As with ID similarity, we find that facial expression diversity can impact FID negatively. To study this, we formed six 4k-image subsets of AffectNet (Neutral A/B and four quasi-uniform samples with VA magnitude bounds of ≤ 0.25, ≤ 0.5, ≤ 0.75, and ≤ 1.0) and computed FID against Neutral A (Table 1). The rising FID scores (8.36 to 13.40) suggest that the metric penalizes expression distribution shifts in addition to measuring realism.

4.1 Facial Expression Editing

Competing Methods We compare identity-preserving facial expression editing against the two state-of-the-art methods GANmut [6] and EmoStyle [7]. EmoStyle edits StyleGAN [58] latents (trained on FFHQ) to control valence and arousal (VA), while GANmut learns its own two dimensional space from AffectNet. To better understand the generalization of each approach, we calculate metrics across three human face datasets: AffectNet [49], FFHQ [58], and CelebA-HQ [4]. Following the recommendation from Darne et al. [10], we use EmotiEffLib [59] to estimate VA values of each image. In Table 2, we see that our method outperforms competing methods on image quality, VA accuracy, expression diversity, and background preservation (LPIPS [60] on non-face regions). Building on the base models robust image quality and background preservation (indicated by FID, IS, and LPIPS), our approach maximizes VA accuracy and diversity while mitigating issues with edit smoothness. Our approach still lags behind in ID similarity, measured using ArcFace, which may in part be explained by our greater expression diversity. In line with lower LPIPS scores, Figure 6 demonstrates better background preservation (e.g., hair, accessories) alongside better coverage of extreme expressions, especially in the top-left VA quadrant (RMSE down 0.337 vs. EmoStyle) and for "sleepy" expressions, which, as noted by Darne et al. [10], prior work struggles to capture. See Section A.6 for additional comparison to the closed source GPT-Image-2 model.

We also implement our approach on Qwen Image Edit 2511 and include results in Table 2 as an additional point of comparison. This implementation beats prior approaches in terms of VA

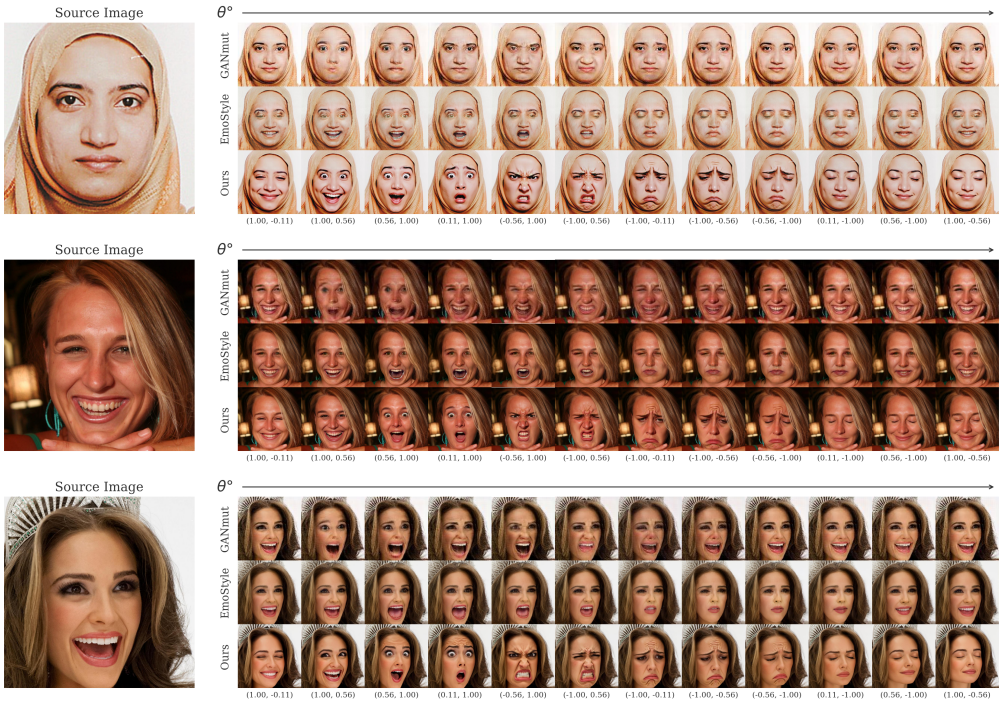


Figure 6: A sweep of VA values around the circumference of the unit circle, where θ is the angle. GANmut, EmoStyle, and Our edits are shown across three datasets AffectNet (top), FFHQ (middle), and CelebA (bottom). Target VA values (shown below columns) apply to EmoStyle and Ours only.

accuracy and diversity, but also improves FID, IS, and ID similarity compared the FLUX.2-based implementation of our method. This demonstrates the potential of our approach to generalize across image editing models, enabling it to benefit from new model releases without retraining.

Non-Human Faces We show VA-based editing results in Figure 1 and 8 for a range of datasets [4, 2, 61, 3]. This shows that our method can be applied to animal faces, cartoon faces, and 3D models of faces as well as to regular human face images. In cases where face morphing cannot be applied, we use WAFT [62] to find the optical flow.

Ablation Study In Table 3, we perform an ablation study by removing or substituting components from our full approach. The full method achieves the best overall VA accuracy, diversity, and smoothness. Removing scaled CFG leads to a drop in VA accuracy and diversity. Substituting Barycentric Guidance with prompt interpolation further degrades VA control, significantly worsening RMSE and smoothness. Finally, removing face morphing results in the lowest smoothness and hence lower diversity. While this setting attains the highest image quality and identity similarity by staying closer to the input, it fails to produce precise affect changes. Overall, these results highlight the diversity–fidelity trade-off: removing components improves FID and ID similarity (by virtue of staying closer to the source), but our full approach is required to maximize VA coverage and control. Despite this, our method still outperforms prior VA-editing approaches on image quality metrics (see Table 2), while achieving state-of-the-art VA accuracy, diversity, and smoothness.

User Study We conduct a user study with 66 participants comparing EmoStyle and our method across 50 CelebA edits. See Section C for full experimental setup. Results in Table 4 show a significant preference ($p < 0.05$, Wilcoxon signed-rank test) for our approach on emotion accuracy (Q1) and non-affective content preservation (Q4). These results provide direct human evidence that our approach more accurately achieves the target affect while better preserving non-affective aspects of the image. While EmoStyle is preferred for realism (Q2) and identity (Q3), this is explained by edit magnitude: EmoStyle's average VA change of $|\Delta\text{VA}| = 0.46$ is much lower than ours, 0.91, against a target of 0.86 (see Figure 23). This suggests EmoStyle remains closer to the input, which is

Table 3: Ablation Study on AffectNet. ⇆: replaces Bary. Guidance with Prompt Interp.

Setting	VA Controllability				Image Quality / Preservation		
	RMSE ↓	VA Entropy ↑	VA Std ↑	S^*_{2D} ↓	FID ↓	IS ↑	ID Sim. ↑
Full Approach	**0.433**	**0.916**	**0.64 / 0.57**	**0.303**	11.02	4.42	0.55
– Scaled CFG	0.439	0.906	0.59 / 0.51	0.305	11.10	4.56	0.61
⇆ Bary. Guidance → Prompt Interp.	0.461	0.869	0.57 / 0.49	0.342	10.75	4.50	0.55
– Face Morphing	0.451	0.756	0.56 / 0.50	0.377	**8.94**	**4.62**	**0.66**

consistent with higher realism and ID scores, whereas our method performs larger, more targeted affect changes, aligning with its advantage on VA accuracy and non-affective content preservation (see also examples in Figure 22). Due to cost, we do not conduct a full VA-plane user study. Instead, we evaluate a representative subset, with broader VA coverage and continuity assessed via grid-based quantitative metrics (see Tables 2 & 6) and qualitative sweeps (see Figures 1, 6, 8, 9, 24 & 28).

Continuous Editing Comparison Finally, we compare against other methods of continuous control for image editing models. Because existing approaches operate exclusively in the 1D setting, we restrict our evaluations to 1D interpolations for a direct comparison. We compare against Kontinuous Kontext [26] (finetuned FLUX.1 Kontext), prompt interpolation, CFG-based control, and FluxSpace [63] (ported to FLUX.1 Kontext). For fair comparison, we reimplement our method on FLUX.1 Kontext. To measure smoothness of the edit we adapt the smoothness metric from [6] which involves interpolating between neutral and a target emotion over 10 steps. We measure LPIPS between consecutive images, reporting the max-to-mean LPIPS ratio, and additionally report RMSE using predicted VA values from EmotiEffLib. As shown in Table 5 and Figure 7, our method yields smoother edits and lower RMSE, outperforming text-based, attention-based, and finetuning-based approaches.

Table 4: User study win rates between our approach and EmoStyle on CelebA facial expression editing. See Section C for exact question wording.

Question	EmoStyle	Ours	Equal	p-value
(1) VA Accuracy	9	40	1	8.1e-6
(2) Realism	33	15	2	2.9e-4
(3) ID Similarity	41	8	1	8.6e-7
(4) Non-Face Pres.	11	36	3	6.4e-5

Table 5: Continuous editing comparison on FLUX.1-Kontext for one dimensional editing.

Method	RMSE	Smooth.	C-LPIPS
Prompt Interp.	0.310	0.707	3.679
CFG	0.316	0.720	3.987
FluxSpace	0.476	0.686	3.490
Kontinuous Kontext	0.321	0.665	3.552
Ours	**0.292**	**0.643**	**1.922**



Figure 7: One dimensional continuous editing for FLUX.1 Kontext.

4.2 Non-Face Emotional Image Editing & Generation

In this section, we extend our approach to two non-face datasets: EmoArt [5] and FindingEmo [64].

EmoArt EmoArt contains 130k images of artworks labeled with emotions designed to evenly cover the VA spectrum. Application of our method to EmoArt, shown in Figure 1e, demonstrates that our VA-based image editing approach can effectively alter scenes, expressions, body language, and overall semantics. Examples include adding angel wings for high valence/arousal, darkening scenes for sad emotions, portraying slumping postures for low arousal, and rendering angry expressions. Quantitative results in Table 6 show that, compared to prompt interpolation on the base model, our approach achieves superior alignment with the VA-model.

Table 6: Non-face results. This includes VA-based editing of artworks on EmoArt [5] and VA-based generation of scenes using FindingEmo [64].

Dataset	Method	VA Controllability				Image Quality / Preservation	
		RMSE ↓	VA Entropy ↑	VA Std ↑	S^*_{2D} ↓	FID ↓	IS ↑
EmoArt	Prompt Interp.	0.851	0.129	0.35 / 0.39	0.502	74.85	**9.64**
	Ours	**0.735**	**0.327**	**0.36 / 0.42**	**0.483**	**67.96**	9.08
FindingEmo	EmotiCrafter	0.796	**0.298**	0.38 / **0.25**	0.540	233.30	8.85
	Ours (klein)	**0.783**	0.163	**0.40** / 0.21	**0.497**	**221.51**	**10.68**

FindingEmo FindingEmo contains 25k images of scenes labeled with a single VA value for the whole scene. We compare our approach against EmotiCrafter [8], which generates images of scenes based on a text prompt and a VA label. We use FLUX.2 Klein 9B as our base model. In Table 6 we get competitive results, outperforming EmotiCrafter on both smoothness and image quality, while also requiring no training. Qualitative results are shown in Figure 9a. We leave additional analysis of scene factors like lighting, composition, and pose as a compelling direction for future work.

5 Conclusion

In this work, we present an approach that enables the foundation image editors to perform valence-arousal-based editing and generation across a range of domains. To this end, we approximate the VA model using a triangle mesh. From this we derive weights, which can then be used for both guidance-based and latent-space-based editing. Results suggest this approach can achieve competitive performance across a wide variety of domains, without requiring retraining the foundation models.

Limitations & Future Work Although the proposed VA mesh can represent complex or mixed states to the extent that they are locally interpolable in VA space, the VA model does not capture the full spectrum of human emotion. For example, compound emotions such as bittersweet involve the simultaneous experience of happiness and sadness, which cannot be adequately represented by a single point in the two-dimensional VA space. In the supplementary materials, we present a simple extension to Barycentric Guidance that utilizes multiple VA points, thereby enabling the representation of compound emotions such as bittersweetness. These initial exploratory results suggest a promising direction for future work. Another avenue would be to extend the Barycentric Guidance framework to alternative models of emotion, such as the VAD model or appraisal-based models.

While our approach does not require training, it may require some adaptation for new domains. Specifically, we found that different domains require different editing prompts. For domains that are more discrete, such as images of text, our continuous editing approach would also not be appropriate.

Finally, while our approach avoids expensive finetuning on individual images, our guidance and latent space editing algorithms add compute time in the form of increased diffusion steps. Future work may look to mitigate this via guidance-based distillation methods or application to few step image editing models.

Acknowledgments and Disclosure of Funding

This research was mainly supported by Studentship from the Faculty of Engineering and Physical Sciences - Electronic and Computer Science at the University of Southampton. The authors acknowledge the use of the IRIDIS X High Performance Computing Facility and the Southampton-Wolfson AI Research Machine (SWARM) GPU cluster, funded by the Wolfson Foundation, together with the associated support services at the University of Southampton in the completion of this work.

References

[1] James A Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161, 1980.

[2] Mert Köklü. Manga facial expressions. https://www.kaggle.com/datasets/mertkkl/manga-facial-expressions, 2021. Accessed: 20/04/2026.

[3] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8188–8197, 2020.

[4] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[5] Cheng Zhang, Hongxia Xie, Bin Wen, Songhan Zuo, Ruoxuan Zhang, and Wen-Huang Cheng. Emoart: A multidimensional dataset for emotion-aware artistic generation. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 12644–12650, 2025.

[6] Stefano d'Apolito, Danda Pani Paudel, Zhiwu Huang, Andrés Romero, and Luc Van Gool. Ganmut: Learning interpretable conditional space for gamut of emotions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 568–577, 2021.

[7] Bita Azari and Angelica Lim. Emostyle: One-shot facial expression editing using continuous emotion parameters. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6385–6394, 2024.

[8] Shengqi Dang, Yi He, Long Ling, Ziqing Qian, Nanxuan Zhao, and Nan Cao. Emoticrafter: Text-to-emotional-image generation based on valence-arousal model. *arXiv preprint arXiv:2501.05710*, 2025.

[9] Kazuaki Mishima, Antoni Bigata Casademunt, Stavros Petridis, Maja Pantic, and Kenji Suzuki. Facecrafter: Identity-conditional diffusion with disentangled control over facial pose, expression, and emotion. *arXiv preprint arXiv:2505.15313*, 2025.

[10] Clément Gérard Daniel Darne, Changqin Quan, and Zhiwei Luo. Generating synthetic facial expression images using emostyle. *Applied Sciences*, 15(19):10636, 2025.

[11] 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.

[12] Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-image technical report, 2025. URL https://arxiv.org/abs/2508.02324.

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

[14] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[15] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[16] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

[17] Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *Advances in neural information processing systems*, 37:84839–84865, 2024.

[18] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

[19] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[20] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

[21] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

[22] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

[23] Black Forest Labs. Flux. `https://github.com/black-forest-labs/flux`, 2024.

[24] Lvmin 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.

[25] 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.

[26] 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.

[27] Jiabin Hua, Hengyuan Xu, Aojie Li, Wei Cheng, Gang Yu, Xingjun Ma, and Yu-Gang Jiang. Pixelsmile: Toward fine-grained facial expression editing. *arXiv preprint arXiv:2603.25728*, 2026.

[28] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8789–8797, 2018.

[29] Albert Pumarola, Antonio Agudo, Aleix M Martinez, Alberto Sanfeliu, and Francesc Moreno-Noguer. Ganimation: Anatomically-aware facial animation from a single image. In *European Conference on Computer Vision*, pages 818–833, 2018.

[30] Paul Ekman. Universals and cultural differences in facial expressions of emotion. In *Nebraska symposium on motivation*. University of Nebraska Press, 1971.

[31] Paul Ekman and Wallace V Friesen. Facial action coding system. *Environmental Psychology & Nonverbal Behavior*, 1978.

[32] Zheng Ding, Xuaner Zhang, Zhihao Xia, Lars Jebe, Zhuowen Tu, and Xiuming Zhang. Diffusionrig: Learning personalized priors for facial appearance editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12736–12746, 2023.

[33] Jiayu Wang, Kang Zhao, Yifeng Ma, Shiwei Zhang, Yingya Zhang, Yujun Shen, Deli Zhao, and Jingren Zhou. Facecomposer: A unified model for versatile facial content creation. *Advances in Neural Information Processing Systems*, 36, 2024.

[34] Yue Han, Junwei Zhu, Keke He, Xu Chen, Yanhao Ge, Wei Li, Xiangtai Li, Jiangning Zhang, Chengjie Wang, and Yong Liu. Face adapter for pre-trained diffusion models with fine-grained id and attribute control. *arXiv preprint arXiv:2405.12970*, 2024.

[35] Tuomas Varanka, Huai-Qian Khor, Yante Li, Mengting Wei, Hanwei Kung, Nicu Sebe, and Guoying Zhao. Towards localized fine-grained control for facial expression generation. *arXiv preprint arXiv:2407.20175*, 2024.

[36] Wenliang Zhao, Yongming Rao, Weikang Shi, Zuyan Liu, Jie Zhou, and Jiwen Lu. Diffswap: High-fidelity and controllable face swapping via 3d-aware masked diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8568–8577, 2023.

[37] Yifang Xu, Benxiang Zhai, Yunzhuo Sun, Ming Li, Yang Li, and Sidan Du. Hifi-portrait: Zero-shot identity-preserved portrait generation with high-fidelity multi-face fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5625–5635, 2025.

[38] Reni Paskaleva, Mykyta Holubakha, Andela Ilic, Saman Motamed, Luc Van Gool, and Danda Paudel. A unified and interpretable emotion representation and expression generation. *arXiv preprint arXiv:2404.01243*, 2024.

[39] Foivos Paraperas Papantoniou, Alexandros Lattas, Stylianos Moschoglou, Jiankang Deng, Bernhard Kainz, and Stefanos Zafeiriou. Arc2face: A foundation model of human faces. *arXiv preprint arXiv:2403.11641*, 2024.

[40] 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. *arXiv preprint arXiv:2106.09685*, 2021.

[41] Jingyuan Yang, Jiawei Feng, Weibin Luo, Dani Lischinski, Daniel Cohen-Or, and Hui Huang. Emoedit: Evoking emotions through image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24690–24699, 2025.

[42] Qing Lin, Jingfeng Zhang, Yew-Soon Ong, and Mengmi Zhang. Make me happier: Evoking emotions through image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16367–16376, 2025.

[43] Peixuan Zhang, Shuchen Weng, Chengxuan Zhu, Binghao Tang, Zijian Jia, Si Li, and Boxin Shi. Affective image editing: Shaping emotional factors via text descriptions. *International Journal of Computer Vision*, 134(1):16, 2026.

[44] Jingyuan Yang, Jiawei Feng, and Hui Huang. Emogen: Emotional image content generation with text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6358–6368, 2024.

[45] Jingyuan Yang, Zihuan Bai, and Hui Huang. Emostyle: Emotion-driven image stylization. *arXiv preprint arXiv:2512.05478*, 2025.

[46] 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.

[47] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

[48] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

[49] 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*, 10(1):18–31, 2017.

[50] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

[51] Jingyuan Yang, Qirui Huang, Tingting Ding, Dani Lischinski, Danny Cohen-Or, and Hui Huang. Emoset: A large-scale visual emotion dataset with rich attributes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20383–20394, 2023.

[52] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021.

[53] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 815–823, 2015.

[54] 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.

[55] 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.

[56] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

[57] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016.

[58] 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.

[59] Andrey Savchenko. Facial expression recognition with adaptive frame rate based on multiple testing correction. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pages 30119–30129. PMLR, 23–29 Jul 2023. URL `https://proceedings.mlr.press/v202/savchenko23a.html`.

[60] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018.

[61] Deepali Aneja, Alex Colburn, Gary Faigin, Linda Shapiro, and Barbara Mones. Modeling stylized character expressions via deep learning. In *Asian Conference on Computer Vision*, pages 136–153. Springer, 2016.

[62] Yihan Wang and Jia Deng. Waft: Warping-alone field transforms for optical flow. *arXiv preprint arXiv:2506.21526*, 2025.

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

[64] Laurent Mertens, Elahe Yargholi, Hans Op de Beeck, Jan Van den Stock, and Joost Vennekens. Findingemo: An image dataset for emotion recognition in the wild. *Advances in Neural Information Processing Systems*, 37:4956–4996, 2024.

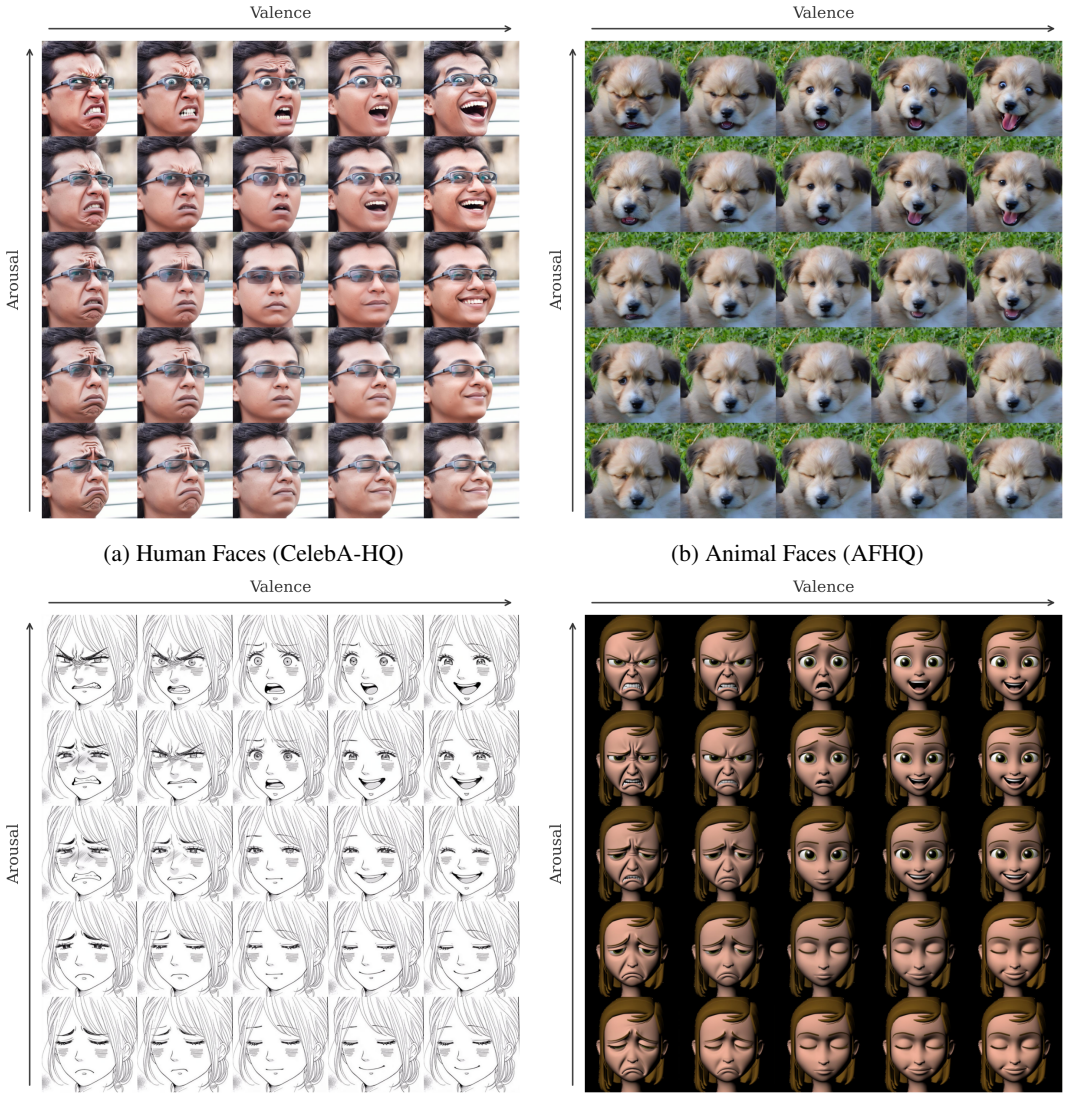
[65] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, et al. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*, 2019.

[66] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[67] Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

A Appendix

Here we provide more details about our method, including additional derivations, experiments, results, and experimental details. In Section A.1 we present qualitative results with additional datasets and models. Section A.2 provides more derivations and experimental details that informed our metrics. Section A.3 presents an experiment exploring the optimal step at which to restart diffusion. Section A.4 extends the Barycentric Guidance method to capture compound emotions. Section A.5 explores the controllability-fidelity trade-off for facial expression editing. Section A.6 compares our approach to a leading closed-source model. Section A.7 details the relative improvements that were discussed in the abstract. Section A.8 goes over the derivation of Equations 2 and 5. Section A.9 contains any experimental details we were not able to fit into the main body of the paper.



(a) Human Faces (CelebA-HQ)

(b) Animal Faces (AFHQ)

(c) Cartoons (Manga Facial Expressions)

(d) 3D Models (FERG-DB)

Figure 8: Our approach can perform valence arousal based emotional image editing across a wide variety of domains. VA range [-1,1] with uniform step size for rows and columns.

A.1 Qualitative Results

In this section, we provide additional qualitative results. In Figure 8, we provide 5 by 5 grids showing a range of valence arousal edits for human face datasets and non-human face datasets. These include



(a) FindingEmo (b) EmoArt

Figure 9: Qualitative results for non-face emotion image editing and generation. The base prompt used for **(a)** is "a man in a green jacket". VA range [-1,1] with uniform step size for rows and columns

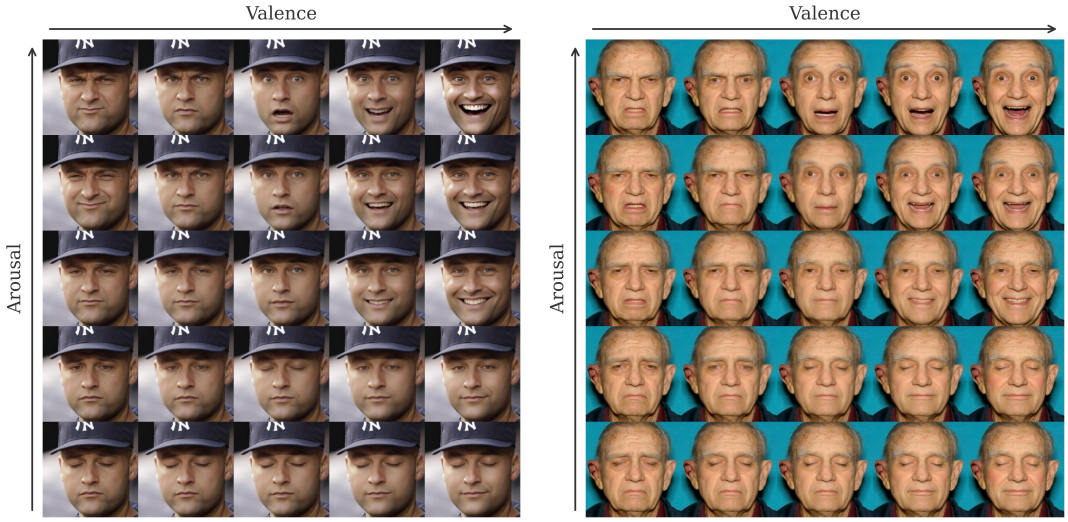


Figure 10: Qualitative results on AffectNet using Qwen Image Edit 2511 as our base model. VA range [-1,1] with uniform step size for rows and columns

CelebA [4], AFHQ [3], FERG-DB [61] and Manga Facial Expressions [2]. In Figure 9, we show qualitative results for non-face datasets FindingEmo [64] and EmoArt [5].

In Figure 10, we apply our full approach to Qwen Image Edit 2511 [12]. These qualitative results show that our method can be applied beyond the FLUX family of models. Throughout the paper, we apply our approach to FLUX.2 dev, FLUX.2 klein, FLUX.1 Kontext, and Qwen Image Edit 2511, suggesting that our approach can be applied broadly across diffusion-based foundation image editing models.

We also provide a further qualitative comparison between our method and EmoStyle [7] on AffectNet [49], CelebA [4], and FFHQ [58] in Figures 28, 29, 26, 27, 24, and 25 for the full 10 by 10 grids used for VA focused metrics. These figures demonstrate more advantages of our method. Firstly, our approach is much better able to represent low arousal emotions. Secondly, our approach is better able to handle occlusion of the face convincingly. Finally, we find that EmoStyle does not deal with

cropped faces well, which are common in the AffectNet dataset. Our approach on the other hand is robust to the cropping of the face.

A.2 More Information on Metrics

This section provides further experimentation and discussion to further support our choice of metrics.

Smoothness To evaluate the smoothness of expression editing over the entire VA space we propose using a Laplacian to measure the curvature of VA predictions of a 10 by 10 grid of image with different uniformly spaced target VA targets. Consider a function $z(x, y)$ which maps from a point on a square lattice (x, y) to a valence arousal point (v, a). From this we can estimate the curvature using a discrete Laplacian, approximated using finite differences, as

$$\Delta z(x, y) = z(x + 1, y) + z(x - 1, y) + z(x, y + 1) + z(x, y - 1) - 4z(x, y). \tag{8}$$

To measure the overall smoothness of our valence and arousal surface, we take the magnitude of these Laplacians

$$S_{2\mathrm{D}} = \frac{1}{(W-2)(H-2)} \sum_{i=2}^{H-1} \sum_{j=2}^{W-1} \sqrt{\Delta z^v(x, y)^2 + \Delta z^a(x, y)^2}, \tag{9}$$

where W and H are the width and height of the grid (i.e. $W = 10$ and $H = 10$), and Δz^v and Δz^a are the valence and arousal components of Δz, respectively. However, if we reduce the range of one dimension, we will reduce the smoothness, even though the rate of change with respect to the total range remains unchanged. In practice, as seen in the toy example in Figure 11, the same set as valence arousal points will results is different S_{2D} values when scaled up. To account for the fact that different models will generated images with different ranges of valence and arousal predictions, we can scale the valence values by the range of valence values r_v, and the arousal values by the range of arousal values r_a. This allows to define a scale-invariant version of the S_{2D} metric

$$S^*_{2\mathrm{D}} = \frac{1}{(W-2)(H-2)} \sum_{i=2}^{H-1} \sum_{j=2}^{W-1} \sqrt{\left(\frac{\Delta z^v(x, y)}{r_v}\right)^2 + \left(\frac{\Delta z^a(x, y)}{r_a}\right)^2}. \tag{10}$$

Now consider a scaled version of the same valence arousal points. We can scale the valence dimension by α and the arousal dimension by β, which means our estimate of curvature becomes

$$\begin{pmatrix}\alpha\\ \beta\end{pmatrix} \cdot \Delta z_{i,j} = \begin{pmatrix}\alpha\\ \beta\end{pmatrix} \cdot (z_{i+1,j} + z_{i-1,j} + z_{i,j+1} + z_{i,j-1} - 4z_{i,j}). \tag{11}$$

Since scaling the the dimensions also scales the minimum and maximum values by the same amount, the range is also scaled by the same amount. Our smoothness metric is then

$$S^*_{2\mathrm{D}} = \frac{1}{(W-2)(H-2)} \sum_{i=2}^{H-1} \sum_{j=2}^{W-1} \sqrt{\left(\frac{\alpha\Delta z^v(x, y)}{\alpha r_v}\right)^2 + \left(\frac{\beta\Delta z^a(x, y)}{\beta r_a}\right)^2}. \tag{12}$$

The scaling factors therefore cancel out for both dimensions leaving us with our original $S^*_{2\mathrm{D}}$ metric. This ensures that our metric is scale invariant and does not unfairly punish more diverse sets of valence arousal points.

Divserity Here we further verify the trade-off between facial expression diversity and ID similarity metrics. In Figure 13 we examine this trade-off on synthetic images using 3500 images generated from AffectNet validation images with random VA values with our full method, except with 14 inference steps and 7 anchor steps. We measure ID similarity using FaceNet and ArcFace between the source and generated images. We reduce the entropy by calculating the radius of each predicted VA values then gradually removing the highest radius points over the course of 50 steps. Once points have been removed, we recalculate the entropy of the remaining VA points and also take the average ArcFace and FaceNet similarity of the remaining points. Figure 13 shows that for the synthetic images, ID similarity clearly trades-off against facial expression diversity. As well as posing problems for evaluation, this also could also cause potential problems when optimizing a facial expression editing model using a facial recognition model.

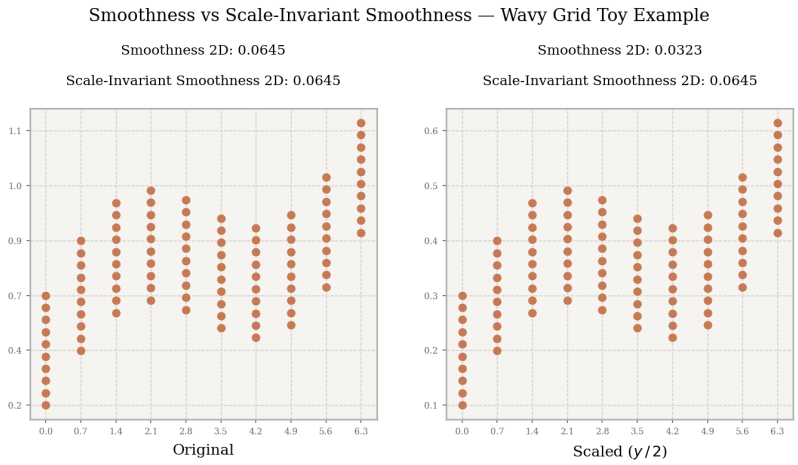
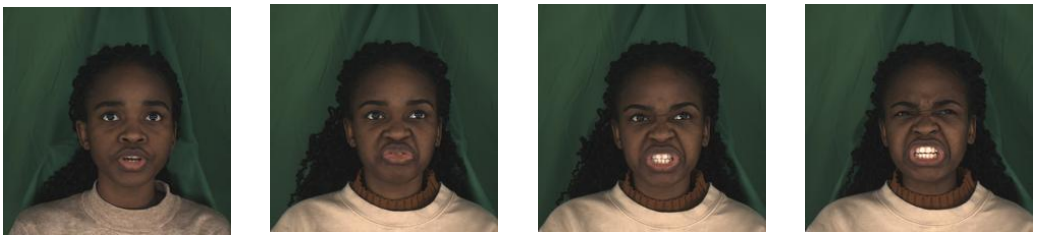


Figure 11: Smoothness Metric Toy Example



Neutral | Disgust (Weak) | Disgust (Medium) | Disgust (Strong)

Figure 12: Comparison of facial expression intensity levels in the MEAD dataset, demonstrating the incremental progression from neutral to strong disgust.

A.3 Morph Step

We explore at which step is best to restart the diffusion process when using geometry-aware latent interpolation. In Figure 14, we edit faces from AffectNet, restarting the diffusion process at $t = \{3, 5, 7, 9, 11, 13\}$ with a total of 15 inference steps. While the morphing step can introduce ghosting artifacts, the reverse diffusion process is capable of fixing this provided we give it enough steps. When $t = 13$, ghosting artifacts are common. Working backwards, we see that by $t = 5$ these artifact have been rectified. We therefore set the morph step to be a third of the way through the reverse diffusion process through out the paper.

A.4 Compound Emotions

As a straightforward extension, Barycentric Guidance can combine two distant VA targets by computing guidance for each and blending them with $\gamma \in [0, 1]$. This can allow for the generation of compound emotions that could not be easily represented using the VA model on its own. Mathematically, this naturally expands Equation 2 to $\gamma(w_1 D_1 + w_2 D_2 + w_3 D_3) + (1 - \gamma)(w_4 D_4 + w_5 D_5 + w_6 D_6)$. Preliminary examples on happy–sad, sad–angry, and happy–disgust show mixed classifier activations consistent with the intended compound emotions. These examples use the exact VA location of the specific target emotions to derive the weights. CelebA images are used as the input images. Qualitative results for these compound emotions are shown in Figure 15.

A.5 Controllability–Fidelity Trade-Off

As discussed in the main body, stronger VA control introduces a controllability–fidelity trade-off. It is not only due to larger edit magnitude, although edit magnitude is an important factor. To better disentangle this, we conducted a matched edit-magnitude analysis on CelebA with the following steps: (1) Pair up EmoStyle edits against Our edits based on having the same or similar (< 0.01

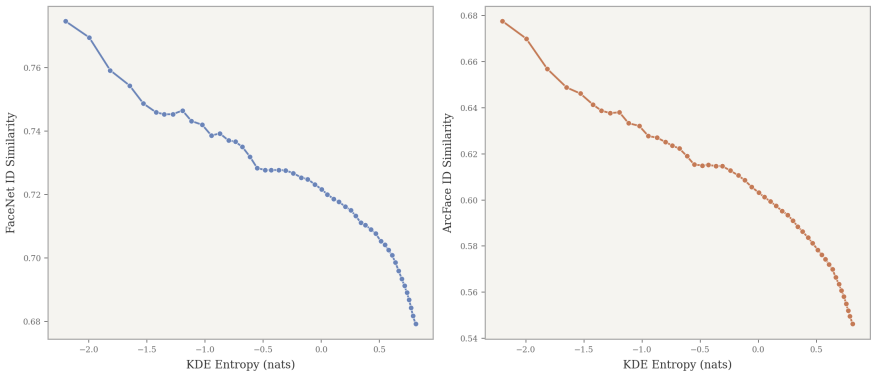


Figure 13: On synthetic images, diversity of facial expression trades-off against ID similarity as measured by FaceNet and ArcFace.



Figure 14: Comparison of different morph step on AffectNet images. The morph step refers to the step at which we restart the diffusion process after face morphing is performed.

distance) edit VA magnitudes. This essentially filters our images so that the distribution of VA edit strengths is the same for both subsets. (2) From these matched pairs, we further filter to strictly include only edits where the absolute VA magnitude is less than 0.2. (3) Recalculate Inception Score (IS) and ArcFace ID similarity for both subsets (yielding a subset of 465 images for each method; FID is omitted here as this sample size is too small for a stable calculation).

Result are shown in Table 7. These results suggest that when edit magnitudes are restricted to comparable, low-intensity ranges, our approach outperforms EmoStyle in both image quality and

Table 7: Restricted edit strength magnitude experiment on EmoStyle and our approach. We compare 465 images with the same small edit strengths (less than 0.2 change in VA distance) the recalculate relevant metrics on this smaller subset. CelebA used for all source images.

Method	IS	ID Similarity
EmoStyle	2.85	0.666
Ours	3.40	0.686



(a) Happy-Sad. (b) Happy-Disgust. (c) Anger-Sad.

Figure 15: Examples of compound emotions generated using two VA points with FLUX.2-dev used as a backbone.

Table 8: Facial Expression Editing comparison between GPT-Image-2 and our approach.

Model	**VA Controllability**				Image Quality/ Preservation			
	RMSE↓	VA Entropy↑	VA Std↑	S^*_{2D}↓	FID↓	IS↑	ID Sim.↑	LPIPS↓
GPT-Image-2	0.530	0.327	0.52 / 0.35	0.402	**9.38**	**4.71**	**0.84**	**0.066**
Ours	**0.433**	**0.916**	**0.64 / 0.57**	**0.303**	11.02	4.42	0.55	0.096

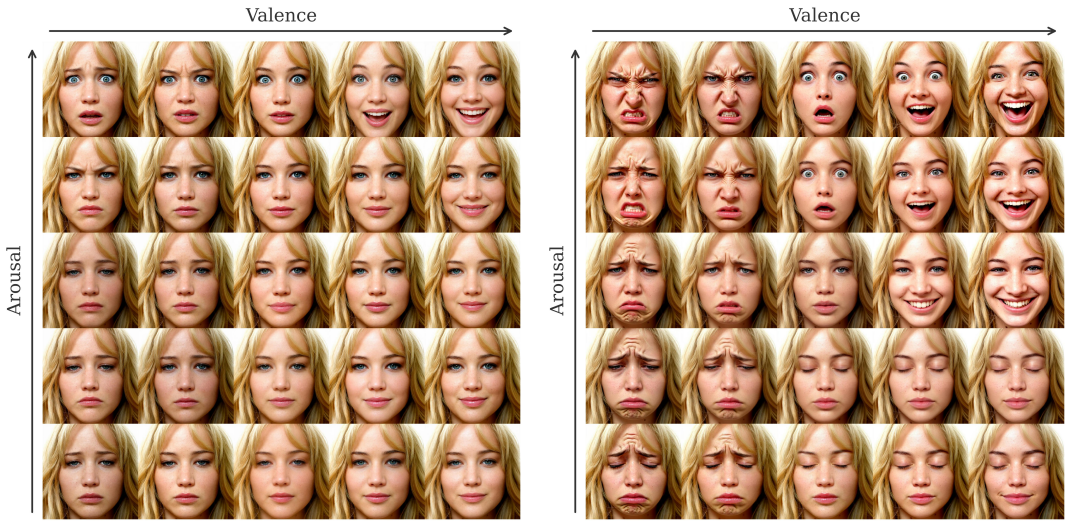
identity preservation. This suggests that the reduced ID similarity observed in the unrestricted user study is a consequence of our method successfully reaching extreme, high-magnitude affective states that EmoStyle is incapable of generating, rather than an inherent flaw in our framework.

Combining this with prior evidence from the main body of the paper allows us to be more precise when analysing the controllability-fidelity trade-off within our approach. Barycentric Guidance provides stronger human-perceived affective controllability, but this comes with a controllability–fidelity trade-off mainly driven by edit magnitude. It is also worth highlighting that our experiments on Qwen Image Edit obtain substantially improved fidelity (FID 10.67, IS 4.71, ID similarity 0.72, LPIPS 0.062) and the second best VA performance (RMSE 0.43, Entropy 0.81, VA std 0.55/0.49, S^*_{2D} 0.28) compared to the FLUX.2 variant of our method. This suggest the underlying editing backbone also plays a significant role in this trade-off.

A.6 GPT-Image-2

As an additional point of comparison, we compare our approach with one of the leading closed source image generators (as of July 2026). Specifically, we evaluate GPT-Image-2's performance on continuous facial expression editing using AffectNet images and the same evaluation setup as in Table 2. However, GPT-Image-2 accepts only text and image inputs. After initial prompt exploration, we found our best performing prompt to be "Edit the provided face image to match target valence-arousal [<valence>, <arousal>]. Preserve the person's identity exactly: same face shape, eyes, nose, mouth structure, skin texture, age, hairstyle, facial hair, accessories, pose, camera angle, lighting, background, and image quality. Only change facial expression muscles enough to express the target affect. Valence is negative-to-positive emotion. Arousal is calm-to-activated energy. Return a photorealistic edited image of the same person. Valence and arousal range from -1.0 to 1.0." Here, <valence> and <arousal> are the target VA values.

Table 8 shows the performance of GPT-Image-2 compared with our approach, which uses FLUX.2-dev as the backbone. While GPT-Image-2 performs extremely well on image quality and identity preservation metrics, it lags behind on all VA controllability metrics, with our approach outperforming GPT-Image-2 in VA accuracy and diversity, as well as VA smoothness. A plausible explanation is that text provides a poor interface for numerical and continuous manipulation of emotions, whereas our latent-space-based approach allows for gradual and controlled changes in emotions. In Figure 16 we present a visual comparison between our approach (with FLUX.2-dev as our backbone) and GPT-Image-2 on AffectNet facial expression editing. In line with the results in Table 8, we see a noticeable increase in diversity while also better capturing extreme emotions.



(a) GPT-Image-2 (b) Ours (FLUX.2-dev)

Figure 16: Qualitative comparison between our approach and GPT-Image-2 for AffectNet facial expression editing. VA range [-1,1] with uniform step size for rows and columns

Table 9: Relative improvements over the strongest baseline. Numbers are derived from the results in Table 2.

Dataset	Entropy ↑	VA Std (V) ↑	VA Std (A) ↑	RMSE ↓
AffectNet	+48.7%	+36.2%	+50.0%	+29.0%
FFHQ	+29.5%	+28.0%	+41.7%	+26.4%
CelebA-HQ	+34.7%	+35.4%	+45.9%	+25.1%

A.7 Relative Improvements

Here we provide additional information regarding the relative percentage improvements presented in the abstract. Table 9 provides the calculation breakdown. Each percentage is computed against the strongest baseline for the corresponding dataset and metric.

A.8 Derivations

In this section, we include the following derivation/proof sketch for Equations 2 and 5. These derivations are based on observations from Song et al. [18] and Ho and Salimans [50] that diffusion models approximate the score function and that Bayes can be applied to conditional generation, respectively. Equation 2 is given as:

$$D^{edit}_{bary}(x_t) = w_1 D(x_t, y_1) + w_2 D(x_t, y_2) + w_3 D(x_t, y_3) \tag{13}$$

But can be rewritten as a summation like so:

$$= \sum_{i=1}^{3} w_i D(x_t, y_i) \tag{14}$$

Song et al. [18] finds that diffusion models learn approximate the score function. For conditional diffusion we therefore have the approximation $D_\theta(x_t, y_i) \approx \nabla_{x_t} \log p_t(x_t|y_i)$, which when plugged into the sum gives us:



(a) Mesh used for faces. (b) Mesh used for artworks. (c) Mesh used for scenes.

Figure 17: Triangle meshes used for different domains. Nodes and their locations are derived from datasets for the specific domain, in this case, AffectNet, EmoArt, and FindingEmo.

$$\approx \sum_{i=1}^{3} w_i \nabla_{x_t} \log p_t(x_t|y_i) \tag{15}$$

Through Bayes rule (the denominator in Bayes rule can be ignored because we are taking the gradient with respect to x_t), we can rewrite this as:

$$= \sum_{i=1}^{3} w_i (\nabla_{x_t} \log p_t(y_i|x_t) + \nabla_{x_t} \log p_t(x_t)) \tag{16}$$

However, since all weights sum to one the second score function can be taken outside of the summation to get the final expression:

$$= \nabla_{x_t} \log p_t(x_t) + \sum_{i=1}^{3} w_i \nabla_{x_t} \log p_t(y_i \mid x_t) \tag{17}$$

For Equation 5 a similar derivation can be done using the same logic. Specifically, using the score function approximation from Song et al. [18] we get:

$$D^{gen}_{bary}(x_t) = (w_{cfg} - s) D_\theta(x_t, y_{base}) + (1 - w_{cfg}) D_\theta(x_t) + s \left(\sum_{i=1}^{3} w_i D_\theta(x_t, y_i) \right) \tag{18}$$

$$\approx (w_{cfg} - s) \nabla_{x_t} \log p_t(x_t|y_{base}) + (1 - w_{cfg}) \nabla_{x_t} \log p_t(x_t) + s \left(\sum_{i=1}^{3} w_i \nabla_{x_t} \log p_t(x_t|y_i) \right) \tag{19}$$

We can apply Bayes rule like with the prior derivation (i.e. $\nabla_{x_t} \log p_t(x_t|y) = \nabla_{x_t} \log p_t(y|x_t) + \nabla_{x_t} \log p_t(x_t)$) then group terms we get the final score function based perspective:

$$\nabla_{x_t} \log p_t(x_t) + (w_{cfg} - s) \nabla_{x_t} \log p_t(y_{base} \mid x_t) + s \sum_{i=1}^{3} w_i \nabla_{x_t} \log p_t(y_i \mid x_t) \tag{20}$$

A.9 Experimental Details

We use a guidance scale of $w_{cfg} = 4.0$ across all modalities, and for FLUX.2-dev we use the default value for distilled guidance of 4.0.

Table 10: The VA accuracy of different emotion words when performing facial expression editing on AffectNet with FLUX.2-dev. Overall, 100 different emotion words were considered. Neither Barycentric Guidance face morphing were used here, just our standard prompt template, where only the emotion word is altered.

	RMSE
Happy	0.054
Sad	0.149
Angry	0.306
Bitter	0.451
Droopy	0.825
At Ease	0.908
Basic Emotions Mean	0.216
Other Emotions Mean	0.501



(a) Original Mesh. (b) Additional Lower Right Anchor.

Figure 18: 10 by 10 grids of our approach (backbone FLUX.2-dev) being used to edit an image from AffectNet. Our standard mesh used for all other facial expression editing results is shown in (a), whereas (b) shows editing results on the same mesh except with an additional anchor, "relaxed", in the bottom right corner.

Mesh Construction In early experiments, we explored using a denser mesh with over 100 emotion words. However, we found that ambiguous anchors (e.g., "bitter") and rare words (e.g., "droopy") led to severe misalignments between the intended target and the generated expression. If we compare the predictive error (RMSE when performing facial expression editing on AffectNet, testing 100 edits per emotion) of basic emotions against rarer, denser mesh words, the drop in accuracy (shown in Table 10) is stark.

The sensitivity to the mesh therefore primarily depends on the text encoder used and how well each discrete anchor will be rendered. This informed our choice of a compact mesh using common, visually grounded, and semantically unambiguous emotion anchors. A denser mesh is not automatically better; it helps only when the added anchors are reliably expressible by the editor.

We also examined VA placement. Adding a new anchor "relaxed" in the lower-right region of the VA space. As shown in Figure 18, this improves the gradual low-arousal transitions, producing more natural sleepy/relaxed expressions, such as progressive eye closure, compared with the current sparser lower-arousal design. This suggest that results may be improved by further refinement of the mesh, which we leave as a promising direction for future work.

Triangle meshes used for editing and generation throughout the paper are shown Figure 17.

Compute Costs We run our core experiments on NVIDIA H200 GPUs, which takes on average 0.64 seconds per inference step when using FLUX.2-dev. For normal generation, the number of function evaluations (NFEs) corresponds to the number of steps (i.e. 30 steps = 30 NFEs). For our full method, in the experiments we have at the moment, we use 30 steps to generate each anchor, then we restart the diffusion process at step 10 out of 30. Barycentric guidance and CFG for the restarted process uses 4 function evaluations per step. In total therefore, our method uses: $30 \times 3 + 20 \times 4 = 180$ NFEs. If no morphing is used we would have $30 \times 4 = 120$ NFEs. Therefore, generation with prompt interpolation takes roughly 19.2 seconds per image, whereas our full approach takes around 76.8 seconds without morphing and 115.2 seconds with morphing.

Facial Expression Editing For GANmut, we trained on AffectNet using the default configuration provided in the source repo. We train on uncropped rather than cropped faces. This allows us to compare background preservation between methods. GANmut utilizes polar instead of cartesian coordinates; for diversity metrics, we therefore sampled ten evenly spaced angles across ten different radii. For EmoStyle we use the default configuration in the original repo and run both phase 1 and phase 2 of the generation process. The prompt template used for our method is "Change this persons facial expression to be <emotion> while maintaining the same facial features, composition, lighting, and camera angle". We use 30 inference steps for our approach for both the anchors (discrete emotions) and the main image generation. The LPIPS metric used to measure the non-identity, non-affect content preservation is calculated by first using MediaPipe [65] to predict face landmarks. These landmarks are then used to create a binary mask, blocking out the face region only. This is done both the source and edited image. Finally, we calculate LPIPS between the face masked source image and face masked edited image.

Continuous Editing Comparison The classifier used for measuring smoothness for our 1D editing comparison was trained on AffectNet. Specifically, we use a VGG model [66] finetuned on AffectNet data for prediction of the 7 basic emotions, undersampling the overrepresented emotions in order to balance the dataset. All approach use the default parameters provided in the official implementations. 14 inference steps are used for all approaches.

EmoArt For the EmoArt dataset we finetune a DINOv3 ConvNeXt-Tiny [67] model to predict the emotion and the art style of the image, ultimately achieving 70.8% accuracy. The classifier predicts dicrete emotion, however, the EmoArt dataset is designed to represent emotion spread evenly over the VA model. We therefore recovered the VA value via a probability-weighted sum over the EmoArt VA points for each emotion. We use the prompt template "Change this artwork to evoke <emotion>, maintaining the art style as <style>, keep the content of the art the same", where "<style>" refers to the art style of the original image. Our approach uses 14 inference steps.

FindingEmo For FindingEmo, we follow [8] and train a linear probe on top of a frozen CLIP model (vit_base_patch32_clip_224) to predict the valence and arousal of an image. FindingEmo contains valence values in the range of $[-3.0, 3.0]$ and arousal values in the range of $[0.0, 6.0]$. We map both dimensions to a range of $[-1.0, 1.0]$ to allow for better comparison with other modalities. For this dataset we use the prompt template "Change the people and environment to feel more <emotion>, it should remain as a realistic photograph". Base prompts are taken from EmotiCrafter's official code base [8]. Our approach uses 14 inference steps.

B Text Encoders

In this section, we further explore the results of performing PCA on different text encoder's embeddings. In the main body, we provide an example of performing PCA on CLIP embeddings. We used the CLIP ViT-L/14 model with the prompt template "a portrait photo of a <emotion> face", where "<emotion>" is replaced with an emotion word. This process recovers a good approximation of the VA model on the principal components. Sometimes principal components need to be rotated or mirrored, but importantly, the ordering of points reliably follows the ordering of points in the VA model. To show that this is not dependent on the prompt template, we replicate this result in Figure 19 using three different prompts. All prompt templates considered recover a good approximation of the VA model on their principal components.

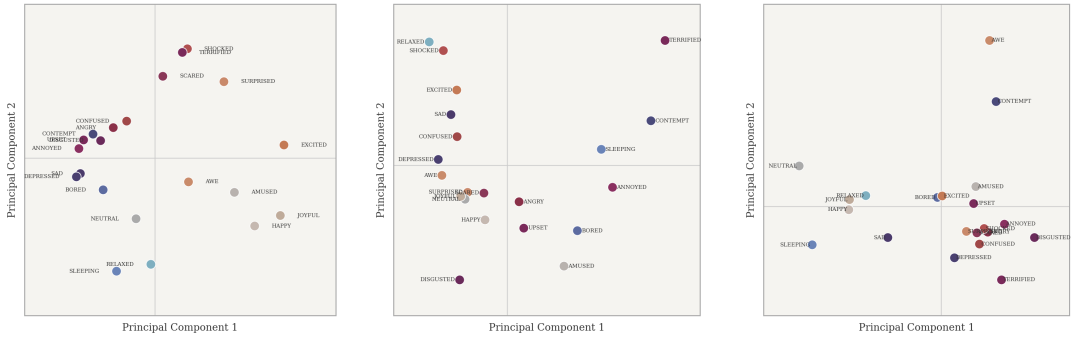


(a) **Prompt Template:** "a picture of a <emotion> man".

(b) **Prompt Template:** "an image of a <emotion> woman".

(c) **Prompt Template:** "<emotion>".

Figure 19: PCA performed on different prompt templates on the CLIP ViT-L/14 model.



(a) OpenCLIP ViT-bigG/14.

(b) T5 v1.1 XXL.

(c) Qwen2.5-VL.

Figure 20: PCA performed text embeddings from different text encoders. Prompt template used was "a portrait photo of a <emotion> face".

In Figure 20, we repeat this experiment on different text encoders with the prompt template "a portrait photo of a <emotion> face". For OpenCLIP ViT-bigG/14, as shown in Figure 20a, we once again recover the VA model. However, for T5 v1.1 XXL and Qwen2_5_VLForConditionalGeneration (shown in Figures 20b & 20c) the process largely fails, with opposite emotions often appearing side by side.

While tradition text-to-image models like Stable Diffusion uses CLIP encoders to condition on text, more modern image editing models do not. For example, FLUX.1-Kontext uses T5 as its primary text encoder. FLUX.2 models use QwenVL and MistralVL text encoders, neither of which provide VA-like subspaces. Finally, Qwen Image Edit also used QwenVL as a text encoder, thus motivating the guidance based approach which we take in this paper.

C Additional User Study Details

Our user study was conducted uses the Qualtrics platform which allows for anonymised survey's to be conducted. Consent and participant information is provided at the start of the online survey. After this, a user must tick a consent tickbox to continue to the survey. If they tick the "I do not consent" tickbox, they will be taken to the end of the survey. Following this there is a short tutorial explaining the survey and the valence-arousal model to users. It includes the following text alongside accompanying visuals:

1. "Emotions can be described by two dimensions: Valence (x-axis, unpleasant → pleasant) and Arousal (y-axis, calm → activated). Every emotion maps to a unique point within the unit circle. Emotions are also continuous under the Valence–Arousal Model, for example in between the astonished and happy, we have excited."

2. “Each point on the VA circle corresponds to a specific emotion. Valence (x-axis) separates positive emotions on the right from negative ones on the left. Here the happy face sits exactly on the Happy point, and the sad face exactly on Sad.”
3. “Arousal (y-axis) runs from calm at the bottom to activated at the top, the sleepy face sits exactly on Sleepy. Emotions are also continuous: the fearful expression sits between Astonished and Angry, illustrating that every point in the circle is a distinct emotional blend.”
4. “Intensity corresponds to distance from the centre. All three faces point toward disgust, but move progressively further out; weak, medium, then strong.”
5. “ Each question shows a source face (blue border) alongside a target emotion marked as a orange cross on the valence-arousal diagram. The blue dot shows the source face’s emotion. Below the divider are two edits (orange border), each produced by a different image-editing method from the same source. You will be asked which edits looks better to you. This is a subjective measure, so if you are not sure, you can also select that they are equal.”

Following the tutorial we then show user 20 sets of images, randomly selected from a total of 50 sets, and shown in a random order. From this we ask four questions:

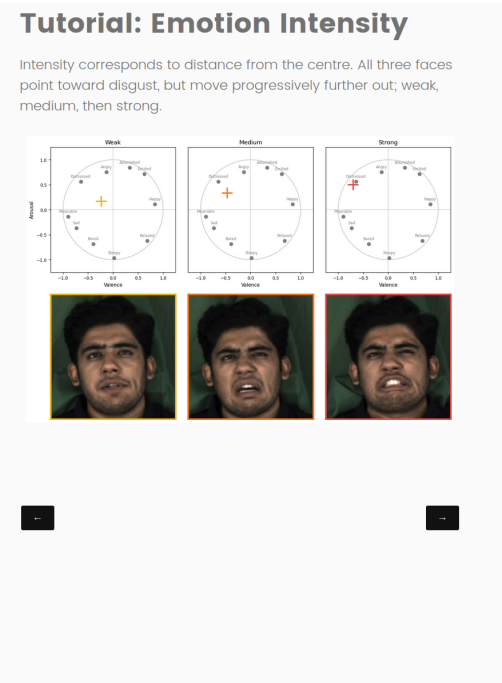
1. “Which image better represents the target emotion?”
2. “Which image looks more realistic? Focus on photo-realism only; ignore the emotion.”
3. “Which image better preserves the identity of the original image?”
4. “Which image better preserves other visual details (excluding identity and emotion), such as background, pose, clothing/accessories, and colour/style?”

For all questions, users must select one of the following options: “Edit A”, “Edit B”, or “They Are Equal”. Screenshots from the user study are shown in Figure 21. We also provide additional images from four more questions in Figure 22. Finally, we analyse the magnitude of VA changes between the source and edited images used in the survey in Figure 23. VA is predicted using EmotiEffLib [59]. From this we can see that EmoStyle is much more limited in the strength of VA edits it can make, and that the size of VA edits made by our model are more closely aligned with the size we are aiming for with our target VA values.

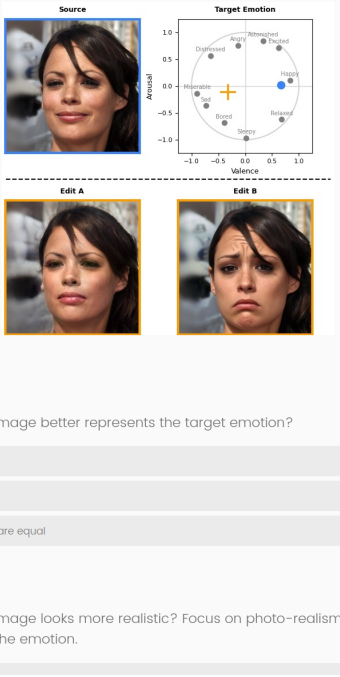
We obtain ethics approval from our Institutional Review Board (IRB) to conduct this user study. The consent form contained this section regarding possible risks for the participants: “No risks are envisaged for this study. Some paired image may only contain subtle differences so please take care to avoid eye stain when considering the images.”.

D Broader Impact

This work introduces a training-free framework for continuous emotional image editing, enabling nuanced creative expression and psychological research across diverse visual domains. While these advancements improve affective controllability, they could be misused to generate deceptive content or deepfakes by manipulating the emotional states of subjects in images. Finally, because our method relies on the internal representations of large foundation models, users should be aware that the range and intensity of generated emotions are ultimately constrained by the patterns present in the models’ original training data.

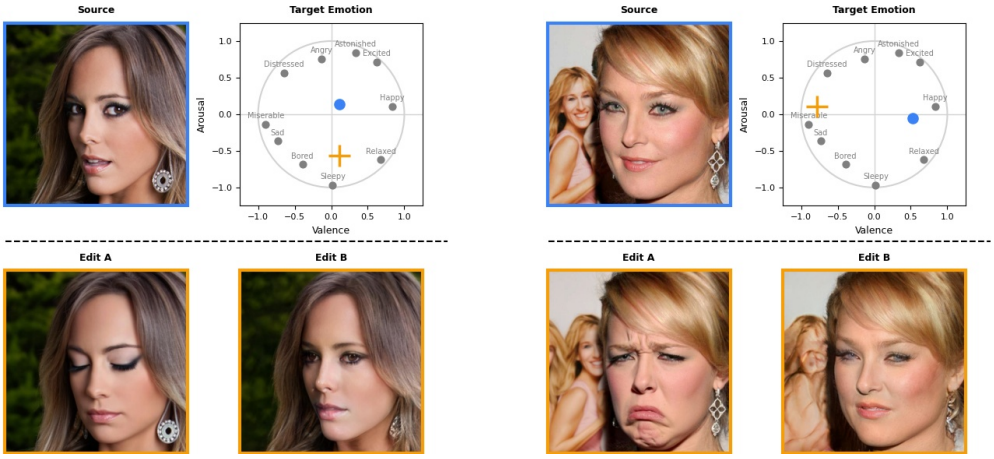


(a) Screenshot from our a user studies tutorial.



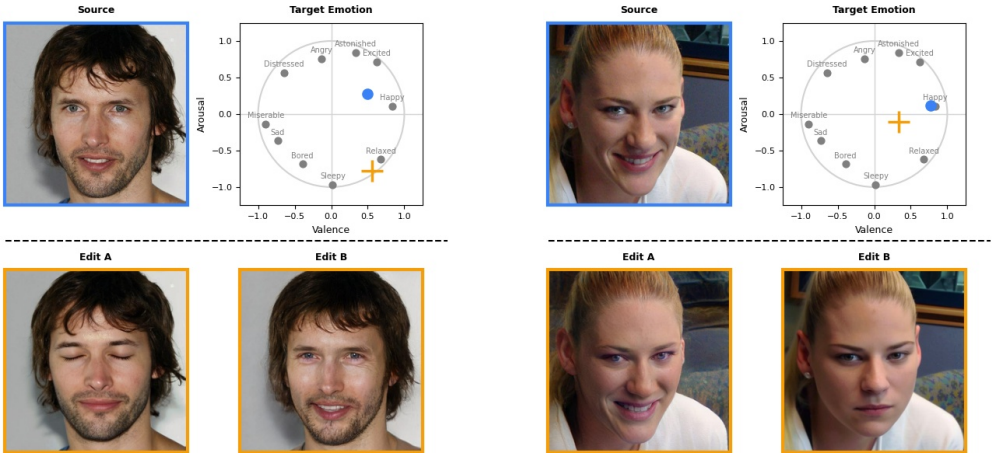
(b) Screenshot of a question from our a user study.

Figure 21: User study screenshots.



(a) Example 1

(b) Example 2



(c) Example 3

(d) Example 4

Figure 22: Examples of image comparisons from our user study. Our edits are Edit A, Edit A, Edit A, and Edit B in (a), (b), (c), and (d), respectively.

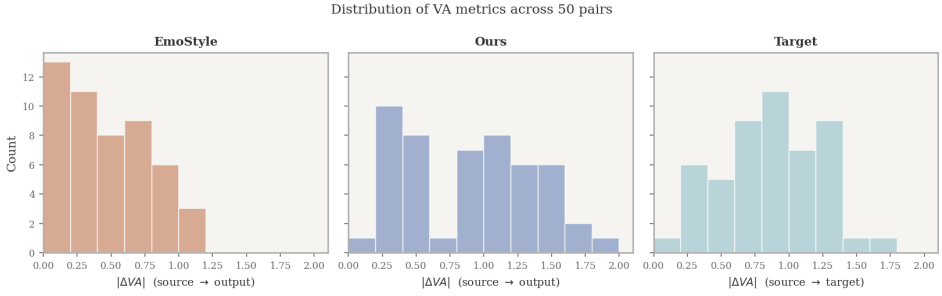


Figure 23: Magnitudes of the valence-arousal (VA) edits shown in our user study, as measured by EmotiEffLib, for EmoStyle (left) and ours (middle). The plot of the right shows the magnitude of the VA edit, as measured by the distance between the source images VA values (measured by EmotiEffLib) and the target VA value that both editing models are aiming at.

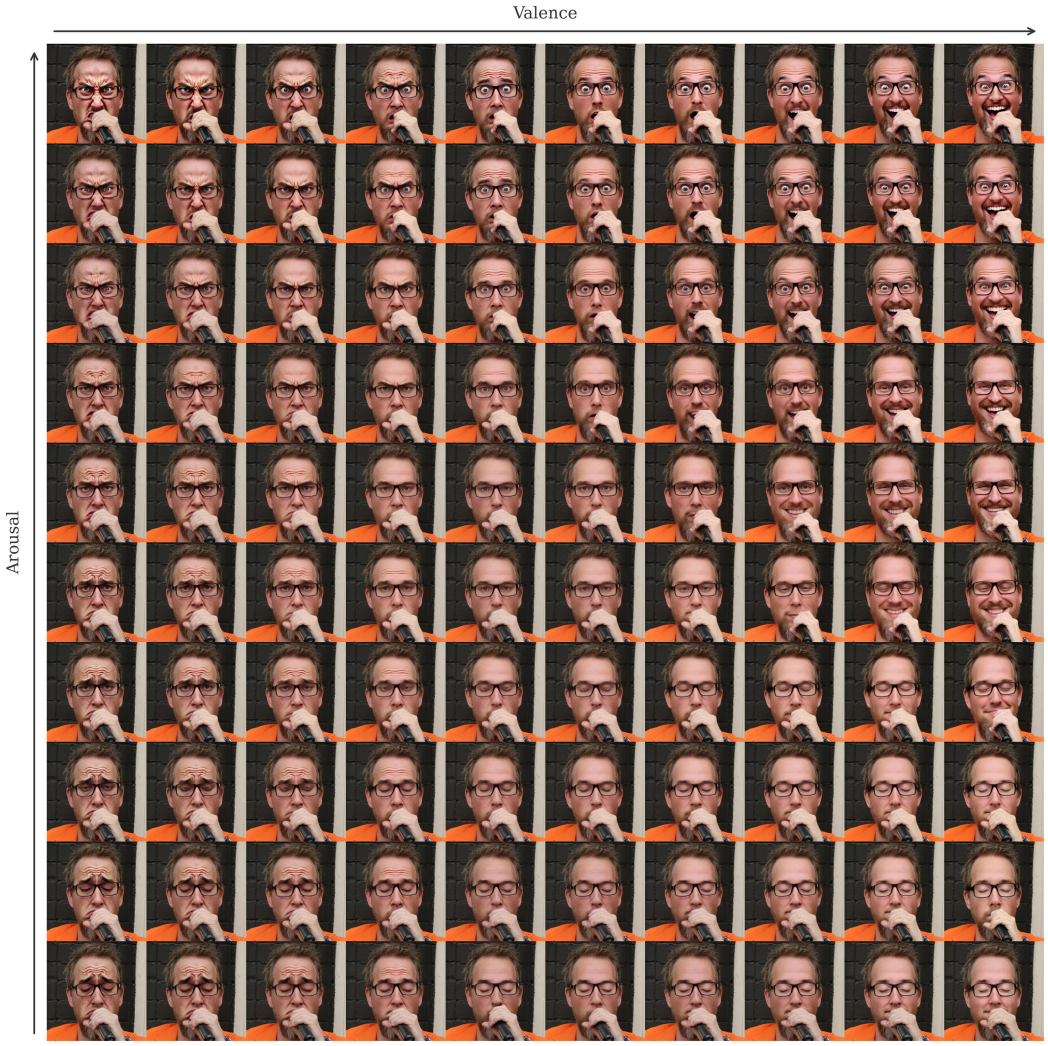


Figure 24: Our approach on FFHQ. VA range [-1,1] with uniform step size for rows and columns.

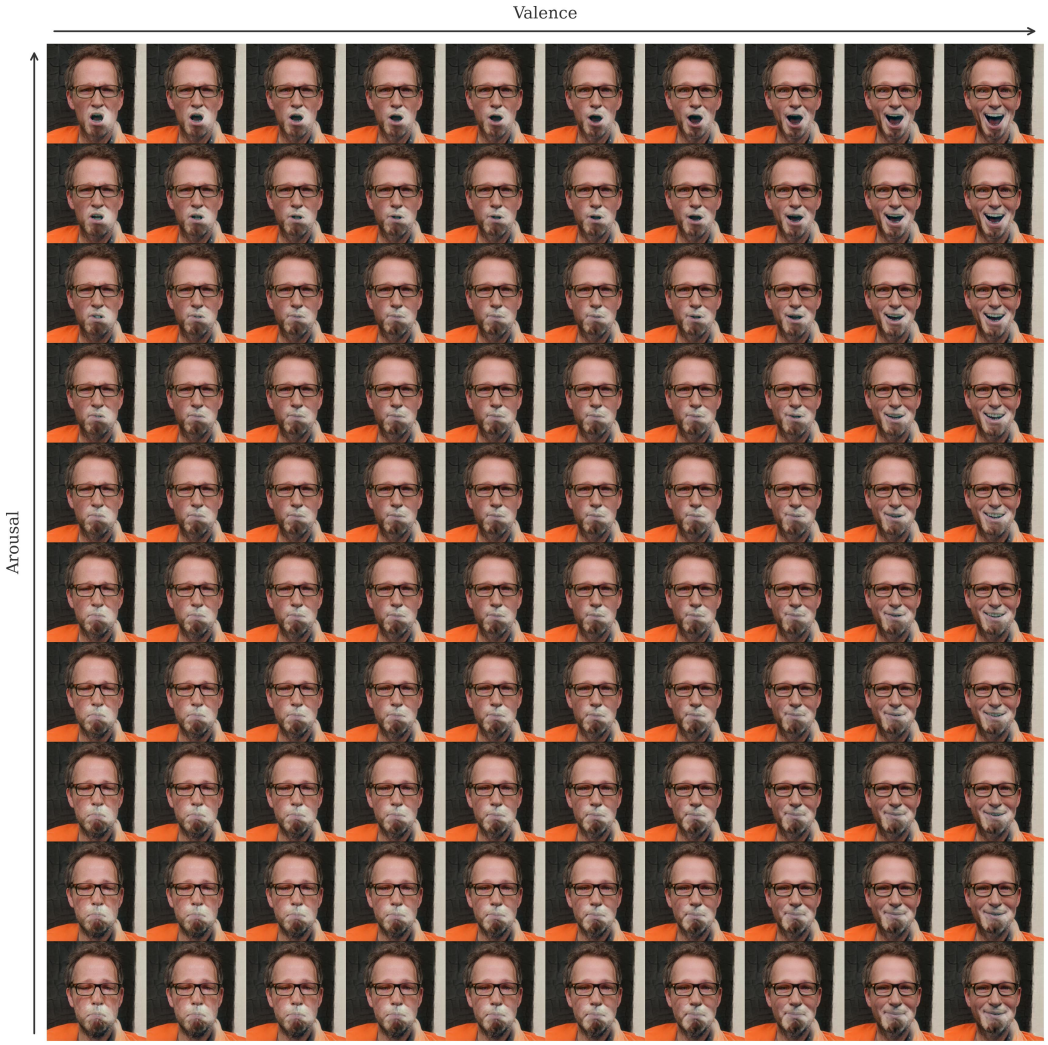


Figure 25: EmoStyle on FFHQ. VA range [-1,1] with uniform step size for rows and columns.



Figure 26: Our approach on CelebA. VA range [-1,1] with uniform step size for rows and columns.

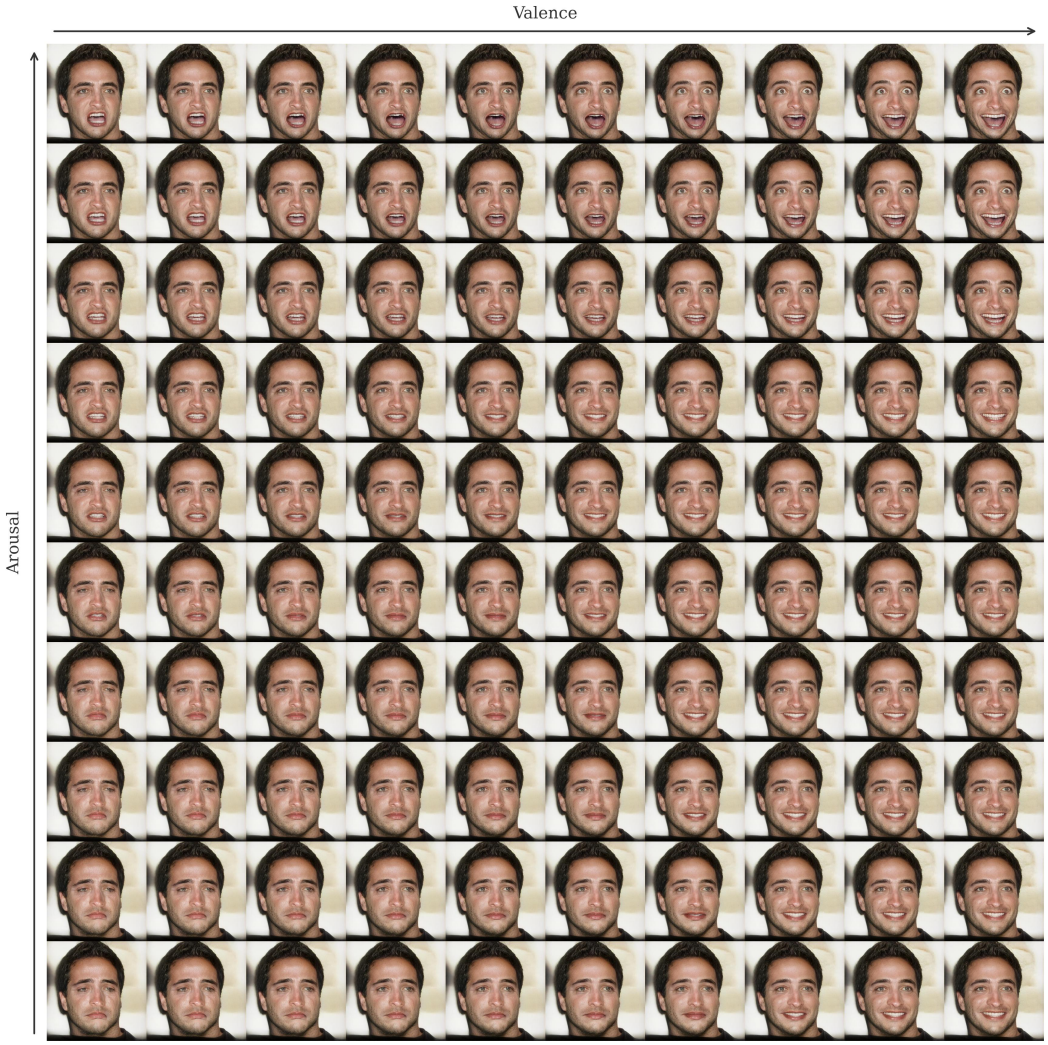


Figure 27: EmoStyle on CelebA. VA range [-1,1] with uniform step size for rows and columns.



Figure 28: Our approach on AffectNet. VA range [-1,1] with uniform step size for rows and columns.

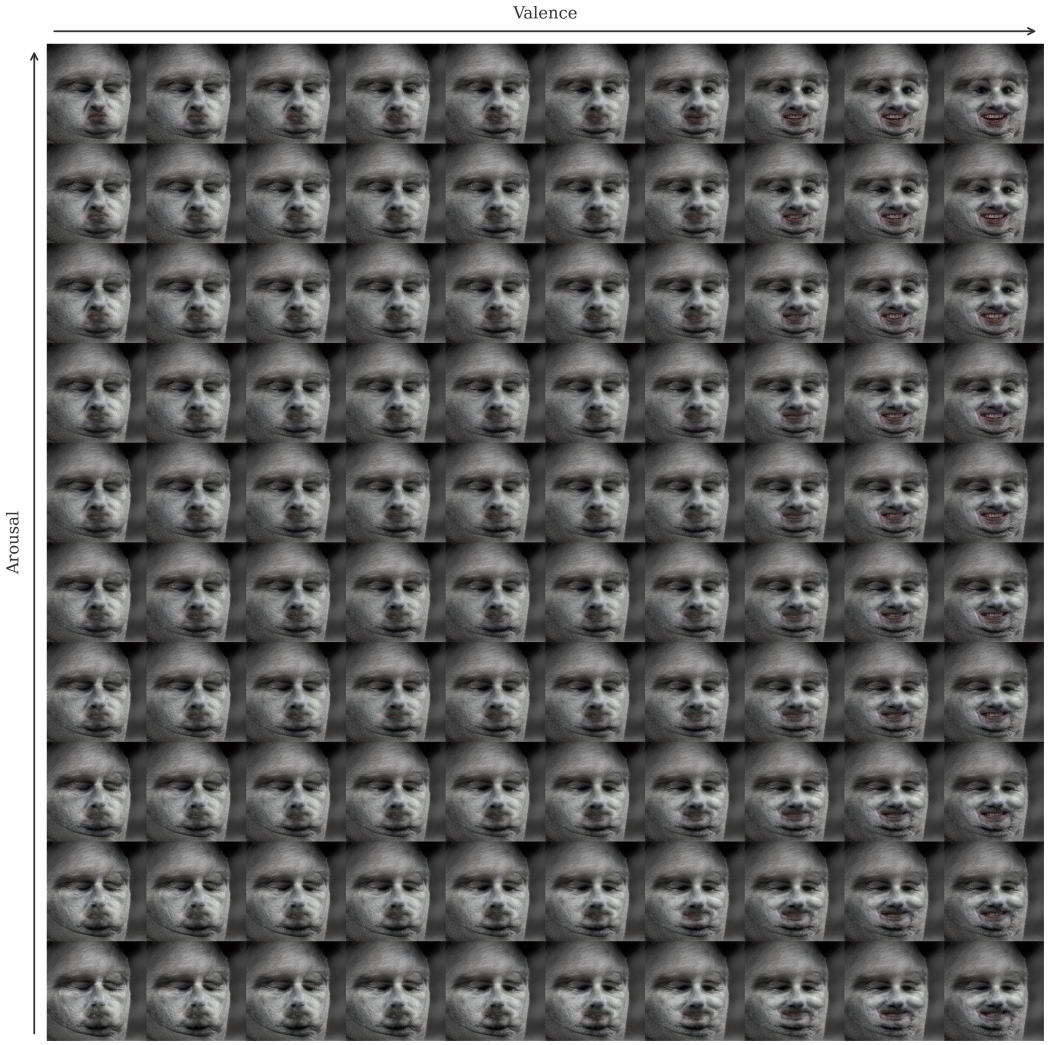


Figure 29: EmoStyle on AffectNet. VA range [-1,1] with uniform step size for rows and columns.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The VA controllability and performance claims are evidenced in Table 2. The claim of working across domains, datasets, models can be shown across Sections 4.1 and 4.2.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: A discussion of limitations can be found in Section 5.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: This paper does not contain any proofs, theorems or lemmas.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Detail for reproducibility are provided throughout the experiments in Section 4. Any details needed for reproducability that do not fit in this section can be found in the supplimentry materials in Section A.9

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Code will be released on GitHub upon acceptance.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: No optimizers used throughout the paper. Hyperparameters and data splits are found in Section A.9.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Error bars and significance levels are shown in cases where they are computationally feasible to calculate, for example in Figure 5 and our user study, but not for metrics like FID for which performing multiple runs would be computationally infeasible.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Relevant information provided in Section A.9.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Our research adheres to the NeurIPS Code of Ethics by using publicly available, cited datasets and transparently discussing the potential for dual-use or negative societal impacts in the appendix.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: A statement of broader societal impact is provided in Section D.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: No datasets or models will be release as a part of this research project.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Dataset and foundation models used throughout the paper have been properly credited. Models and datasets with license agreements have been adhered to.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: No assets are released as part of the paper in its current form. Once accepted, code will be released which will contain documentation in the form of markdown files containing instruction for setup and running of the code.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: Our user study gathers anonymised responses from human subjects. We include text instructions and screenshots in Section C. The user study was conducted on a voluntary basis, and so no compensation was provided.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [Yes]

    Justification: Details on the ethics review process and participant risk are provided in Section C.

    Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs were not used for the development of our method.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.