

ACE: Affective Continuous Editing via Object–Atmospheric Control for Fine-Grained Image Editing

Hao Li Kai Shu Gang Yang* Jingyang Jia Xun Chen Aiping Liu*

School of Information Science and Technology

University of Science and Technology of China

Hefei, Anhui, China

{haoli2101, sk18872125513, jjygood}@mail.ustc.edu.cn

{yg1997, xunchen, aiping1}@ustc.edu.cn

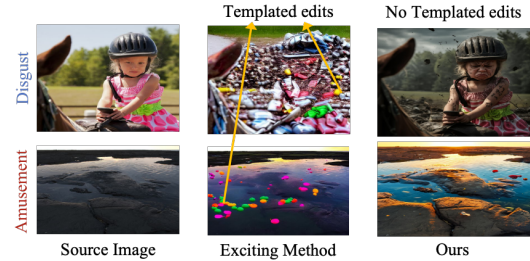
Abstract

Affective image editing aims to transform an image toward a target emotional state. Existing emotion-based editing methods typically represent emotion as a global condition, which often leads models to produce templated edits, limiting their ability to control fine-grained affective variations. To address these limitations, we propose Affective Continuous Editing Framework (ACE), an object–atmospheric framework for modeling visual emotion. To reduce templated editing, our framework employs an Affective Disentanglement Agent (ADA) to translate abstract affective intent into region-level editing instructions and counterfactual supervision signals. Guided by this data, we design a decoupled regional LoRA module with two independent lightweight branches, Object LoRA (OB-LoRA) and Atmosphere LoRA (AT-LoRA), which respectively learn object and atmospheric editing directions. In addition, we introduce a counterfactual regional suppression loss, encouraging the model to learn disentangled regional features and continuous, fine-grained affective variations. Experiments demonstrate that ACE achieves better identity preservation, region-level controllability, and fine-grained emotional expression than existing emotion-conditioned editing methods. Quantitative analyses further confirm that the two learned control directions are monotonic and largely confined to their intended regions. The dataset and code are available at <https://github.com/xixixixixxxx/ACE>.

1 Introduction

Affective image editing aims to transform a source image toward a user-specified emotional state while preserving its structure, objects, and subject identity. Unlike conventional object editing, emotion does not have explicit visual boundaries. Instead, it emerges from the joint effects of subjects, actions,

* Corresponding authors.



(a) Existing methods tend to generate templated edits that are strongly associated with the target emotion but weakly grounded in the original image content.



(b) Continuous editing from sadness to contentment with ACE.

Figure 1: Motivation of ACE. (a) Existing methods tend to generate target-emotion-specific but image-agnostic templated edits. (b) ACE enables fine-grained target-emotion editing through object–atmospheric control while preserving source identity and layout.

relationships, illumination, color tone, weather, and ambience (Machajdik and Hanbury, 2010; Yuan et al., 2013; Zhao et al., 2018; Fan et al., 2018; Yang et al., 2021).

Recent emotion-oriented editing methods have made notable progress (Weng et al., 2023; Lin et al., 2025; Yang et al., 2025; Mao et al., 2026; Zhang and Fan, 2026); however, most methods inject the target emotion as a global prompt or implicit condition. This formulation lacks an explicit bridge between abstract affective intent and image-specific

visual editing operations. As a result, models may associate emotions with a limited set of frequent objects or scene templates, thereby producing templated edits. As shown in Figure 1a, existing methods tend to generate target-emotion-specific but image-agnostic templated edits, indicating that the learned affective transformation is often weakly grounded in the source image content.

Moreover, such global emotion modeling (Lin et al., 2025; Yang et al., 2025) can also collapse fine-grained emotional expression: existing methods distinguish coarse affective directions such as positive versus negative reasonably well, but when the target is specified into fine-grained categories such as amusement, contentment, awe, and excitement, their results exhibit substantial confusion.

Studies in visual cognition suggest that affective appraisal in everyday scenes is primarily driven by object representations and their object relationships, whereas scene-level cues provide complementary evidence when object information is insufficient (Gao et al., 2025). This finding indicates that image emotion is shaped by both object and atmospheric factors. Motivated by this observation, we propose ACE, an explicitly object-atmospheric framework for continuous affective image editing. As illustrated in Figure 1b, ACE enables continuous affective editing from sadness to contentment by jointly controlling object and atmospheric factors, supporting fine-grained target-emotion editing while preserving the source identity and layout.

To translate abstract emotions into image-dependent executable visual edits and mitigate template-like editing, we construct an *Affective Disentanglement Agent* (ADA) driven by large-scale foundation models. The agent analyzes the input image with visual foundation models to identify low-level visual cues and high-level semantic information. In this way, abstract affective intent is converted into a region-aware, attribute-level editing plan and further decomposed into object-level and atmospheric changes (Wu et al., 2023; Cui et al., 2023; Mao et al., 2026).

We further introduce an *Object-Atmospheric Affective Continuous Editing model* (OACE). It formulates emotional transformation as the coordinated interaction between object-level and atmospheric factors, controlled along two continuous dimensions. To achieve disentangled and continuous control, we design a counterfactual regional suppression loss, which encourages affective expression within the target region while suppressing

irrelevant emotional changes in non-target regions. After training, the lightweight adaptation modules form two smoothly scalable control directions, enabling interpretable modulation of object-level and atmospheric factors.

Our contributions are summarized as follows:

- We propose an object atmospheric continuous paradigm for affective image editing, moving beyond global emotion-label conditioning toward controllable modulation of affect-bearing object and atmospheric factors.
- We develop an Affective Disentanglement Agent that translates abstract affective intent into region-aware, attribute-level editing instructions and counterfactual supervision, bridging high-level affective intent with executable visual operations.
- We introduce OACE, a lightweight continuous affective editing model that learns scalable object and atmospheric control directions with a counterfactual regional suppression loss, enabling smooth, localized, and interpretable emotion editing.

2 Related Work

2.1 Image Editing

Diffusion-based image editing has advanced through pretrained generative models (Ho et al., 2020; Dhariwal and Nichol, 2021; Rombach et al., 2022), attention or feature manipulation (Hertz et al., 2023; Tumanyan et al., 2023; Mokady et al., 2023), and subject-preserving representations such as BLIP-Diffusion (Li et al., 2023). Instruction-based and conversational methods, including InstructPix2Pix (Brooks et al., 2023), Visual ChatGPT (Wu et al., 2023), and ChatEdit (Cui et al., 2023), align natural-language commands with visual modifications (Sun et al., 2025; Li et al., 2025), and recent foundation models such as Qwen-Image and Qwen-Image-Edit (Wu et al., 2025) support high-resolution generation and complex object or appearance editing. Region-aware and controllable methods such as Blended Diffusion (Avrahami et al., 2022), Prompt-to-Prompt, ControlNet (Zhang et al., 2023b), and drag-based editing (Fu et al., 2025) localize edits with user-provided masks, attention maps, point handles, or control signals, and LoRA adapters such as

Concept Sliders (Gandikota et al., 2024) and SliderEdit (Zarei et al., 2025) expose continuous control over a single attribute. These methods target semantic edits with explicit boundaries, require the user to specify what and where to edit, and rarely model affective factors. In contrast, ACE takes only a source image and a target emotion, derives image-specific object and atmospheric edits automatically, and learns two composable LoRA directions confined to disjoint regions.

2.2 Image Emotion Editing

Early image emotion editing relied on emotional color transfer and style manipulation with predefined affective color statistics (Liu et al., 2018, 2022). Later methods incorporated human-centric cues such as facial expressions and face geometry (Li et al., 2024b), but still focused largely on global appearance shifts rather than localized affective reasoning (Ibarrola et al., 2024; Yang et al., 2024a). Affective Image Filter (AIF) maps textual emotion descriptions to visual styles while preserving object content (Weng et al., 2023); EmoEdit constructs a large-scale emotion editing dataset and trains an Emotion Adapter for content-level modifications (Yang et al., 2025); Make Me Happier learns an affective semantic mapping that shifts source image representations toward target emotions (Lin et al., 2025); and EmoAgent (Mao et al., 2026) and EmoKGEdit (Zhang and Fan, 2026) diversify affective edits with multi-agent planning or visual-cue knowledge. All of these methods condition on a global emotion label or description; none separates affect-bearing object factors from atmospheric factors or supervises them with region-level counterfactual targets.

2.3 LLM-As-Agent

LLM-based agents have demonstrated strong reasoning, planning, and tool-use capabilities (Wei et al., 2022; Yao et al., 2023; Schick et al., 2023; Wang et al., 2024a; Liu et al., 2024b). Multimodal agents extend them to visual tasks (Xie et al., 2024): Visual ChatGPT, MM-REACT, and HuggingGPT orchestrate vision foundation models through an LLM controller (Wu et al., 2023; Yang et al., 2023b; Shen et al., 2023); VisProg and ViperGPT compose visual programs (Gupta and Kembhavi, 2023; Surís et al., 2023); and ChatEdit and GenArtist apply agents to image editing (Cui et al., 2023; Wang et al., 2024b). Most visual agents focus on explicit semantic edits, whereas affective editing

requires reasoning over abstract emotions and their image-specific manifestations. Our ADA converts affective intent into region-aware instructions and counterfactual supervision; unlike EmoAgent (Mao et al., 2026), which invokes agents at inference time, it is used offline, so the resulting editor requires no agent or tool calls.

3 Method

3.1 Affective Disentanglement Agent

To obtain region-level counterfactual supervision for object–atmospheric affective control and mitigate the template-like editing, we introduce ADA, a multi-tool-augmented multimodal visual reasoning agent that translates a source image and a target emotion into context-aware, region-specific, and executable editing plans. ADA is used only offline for data construction; it is not invoked during OACE training or inference.

As shown in Figure 2, ADA operates in three stages. First, it performs structured visual perception by integrating Depth Anything V2 (Yang et al., 2024b), InSPyReNet (Kim et al., 2022), Grounding DINO (Liu et al., 2024a), and Semantic-SAM (Li et al., 2024a). These modules provide spatial relations, salient-subject localization, open-vocabulary detection, and region segmentation as structural cues and region-level evidence for the input image.

In the affective editing reasoning stage, ADA decomposes the editing intent into object adjustment and atmospheric modulation based on the target emotion and image context. Object adjustment refers to high-level changes of the affect-bearing subjects, such as facial expression, pose, action, object state, and inter-subject relationships. Atmospheric modulation refers to low- and mid-level visual changes, such as illumination, color tone, weather conditions, and overall ambience.

In the counterfactual editing generation stage, ADA generates object and atmospheric editing instructions separately and constructs three types of counterfactual targets: object-only edits, atmospheric-only edits, and joint object–atmospheric edits. Each constructed sample contains the source image x_{src} , the target emotion e , object and atmospheric masks M_s and M_a , the ADA-generated object and atmospheric instructions τ_s and τ_a , and counterfactual targets x_s , x_a , and x_f , which correspond to object-only, atmospheric-only, and full-image edits, respectively. The detailed instructions τ_s and τ_a are used

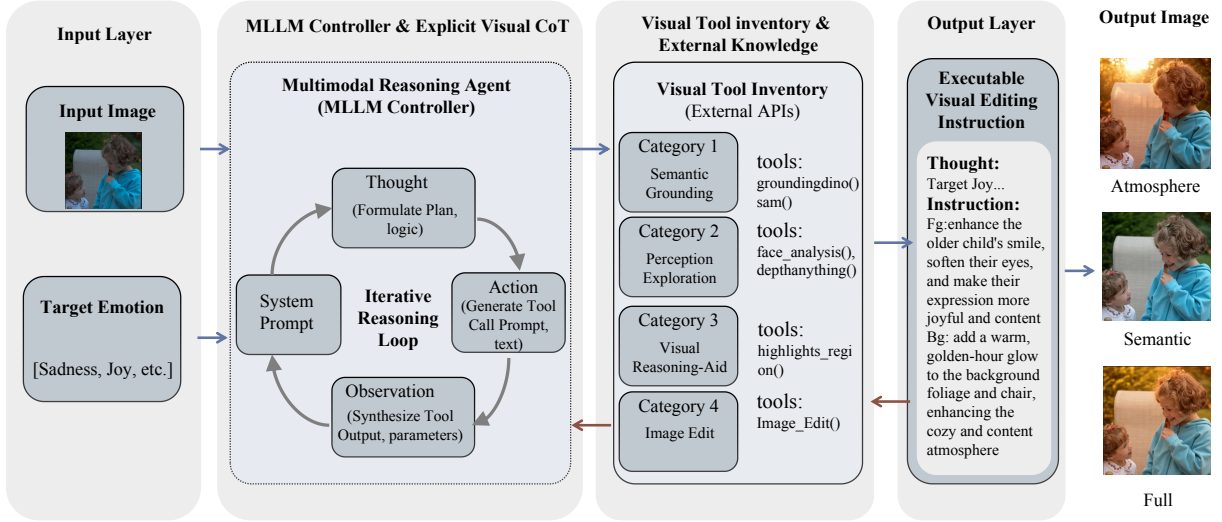


Figure 2: Overview of the ADA multimodal editing agent. Given an input image and a target emotion, the agent performs structured visual perception, iterative multimodal reasoning, and tool-augmented editing generation to produce executable region-level affective editing instructions.

only to produce the counterfactual targets; as described in Section 3.2, OACE itself is conditioned on concise emotion prompts, so no ADA output is required at inference.

Overall, ADA produces a counterfactual object-atmospheric dataset comprising 4,609 source images, 6,945 full-image edits, 13,419 object and atmospheric region-level edits, and 8,226 masks. On average, each source image is associated with 1.51 target-emotion editing directions, and each editing direction is paired with explicit object and atmospheric instructions. The source images come from existing datasets, and the generation, filtering, and editing models are all distinct, so no model consumes its own supervision. Image filtering, the agent prompt, tool configurations, quality control, and rejection statistics are detailed in Appendix A.

3.2 Object-Atmospheric Affective Continuous Editing Model

Using the counterfactual object-atmospheric dataset constructed by ADA, we build the Object-Atmospheric Affective Continuous Editing model (OACE) on a frozen Qwen-Image-Edit backbone with an MM-DiT architecture. Instead of modeling the target emotion as a global condition, OACE factorizes affective editing into two composable directions, object-level cues and atmospheric cues, as defined in Section 3.1.

Given a training sample, OACE is conditioned on concise region-level emotion prompts rather than one-hot emotion labels, global emotion em-

beddings, or the detailed ADA instructions. We consider three training states: object editing, atmospheric editing, and full affective editing. Their textual prompts are defined as $p_s = t_s$, $p_a = t_a$, and $p_f = t_s \oplus t_a$, respectively, where $\oplus$ denotes text concatenation. Specifically, t_s follows the foreground-oriented template “Edit the foreground of the image to convey {emotion},” while t_a follows the background-oriented template “Edit the background of the image to convey {emotion}.” The same templates are used at inference, so the model receives only the target emotion and no additional regional or attribute information. The prompts are encoded by the frozen text encoder $\phi(\cdot)$ to obtain c_s , c_a , and c_f .

During training, a state $k \in \{s, a, f\}$ is sampled, and a state router $\mathcal{R}$ selects the corresponding condition c_k , target image x_k , mask M_k when available, and LoRA strengths (λ_s, λ_a) . For object-only, atmosphere-only, and full editing states, the router sets (λ_s, λ_a) to $(1, 0)$, $(0, 1)$, and $(1, 1)$, respectively, thereby activating OB-LoRA, AT-LoRA, or both, so that each counterfactual target supervises only the branch (or composition) responsible for it.

Decoupled regional LoRA. Standard LoRA (Hu et al., 2022) adapts pretrained weights by introducing low-rank incremental updates. For the l -th linear layer in MM-DiT, the update is defined as

$$W'_l = W_l + \Delta W_l, \quad \Delta W_l = B_l A_l, \quad (1)$$

where W_l denotes the frozen pretrained weight, and A_l and B_l are trainable low-rank matrices.

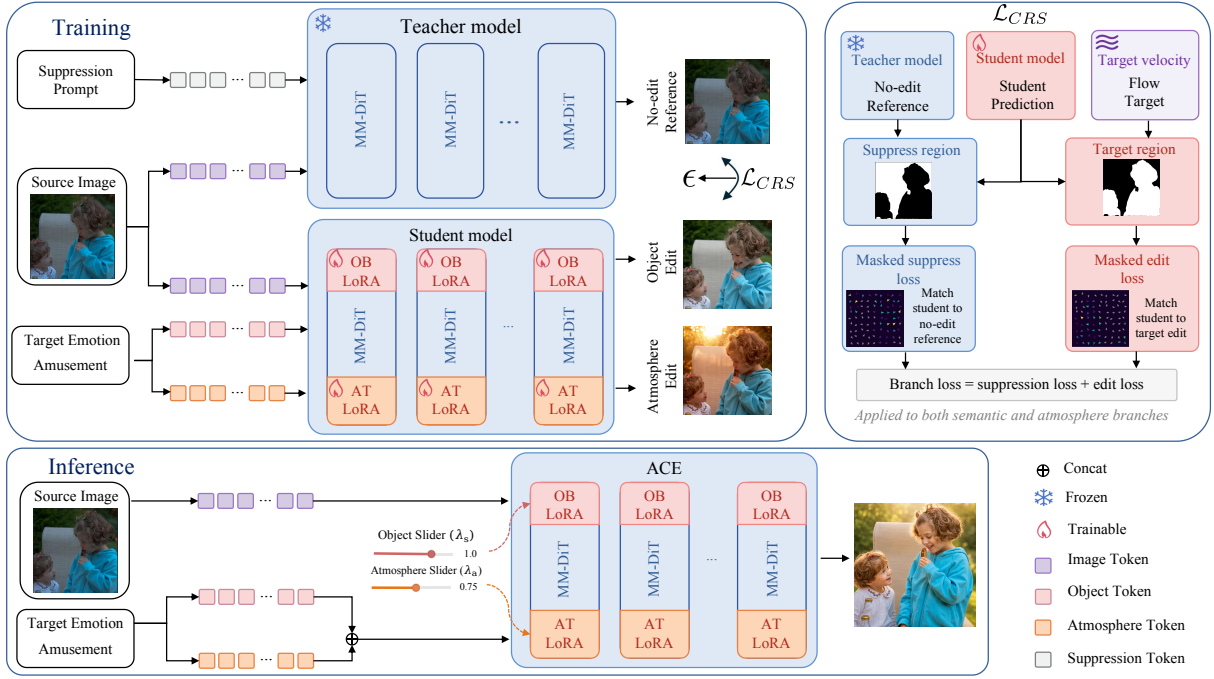


Figure 3: Overview of OACE. The upper part shows the training stage, where the teacher and student models are optimized with the designed training losses. The lower part shows the inference stage, where the learned modules are used for controllable emotion editing. The right part illustrates the loss design, including reconstruction, suppression, and edit-related supervision.

In affective image editing, a single LoRA module often entangles emotion cues from different sources. It may change color tone when only facial expression should be edited, or weaken identity when only the atmosphere should change. To address this, we decompose the affective parameter update into two independent directions:

$$W_l' = W_l + \lambda_s \underbrace{B_l^s A_l^s}_{\Delta W_l^s} + \lambda_a \underbrace{B_l^a A_l^a}_{\Delta W_l^a}. \quad (2)$$

Here, ΔW_l^s denotes the object editing direction learned by OB-LoRA, whereas ΔW_l^a denotes the atmospheric editing direction learned by AT-LoRA. The coefficients λ_s and λ_a control the strengths of the object and atmospheric sliders, respectively. Unlike single-adapter LoRA editing, the two branches are trained under the router with paired counterfactual targets and thus form composable control directions rather than parallel adapters.

Counterfactual regional suppression (CRS) loss. Although the counterfactual targets provide direct supervision for the two affective factors, naively fitting them over the entire image can still induce cross-region leakage. For example, the object branch may learn atmospheric color shifts from an object-only target, whereas the atmospheric branch

may modify object content from an atmospheric-only target. We therefore retain each LoRA branch as a positive editing direction and use regional suppression only as a leakage regularizer outside the intended edit region.

For a sampled training state $k \in \{s, a, f\}$, we obtain a noisy latent z_t^k from the corresponding counterfactual target x_k and denote its flow-matching velocity (Lipman et al., 2023) by v_k^* . The student prediction is computed as $v_{\text{stu}}^k = F_{\theta_0 + \lambda_s \Delta \theta_s + \lambda_a \Delta \theta_a}(z_t^k, c_k)$, where θ_0 is the frozen pretrained editing model and (λ_s, λ_a) is selected by the router, as described above. For the two regional states $k \in \{s, a\}$, the intended edit region is M_k , and the active branch is trained to match the counterfactual edit within that region:

$$\mathcal{L}_{\text{edit}}^k = \left\| D(M_k) \odot (v_{\text{stu}}^k - v_k^*) \right\|_2^2, \quad (3)$$

where $D(\cdot)$ downsamples an image-space mask to the latent or token resolution.

To suppress unintended changes outside the target region, we construct a no-edit teacher prediction $v_{\text{ref}}^k = F_{\theta_0}(z_t^k, c_{\text{neu}})$ with the frozen base model and a neutral preservation prompt c_{neu} . The preservation regularizer then matches the student

prediction to this neutral reference on the complement of the edit mask:

$$\mathcal{L}_{\text{sup}}^k = \|D(1-M_k) \odot (v_{\text{stu}}^k - \text{sg}(v_{\text{ref}}^k))\|_2^2, \quad (4)$$

where $\text{sg}(\cdot)$ denotes the stop-gradient operation. This term does not reverse the editing direction; it only discourages the active branch from producing affective changes in regions that should remain unchanged. Unlike pixel-space preservation or mask-reconstruction losses, it operates in the flow-matching prediction space and matches a frozen no-edit reference rather than the source pixels.

For the full editing state f , both LoRA branches are activated and the target is the joint counterfactual edit x_f . Since no region is held out, we supervise the full prediction directly:

$$\mathcal{L}_{\text{edit}}^f = \|v_{\text{stu}}^f - v_f^*\|_2^2. \quad (5)$$

The state-wise loss is $\mathcal{L}^k = \lambda_{\text{edit}}\mathcal{L}_{\text{edit}}^k + \lambda_{\text{sup}}\mathcal{L}_{\text{sup}}^k$ for $k \in \{s, a\}$ and $\mathcal{L}^f = \mathcal{L}_{\text{edit}}^f$, and the overall objective is $\mathcal{L} = \mathbb{E}_{k \sim p(k)}[\mathcal{L}^k]$, where $p(k)$ is the state-sampling distribution used by the router.

4 Experiments

We evaluate ACE from four aspects: whether the edits express the requested target emotion, whether they preserve the source content and visual quality, whether human raters prefer the edits, and whether the decoupled object and atmospheric controls are monotonic, disentangled, and interpretable.

4.1 Dataset

To evaluate affective image editing across diverse visual scenarios, including natural scenes, indoor and outdoor objects, and human activities, we construct a test set containing 346 source images. The images are sampled from EmoSet (Yang et al., 2023a), MagicBrush (Zhang et al., 2023a), and MA5K (Shi et al., 2021), and are manually filtered and deduplicated to balance emotional relevance, source diversity, and editability. Each image is targeted with 8 emotion directions, resulting in a total of 2,768 image pairs.

4.2 Evaluation Metrics

Affective image editing requires a balance between emotional transformation and source preservation. We therefore report emotion, preservation, and perceptual-quality metrics. For affective expression, Emo-Acc8 (Yang et al., 2024a) measures

whether an edited image is assigned to the requested emotion among the eight target categories. TEA (Zhang and Fan, 2026) measures the activation strength of the target emotion in the edited result, while Emo-S (Yang et al., 2024a) measures the gain of the target emotion response relative to the source image. For source preservation, DINO-I (Caron et al., 2021) measures high-level semantic similarity between the edited and source images, and LPIPS (Zhang et al., 2018) measures perceptual distance in deep feature space. We also report AesScore to assess overall visual quality. Finally, ID measures identity consistency on person-centric samples using identity features, including ArcFace (Deng et al., 2019) facial features when faces are available, so that subject drift is reflected in the evaluation. All methods are evaluated on the same 2,768 samples with identical evaluators; implementation details of the evaluators are given in Appendix B.3.

4.3 Baselines

We compare ACE with representative general image editing and affective editing methods. The baselines include two instruction-based image editing methods, InstructPix2Pix (Brooks et al., 2023) and Qwen-Image-Edit (Wu et al., 2025); two affective transfer methods, AIF (Weng et al., 2023) and CLVA (Fu et al., 2022); and two affective image editing methods, EmoEdit (Yang et al., 2025) and EmoEditor (Lin et al., 2025). Qwen-Image-Edit receives the same concise emotion prompt as ACE without adaptation and serves as a same-backbone, no-training reference; same-backbone LoRA alternatives are compared in Section 4.8.

4.4 Qualitative Analysis

Figure 4 presents the qualitative comparisons. ACE produces target-consistent affective edits while preserving the main content and spatial layout of the source images. General editing methods often make limited emotion-related changes, affective transfer methods rely mainly on global stylization and may weaken source objects, and existing affective editing methods introduce stronger changes that are sometimes less localized or less aligned with the target emotion. In contrast, ACE adjusts object and atmospheric cues according to the image context, leading to more expressive and better-preserved edits.



Figure 4: Qualitative comparison with general image editing, affective transfer, and affective image editing baselines. The examples cover amusement, excitement, fear, and sadness across object-centric, human-centric, scene-centric, and animal-centric inputs.

4.5 Quantitative Analysis

Table 1 presents the quantitative comparison. ACE achieves the best results on all affective metrics, indicating stronger target-emotion alignment than the compared methods. Compared with EmoEdit, the strongest affective baseline, ACE improves Emo-Acc8 from 0.300 to 0.564 and Emo-S from 1.914 to 6.327. ACE also improves DINO-I, ID, and AesScore over EmoEdit, showing better content consistency, identity preservation, and visual quality; EmoEditor achieves the highest ID but remains weaker on emotion metrics and DINO-I.

Expression–preservation trade-off. ACE does not obtain the best DINO-I or LPIPS. Both metrics compare the edit directly with the source and thus favor methods that change little: InstructPix2Pix and Qwen-Image-Edit obtain the lowest LPIPS (0.129 and 0.195) but also the lowest Emo-S (0.430 and 0.491), i.e., they largely leave the image unchanged. Conveying a new emotion requires visible changes in expression, pose, illumination, and color tone, so a moderate increase in perceptual distance is an expected consequence of successful affective editing rather than a preservation failure. We therefore treat DINO-I and LPIPS as safeguards against excessive drift: compared with EmoEdit, the only baseline with a substantial affective gain, ACE achieves a much larger Emo-S together with higher DINO-I, ID, and AesScore, and its LPIPS (0.553) stays below that of AIF (0.663) and EmoEd-

itor (0.623), whose changes are less targeted. The user study (Section 4.7) confirms that raters judge this operating point as the best balance.

4.6 Continuous Control and Disentanglement

Qualitative sweeps. For a fixed source image and target emotion, we vary the object strength λ_s and the atmospheric strength λ_a independently. As shown in Figure 5, increasing λ_s gradually strengthens affect-bearing object cues while keeping the atmospheric context largely stable, whereas increasing λ_a progressively modulates illumination, color tone, and ambience without replacing the object content. Figure 6 further shows that the two sliders can be assigned to different target emotions, enabling counterfactual affective compositions that are not available in a single-label editing formulation. In these examples, object changes and atmospheric changes evolve along different emotional directions while the subject identity and spatial layout remain largely preserved. Additional sweeps are provided in Appendix C.

Monotonicity and smoothness. To verify that the sliders provide gradual rather than abrupt control, we evaluate each branch on 352 source image–emotion sequences covering all eight emotions, sweeping its strength over five evenly spaced levels while the other branch is disabled. Table 2 reports Spearman correlations between slider strength and regional LPIPS change (visual monotonicity) or

Method	Main Test Set							EmoEdit Test Set						
	Emo-Acc8	TEA	Emo-S	DINO-I	LPIPS	AesScore	ID	Emo-Acc8	TEA	Emo-S	DINO-I	LPIPS	AesScore	ID
InstructPix2Pix	0.152	0.147	0.430	0.964	0.129	5.492	0.449	0.150	0.137	0.447	0.952	0.132	5.733	0.626
Qwen-Image-Edit	0.160	0.145	0.491	0.938	0.195	5.534	0.338	0.152	0.148	0.525	0.935	0.278	6.394	0.645
AIF	0.121	0.127	0.161	0.755	0.663	4.711	0.670	0.127	0.125	0.160	0.743	0.647	5.035	0.821
CLVA	0.136	0.125	0.093	0.702	0.531	4.964	0.712	0.136	0.137	0.144	0.720	0.583	5.309	0.786
EmoEdit	<u>0.300</u>	<u>0.345</u>	<u>1.914</u>	0.725	0.280	5.510	0.488	<u>0.436</u>	<u>0.355</u>	<u>3.025</u>	0.744	0.368	6.170	0.637
EmoEditor	0.152	0.244	0.513	0.542	0.623	5.106	0.858	0.145	0.190	0.417	0.656	0.550	5.542	0.869
ACE	0.564	0.353	6.327	0.755	0.553	5.583	<u>0.807</u>	0.531	0.369	5.722	0.740	0.433	<u>6.173</u>	0.898

Table 1: Main quantitative comparison on the main test set and the EmoEdit test set. Higher is better for Emo-Acc8, TEA, Emo-S, DINO-I, AesScore, and ID. Lower is better for LPIPS.

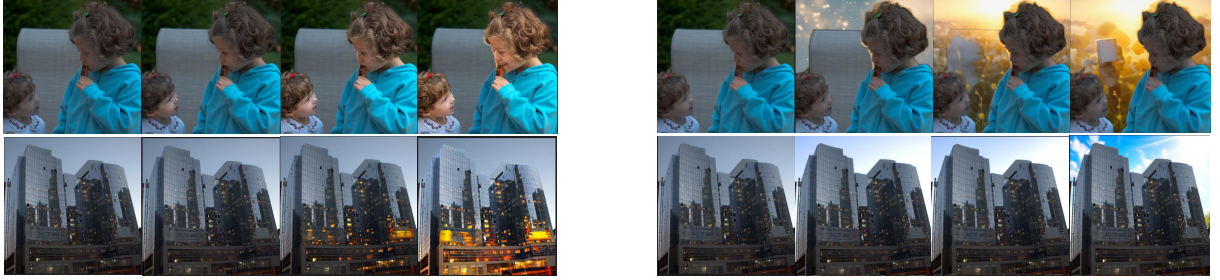


Figure 5: Continuous control visualization of ACE on two source images. Each row shows one sample: the left sweep varies the object slider and the right sweep varies the atmospheric slider.

Metric	OB-LoRA	AT-LoRA	Better
Visual Monotonicity (ρ)	0.990	0.977	$\uparrow$
median / success rate	1.000 / 0.989	1.000 / 0.975	$\uparrow$
Emotion Monotonicity (ρ)	0.678	0.637	$\uparrow$
median / success rate	0.900 / 0.773	0.800 / 0.766	$\uparrow$
Step-CV	0.419	0.524	$\downarrow$
Path Straightness	0.743	0.623	$\uparrow$
Emotion Dynamic Range	5.429	4.623	$\uparrow$

Table 2: Continuous-control analysis over 352 sequences with strengths $\{0, 0.25, 0.5, 0.75, 1.0\}$. ρ : mean Spearman correlation between strength and measured change; success rate: fraction of sequences with $\rho > 0$. Definitions are in Appendix B.4.

Branch	Target Δ	Non-target Δ	Ratio	Leakage
OB-LoRA	FG: 0.152	BG: 0.091	1.68 $\times$	0.434 $\pm$ 0.124
AT-LoRA	BG: 0.274	FG: 0.058	4.76 $\times$	0.258 $\pm$ 0.106
Disentangle Score (chance level: 0.500)				0.729 $\pm$ 0.123

Table 3: Cross-region leakage under single-branch editing. Δ : mean regional LPIPS change inside the foreground (FG) or background (BG) mask; Leakage: per-sample non-target/target ratio ($\downarrow$); Disentangle Score: proportion of change inside the intended region ($\uparrow$).

Emo-S (emotion monotonicity), the coefficient of variation of adjacent-step changes (Step-CV), trajectory straightness in DINO feature space, and the emotion dynamic range over the full sweep. Both branches are almost perfectly monotonic in visual change (median $\rho = 1.0$) and positively monotonic in emotion, and the non-trivial dynamic ranges rule out a degenerate no-change solution. OB-LoRA has a lower Step-CV and a straighter trajectory than AT-LoRA, which varies slightly more because it jointly modifies spatially distributed attributes such

as lighting, color tone, and ambience.

Cross-region leakage and disentanglement. Table 3 quantifies whether each branch modifies only its intended region. Under OB-only editing the foreground changes $1.68\times$ more than the background, and under AT-only editing the background changes $4.76\times$ more than the foreground, giving leakage ratios of 0.434 and 0.258. The Disentangle Score of 0.729 is well above the 0.5 chance level of a non-selective edit, confirming that the two branches primarily act within their own regions.

4.7 User Study

We further conduct a user study with 24 participants, each answering 80 questions, yielding 1,920 human judgments in total. Participants are shown edited results from ACE and three affective baselines, AIF, EmoEdit, and EmoEditor, under the same source image and target emotion. For each comparison, participants are asked to select the preferred result from four aspects: structural similarity (StructSim), emotion activation (EmoAct), balance between content preservation and emotional expression, and aesthetic quality (Aesthetics).

As shown in Figure 7, ACE is consistently preferred across all four criteria. The gains are most evident in Balance and Aesthetics, suggesting better content preservation and visual quality. Higher preferences in StructSim and EmoAct further show that object-atmospheric control improves both



Figure 6: Counterfactual continuous slider editing. Object and atmospheric sliders are independently assigned different target emotions to generate continuous counterfactual images.

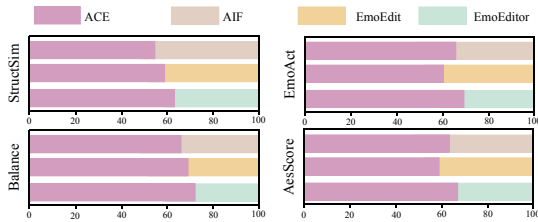


Figure 7: User study results on pairwise preference selection between ACE and each comparative method in terms of structural similarity, emotion activation, balance and aesthetics.

structural consistency and target-emotion expression, aligning with the automatic evaluation. The study is limited in scale and demographic coverage and serves as complementary human validation (see Limitations).

Variant	Emo-Acc8	TEA	Emo-S	DINO-I	LPIPS	AesScore	ID
Global Joint LoRA	0.402	0.337	3.214	0.712	0.534	5.421	0.781
Single LoRA	0.467	0.348	4.722	0.741	0.578	5.532	0.779
Dual LoRA w/o Supp	0.502	0.350	5.232	0.736	0.570	5.556	0.780
ACE	0.564	0.353	6.327	0.755	0.553	5.583	0.807

Table 4: Ablation results of ACE. Higher is better for Emo-Acc8, TEA, Emo-S, DINO-I, AesScore, and ID. Lower is better for LPIPS.

4.8 Ablation Study

Table 4 evaluates the contribution of each component in ACE. All variants share the same frozen backbone, ADA-constructed data, prompts, and training settings, isolating the parameterization and loss design. Global Joint LoRA trains a single adapter on the full-image targets x_f only; Single LoRA trains one adapter on all three counterfactual states; Dual LoRA w/o Supp uses the two routed branches without the CRS loss. Compared with Global Joint LoRA, Single LoRA improves affective expression by learning region-aware editing cues. Separating the adaptation into dual LoRA branches further strengthens the target emotion, while the full ACE achieves the best overall trade-off after adding regional suppression, which also improves DINO-I and ID. ADA, the counterfactual targets, and the masks are not ablated individually, since they jointly define the training states and regional losses.

The qualitative comparison in Appendix C.3 shows the same trend: Dual LoRA without suppression may over-amplify atmospheric cues, whereas ACE expresses the target emotion clearly while keeping region-specific edits under control.

5 Conclusion

We presented ACE, an object-atmospheric continuous control framework for fine-grained affective image editing. By decomposing affective expression into object and atmospheric cues and training two lightweight LoRA branches with counterfactual regional supervision and a suppression loss, ACE enables controllable affective editing with reduced region leakage. Experiments show improved target-emotion expression with competitive preservation, quality, and identity consistency, and quantitative analyses confirm that the two learned directions are monotonic and largely disentangled. These findings hold for the eight emotions, test sets, and participants evaluated here and are not universal claims about affective perception.

Limitations

Although ACE demonstrates promising results for controllable affective image editing, several aspects can be further explored. First, visual emotion is inherently complex, and real-world affective states go beyond the eight emotion categories considered in this work. Extending the range of affective factors and target emotion categories may further improve the diversity and expressiveness of editing results. Second, affective perception is subjective, and viewers with different cultural backgrounds or personal experiences may interpret the same visual cues in different ways. Our user study involves 24 participants and does not cover diverse cultural backgrounds, so its results should be understood as complementary human validation rather than as a universal measure of affective preference. Future work may consider more personalized or context-aware affective representations, together with larger and more diverse human evaluations, to better reflect such diversity in human emotion perception. Third, the training supervision is produced by an automated agent built on foundation models and filtered with automatic metrics. Although the generation, filtering, and editing models are distinct and the evaluation uses an independent test set and human judgments, the learned editor may still inherit biases and failure modes of these components (see Appendix C.5), our claim is therefore that ADA provides scalable and effective region-level supervision for the evaluated emotions, not that it captures universal principles of human affect. Fourth, the default slider strengths are set per emotion (Table 5), while the sliders are shown to be monotonic, choosing these defaults automatically remains open. Finally, automatic emotion metrics depend on learned classifiers and may reward stereotyped cues, so we complement them with preservation metrics and human feedback; more interactive and personalized evaluation protocols would further strengthen affective editing research.

Acknowledgments

This work was partially supported by the National Natural Science Foundation of China under Grant W2432042, the Anhui Province Science and Technology Innovation Project under Grant 202523009050013, and the USTC Research Funds of the Double First-Class Initiative under Grant YD2105202623.

References

- Omri Avrahami, Dani Lischinski, and Ohad Fried. 2022. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18208–18218.
- Tim Brooks, Aleksander Holynski, and Alexei A. Efros. 2023. InstructPix2Pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402.
- Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. 2021. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9650–9660.
- Xing Cui, Zekun Li, Pei Li, Yibo Hu, Hailin Shi, Chunshui Cao, and Zhaofeng He. 2023. [ChatEdit: Towards multi-turn interactive facial image editing via dialogue](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14567–14583, Singapore. Association for Computational Linguistics.
- Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. 2019. ArcFace: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4690–4699.
- Prafulla Dhariwal and Alexander Nichol. 2021. Diffusion models beat GANs on image synthesis. In *Advances in Neural Information Processing Systems*.
- Shaojing Fan, Zhiqi Shen, Ming Jiang, Bryan L. Koenig, Juan Xu, Mohan S. Kankanhalli, and Qi Zhao. 2018. Emotional attention: A study of image sentiment and visual attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7521–7531.
- Tsu-Jui Fu, Xin Eric Wang, and William Yang Wang. 2022. Language-driven artistic style transfer. In *European Conference on Computer Vision*, pages 717–734. Springer.
- Yuxiang Fu, Jiakun Ding, Renzhi Wang, Qian Fu, Ivor W. Tsang, Ming-Ming Cheng, and Qing Guo. 2025. [Benchmarking Drag* for eye direction transformation and beyond](#). *Visual Intelligence*, 3(29).
- Rohit Gandikota, Joanna Materzyńska, Tingrui Zhou, Antonio Torralba, and David Bau. 2024. Concept sliders: LoRA adaptors for precise control in diffusion models. In *European Conference on Computer Vision*.
- Chuanji Gao, Susan Ajith, and Marius V Peelen. 2025. Object representations drive emotion schemas across a large and diverse set of daily-life scenes. *Communications Biology*, 8(1):697.

- Tanmay Gupta and Aniruddha Kembhavi. 2023. Visual programming: Compositional visual reasoning without training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14953–14962.
- Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. 2023. Prompt-to-prompt image editing with cross-attention control. In *International Conference on Learning Representations*.
- Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.
- Francisco Ibarrola, Rohan Lulham, and Kazjon Grace. 2024. Affect-conditioned image generation. *IEEE Transactions on Affective Computing*, 15(4):2169–2179.
- Taehun Kim, Kunhee Kim, Joonyeong Lee, Dongmin Cha, Jiho Lee, and Daijin Kim. 2022. Revisiting image pyramid structure for high resolution salient object detection. In *Proceedings of the Asian Conference on Computer Vision*, pages 108–124.
- Dongxu Li, Junnan Li, and Steven C. H. Hoi. 2023. BLIP-Diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. In *Advances in Neural Information Processing Systems*.
- Feng Li, Hao Zhang, Peize Sun, Xueyan Zou, Shilong Liu, Chunyuan Li, Jianwei Yang, Lei Zhang, and Jianfeng Gao. 2024a. Segment and recognize anything at any granularity. In *European Conference on Computer Vision*, pages 467–484.
- Suyan Li, Fuxiang Huang, and Lei Zhang. 2025. A survey of multimodal composite editing and retrieval. *Visual Intelligence*, 3(15).
- Zonglin Li, Xiaoqian Lv, Wei Yu, Qinglin Liu, Jingbo Lin, and Shengping Zhang. 2024b. Face shape transfer via semantic warping. *Visual Intelligence*, 2(26).
- Qing Lin, Jingfeng Zhang, Yew-Soon Ong, and Mengmi Zhang. 2025. Make me happier: Evoking emotions through image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16367–16376.
- Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. 2023. Flow matching for generative modeling. In *International Conference on Learning Representations*.
- Da Liu, Yaxi Jiang, Min Pei, and Shiguang Liu. 2018. Emotional image color transfer via deep learning. *Pattern Recognition Letters*, 110:16–22.
- Shiguang Liu, Huixin Wang, and Min Pei. 2022. Facial-expression-aware emotional color transfer based on convolutional neural network. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 18(1):8:1–8:19.
- Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. 2024a. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*.
- Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng Shen, Tianjun Zhang, Yu Su, Huan Sun, and 3 others. 2024b. AgentBench: Evaluating LLMs as agents. In *International Conference on Learning Representations*.
- Jana Machajdik and Allan Hanbury. 2010. Affective image classification using features inspired by psychology and art theory. In *Proceedings of the ACM International Conference on Multimedia*, pages 83–92.
- Qi Mao, Haobo Hu, Yujie He, Difei Gao, Haokun Chen, and Libiao Jin. 2026. EmoAgent: A multi-agent framework for diverse affective image manipulation. *IEEE Transactions on Affective Computing*, 17(2):1618–1635.
- Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. 2023. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6038–6047.
- Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695.
- Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems*, 36:68539–68551.
- Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. 2023. Hugging-GPT: Solving AI tasks with ChatGPT and its friends in Hugging Face. In *Advances in Neural Information Processing Systems*.
- Jing Shi, Ning Xu, Yihang Xu, Trung Bui, Franck Dérioncourt, and Chenliang Xu. 2021. Learning by planning: Language-guided global image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13590–13599.

- Yasheng Sun, Bohan Li, Mingchen Zhuge, Deng-Ping Fan, Salman Khan, Fahad Shahbaz Khan, and Hideki Koike. 2025. [Connecting dreams with visual brainstorming instruction](#). *Visual Intelligence*, 3(12).
- Dídac Surís, Sachit Menon, and Carl Vondrick. 2023. ViperGPT: Visual inference via python execution for reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11888–11898.
- Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. 2023. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1921–1930.
- Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, Wayne Xin Zhao, Zhewei Wei, and Jirong Wen. 2024a. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6):186345.
- Zhenyu Wang, Aoxue Li, Zhenguo Li, and Xihui Liu. 2024b. GenArtist: Multimodal LLM as an agent for unified image generation and editing. In *Advances in Neural Information Processing Systems*.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.
- Shuchen Weng, Peixuan Zhang, Zheng Chang, Xinlong Wang, Si Li, and Boxin Shi. 2023. Affective image filter: Reflecting emotions from text to images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10810–10819.
- 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, and 20 others. 2025. Qwen-Image technical report. *arXiv preprint arXiv:2508.02324*.
- Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. 2023. Visual ChatGPT: Talking, drawing and editing with visual foundation models. *arXiv preprint arXiv:2303.04671*.
- Junlin Xie, Zhihong Chen, Ruifei Zhang, Xiang Wan, and Guanbin Li. 2024. Large multimodal agents: A survey. *arXiv preprint arXiv:2402.15116*.
- Jingyuan Yang, Jiawei Feng, and Hui Huang. 2024a. 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.
- Jingyuan Yang, Jiawei Feng, Weibin Luo, Dani Lischinski, Daniel Cohen-Or, and Hui Huang. 2025. EmoEdit: Evoking emotions through image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24690–24699.
- Jingyuan Yang, Xinbo Gao, Leida Li, Xiumei Wang, and Jinshan Ding. 2021. SOLVER: Scene-object interrelated visual emotion reasoning network. *IEEE Transactions on Image Processing*, 30:8686–8701.
- Jingyuan Yang, Qirui Huang, Tingting Ding, Dani Lischinski, Danny Cohen-Or, and Hui Huang. 2023a. EmoSet: A large-scale visual emotion dataset with rich attributes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20383–20394.
- Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. 2024b. Depth anything V2. In *Advances in Neural Information Processing Systems*.
- Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. 2023b. MM-REACT: Prompting ChatGPT for multimodal reasoning and action. *arXiv preprint arXiv:2303.11381*.
- Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*.
- Jianbo Yuan, Sean Mcdonough, Quanzeng You, and Jiebo Luo. 2013. Sentribute: Image sentiment analysis from a mid-level perspective. In *Proceedings of the Workshop on Issues of Sentiment Discovery and Opinion Mining*, pages 10:1–10:8.
- Arman Zarei, Samyadeep Basu, Mobina Pourneamat, Sayan Nag, Ryan Rossi, and Soheil Feizi. 2025. [SliderEdit: Continuous image editing with fine-grained instruction control](#). *arXiv preprint arXiv:2511.09715*.
- Jing Zhang and Bingjie Fan. 2026. [EmoKGEdit: Training-free affective injection via visual cue transformation](#). *arXiv preprint arXiv:2601.12326*.
- Kai Zhang, Lingbo Mo, Wenhui Chen, Huan Sun, and Yu Su. 2023a. MagicBrush: A manually annotated dataset for instruction-guided image editing. In *Advances in Neural Information Processing Systems*.
- Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023b. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. 2018. 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.

Sicheng Zhao, Guiguang Ding, Qingming Huang, Tat-Seng Chua, Björn W. Schuller, and Kurt Keutzer. 2018. Affective image content analysis: A comprehensive survey. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 5534–5541.

A Dataset Construction Details

This section provides additional details about the construction of the counterfactual object–atmospheric editing dataset. We describe the image sources, the two-stage filtering pipeline, the agent prompt and tool configurations, the quality-control criteria and rejection statistics, and the computational cost of data generation.

A.1 Source Image Collection

We collect candidate images from three large-scale image datasets: EmoSet, MagicBrush, and MA5K. These datasets provide diverse emotion-related, instruction-editing, and high-quality photographic images, covering both subject-centric and scene-centric cases. We then apply a two-stage filtering pipeline to identify images that are suitable for region-level affective editing.

A.2 Geometric and Saliency Filtering

The first stage uses lightweight visual tools to remove images that are unlikely to support reliable region-level editing. Specifically, we use InSPyReNet to estimate salient foreground regions and Depth Anything V2 to estimate monocular depth. These cues help identify images with clear foreground–background separation, compact editable subjects, and sufficient spatial structure. Images with ambiguous foregrounds, fragmented subjects, or weak geometric separation are removed or sent for manual review.

A.3 Second-stage VLM Editability Assessment

The second stage uses a vision-language model to assess whether the remaining images have sufficient potential for high-level secondary editing and emotional activation. Each image is evaluated together with a short reference description of the original foreground or scene. The model is

prompted to judge three aspects on a 0–3 scale: foreground–background separation, editability under background or lighting changes, and emotion activation potential after color grading or visual effects. Based on these judgments, images are assigned to keep, review, or reject groups. This stage filters out visually flat, cluttered, or emotionally limited images that are unlikely to benefit from object–atmospheric editing.

A.4 ADA Agent Design

Tool configurations. In our implementation, ADA uses Qwen3-VL-32B-Instruct as the multimodal reasoning model, and the `image_editor` tool is instantiated with FLUX.2 [dev] for high-quality image editing. The perception tools are Grounding DINO (Liu et al., 2024a) for open-vocabulary detection, Semantic-SAM (Li et al., 2024a) for instance segmentation, Depth Anything V2 (Yang et al., 2024b) for monocular depth, and InSPyReNet (Kim et al., 2022) for saliency; all use their official checkpoints and default settings. The object mask M_s is the union of the Semantic-SAM instance masks of the subjects selected by the agent, refined by the InSPyReNet saliency map, and the atmospheric mask is its complement, $M_a = 1 - M_s$. The agent runs the ReAct-style loop below with at most 20 tool calls per sample; samples that exceed this budget or fail a verification step are discarded.

Instruction generation. Given the perception results, the agent first writes an object instruction τ_s describing concrete, visually observable changes to the salient subjects (expression, pose, action, object state, or inter-subject relation) and an atmospheric instruction τ_a describing changes to illumination, color tone, weather, or ambience. Both instructions must reference entities detected in the image, which grounds the edit in the source content and discourages templated, image-agnostic edits. The object-only target x_s is produced by applying τ_s with a relatively high editing strength, the atmospheric-only target x_a by applying τ_a to the source with a lower strength, and the joint target x_f by applying τ_a on top of x_s .

Prompt template. The following prompt template is used to guide ADA. We divide the prompt into four parts for readability: agent role, tool specification, operating procedure, and response format.

[Role and Objective]

You are an autonomous Emotional Image Editor Agent. You will receive an initial image, its object

description, and a target emotion. Your goal is to orchestrate visual perception tools and image editing tools to make the image convey the target emotion, while strictly preserving the main subject identity and the core spatial structure.

[Toolbox]

`grounding_dino`: open-vocabulary object detection. Use it to locate main subjects and key objects, and to verify whether object positions and counts are preserved after editing.

`depth_anything`: monocular depth estimation. Use it to record foreground-background depth structure and to verify that style edits do not change the subject depth layer.

`object_sam`: pixel-level instance segmentation. Use it to obtain object masks, check subject connectivity, and compare subject masks before and after editing.

`inspyrenet`: saliency detection. Use it to identify the protected salient subject region and to verify that editing does not erode the main foreground.

`facial_action_analysis`: face detection and facial action unit analysis. Use it only when faces are present, and verify whether expression-related action units match the target emotion.

`image_editor`: diffusion-based image editing. The editing prompt must describe concrete and visually observable changes. It should avoid abstract emotional phrases such as “make it terrifying” and instead specify actionable visual modifications.

[Standard Operating Procedure]

Step 1: Deep observation. Do not edit yet. Use perception tools to verify the initial description and record subject, mask, depth, and expression cues.

Step 2: Object planning. Compare the observed image with the target emotion and plan the object edit.

Step 3: Object editing. Call `image_editor` with a subject-focused instruction and a relatively high editing strength.

Step 4: Object verification. Use perception tools again. If subject identity or spatial structure is damaged, stop and mark the sample as failed.

Step 5: Atmospheric editing. Call `image_editor` with a background, lighting, color, or ambience instruction using a lower editing strength.

Step 6: Atmospheric verification. Verify that the atmosphere is changed while the main subject and structure remain intact.

Step 7: Finish. Output `done` when the target emotion is successfully evoked.

[Response Format]

The agent must output only one valid JSON object:

```
{
  "thought": "SOP step.",
  "action": "Tool name or done",
  "params": {"key": "value"}
}
```

A.5 Quality Control

Beyond the verification and revision steps performed by the agent itself, we apply an additional automatic quality validation stage to the generated counterfactual edits. We first use CLIP image similarity to constrain source preservation, keeping edits whose similarity to the source image falls within the range of 0.75 to 0.90. We then use CLIP text-image similarity to measure whether the edited image is consistent with the corresponding editing instruction, and retain samples with scores between 0.25 and 1.00. For object-only and atmospheric-only regional edits, both CLIP-based scores are computed on the corresponding masked regions rather than on the full image. Finally, we evaluate target-emotion activation with an emotion classifier (the EmoSet-trained ResNet/CLIP classifier used for Emo-Acc8) and require the softmax score of the target emotion to be higher than 0.30. Samples failing any of these checks are discarded from the training set. The image-similarity range excludes both near-identity edits (similarity above 0.90, i.e., the instruction was not executed) and excessive edits (below 0.75, i.e., the source content was replaced); the instruction-consistency and emotion thresholds ensure that retained edits are both faithful to the instruction and affectively effective.

A.6 Computational Cost of Data Generation

ADA is executed once, offline, to build the dataset, and none of its components (the multimodal reasoning model, the perception tools, or the FLUX.2 editor) is used during OACE training or inference. Generating one training sample requires one multimodal reasoning loop and, on average, 15 perception tool calls and two to three `image_editor` calls, which takes approximately 200 seconds on a single H100; the complete dataset was generated in approximately 168 GPU-hours. Once the dataset is built, OACE has the same cost as standard LoRA fine-tuning of the backbone: with four NVIDIA H100 GPUs, training takes approximately 30 hours, compared with 20 hours for EmoEdit and 38 hours for EmoEditor under the same hardware. At inference, OACE adds only the two rank-32 LoRA branches to the frozen backbone, so its latency and memory footprint are essentially those of Qwen-Image-Edit.

B Training and Evaluation Details

This section documents the implementation details of ACE, including the pretrained backbone, LoRA insertion layers, optimization settings, training schedule, inference-time slider control, and the evaluation protocol.

B.1 Adapter Configuration and Optimization

We use the pretrained Qwen-Image-Edit model as the frozen editing backbone. OB-LoRA and AT-LoRA are inserted into the linear layers of every MM-DiT block, with the LoRA rank set to 32 and $\alpha = 1$ for both branches. The remaining backbone parameters are kept frozen during training. We train with AdamW at a learning rate of $1e-4$, a batch size of 1, and 500 training steps at a resolution of 1024×1024 , on four NVIDIA H100 GPUs. The state router samples the three training states with probabilities $p(s) = 1/3$, $p(a) = 1/3$, and $p(f) = 1/3$, and the loss weights are $\lambda_{\text{edit}} = 1$ and $\lambda_{\text{sup}} = 1$. The neutral preservation prompt c_{neu} is "Keep the image same".

B.2 Inference and Slider Scaling

At inference time, the object and atmospheric sliders are controlled by scaling the two learned LoRA directions with λ_s and λ_a . We use emotion-specific default strengths for the object slider (OB) and atmospheric slider (AT), as shown in Table 5. These values provide the default inference configuration for each target emotion, while users can further adjust the two sliders for finer control. The defaults were chosen empirically, with stronger atmospheric modulation for high-arousal negative emotions such as fear and anger and milder settings for contentment; Section 4.6 shows that the sliders are monotonic, so the defaults act as an operating point rather than as a requirement for controllability.

B.3 Evaluation Protocol and Metric Implementation

All methods are evaluated on the same 2,768 (source image, target emotion) pairs of the main test set and on the EmoEdit test set with identical evaluators and image resolution. Emo-Acc8 and Emo-S follow EmoGen (Yang et al., 2024a): an eight-way emotion classifier predicts the emotion distribution of the edited image; Emo-Acc8 is the fraction of edits whose argmax equals the target emotion, and Emo-S is the increase of the target-emotion logit relative to the source image.

TEA follows EmoKGEdit (Zhang and Fan, 2026) and measures the activation strength of the target emotion. DINO-I is the cosine similarity between DINO ViT (Caron et al., 2021) CLS features of the edited and source images; LPIPS uses the AlexNet variant (Zhang et al., 2018); AesScore uses the LAION aesthetic predictor. ID is computed on person-centric samples using ArcFace (Deng et al., 2019) embeddings when a face is detected and CLIP embeddings of the salient subject otherwise. Baselines use their official code and settings.

B.4 Continuous-Control Metrics

For the analysis in Section 4.6, we use 352 (source image, target emotion) sequences covering all eight emotions. For each sequence and branch $b \in \{s, a\}$, we set the other branch to zero and sweep $\lambda_b \in \{0, 0.25, 0.5, 0.75, 1.0\}$, producing images $\{y_0, \dots, y_4\}$ with $y_0 = x_{\text{src}}$. Let M_b be the target mask of the branch (M_s for OB-LoRA and $M_a = 1 - M_s$ for AT-LoRA), and let $d_M(x, y)$ denote the LPIPS distance inside mask M .

Visual monotonicity is the Spearman correlation between the slider strength λ_b and the regional change $d_{M_b}(x_{\text{src}}, y_i)$ across the five levels, averaged over sequences; we also report the median and the success rate, i.e., the fraction of sequences with positive correlation.

Emotion monotonicity is defined analogously with Emo-S of y_i in place of the regional change.

Step-CV is the coefficient of variation (standard deviation divided by mean) of the adjacent-step changes $\{d_{M_b}(y_{i-1}, y_i)\}_{i=1}^4$; lower values indicate more uniform transitions between adjacent levels.

Path straightness is the ratio between the direct distance and the path length of the editing trajectory in DINO feature space, $\|f(y_4) - f(y_0)\|_2 / \sum_{i=1}^4 \|f(y_i) - f(y_{i-1})\|_2$, where $f(\cdot)$ is the DINO CLS embedding; a value of 1 corresponds to a perfectly straight trajectory.

Emotion dynamic range is the Emo-S of y_4 , i.e., the total change of target-emotion intensity over the full sweep.

Leakage and Disentangle Score. Let Δ_{fg} and Δ_{bg} denote the regional change $d_M(x_{\text{src}}, y)$ computed inside the foreground mask M_s and the background mask M_a for a single-branch edit y at

Target Emotion	OB	AT
Amusement	0.75	0.50
Awe	0.75	1.00
Contentment	0.50	0.50
Excitement	1.00	0.75
Disgust	1.00	0.75
Anger	1.00	1.00
Fear	1.00	1.25
Sadness	1.00	1.00

Table 5: Default object (OB) and atmospheric (AT) slider strengths for each target emotion.

$\lambda_b = 1$. Under OB-only editing,

$$\text{OB-Leak} = \frac{\Delta_{bg}}{\Delta_{fg}}, \quad (6)$$

and under AT-only editing,

$$\text{AT-Leak} = \frac{\Delta_{fg}}{\Delta_{bg}}. \quad (7)$$

The Disentangle Score averages, over both branches, the proportion of change that falls inside the intended region,

$$\text{DS} = \frac{1}{2} \left(\frac{\Delta_{fg}^s}{\Delta_{fg}^s + \Delta_{bg}^s} + \frac{\Delta_{bg}^a}{\Delta_{fg}^a + \Delta_{bg}^a} \right), \quad (8)$$

where superscripts s and a denote OB-only and AT-only edits. A non-selective edit that changes both regions equally yields $\text{DS} = 0.5$. All leakage scores are computed per sample and then averaged (mean $\pm$ standard deviation in Table 3), whereas the reported foreground and background changes are averaged independently; the mean leakage therefore does not exactly equal the ratio of the two mean changes.

C Additional Results and Analysis

This section collects additional qualitative and quantitative results that support the main claims of the paper.

C.1 Continuous Control Analysis

Figures 8 and 9 provide additional one-dimensional slider sweeps for the two learned control directions. The object slider mainly changes foreground affective cues, whereas the atmospheric slider gradually adjusts background illumination, color tone, and ambience. Figures 10 and 11 further show two-dimensional OB–AT sweeps for contentment and sadness, demonstrating that the two sliders can jointly compose emotion-specific object and atmospheric variations. The quantitative counterpart of these sweeps is reported in Tables 2 and 3.

C.2 Additional Qualitative Examples

Figure 12 shows neutral-to-emotion editing results, where ACE introduces target-consistent object details and atmospheric changes while preserving the original layout and main subject. Negative emotions are expressed through darker illumination, desaturated tones, and adverse visual cues, whereas positive or high-arousal emotions are conveyed through brighter lighting, warmer or more vivid colors, and emotion-relevant object cues.

Figure 13 further shows emotion-to-emotion transfer results. These examples indicate that ACE does not simply amplify the source emotion, but can redirect the affective state by jointly adjusting object-level cues and atmospheric factors, enabling flexible transformations across different emotional starting points.

C.3 Qualitative Ablation

Figure 14 shows qualitative results of the ablation variants. Global Joint LoRA often produces weak affective changes, whereas Single LoRA introduces clearer object or atmospheric cues. Dual LoRA without suppression can generate stronger edits, but may over-amplify atmospheric changes or introduce excessive emotional cues. In contrast, ACE expresses the target emotion more clearly while keeping the source layout and region-specific edits better controlled.

C.4 Challenging Scenes

Affective editing is most difficult when the object factor is not concentrated on a single salient subject. We observe three recurring hard cases. First, in crowded scenes with many small subjects (e.g., street crowds or group photos), the object mask M_s covers many disconnected regions, and OB-LoRA tends to modify only the largest subjects while the remaining ones stay neutral; the atmospheric branch still shifts the overall mood, so the edit reads as a scene-level change rather than a coordinated change of all subjects. Second, in images without a clear foreground (e.g., wide landscapes or textures), the object mask degenerates to a small or empty region, and the edit is carried almost entirely by AT-LoRA, so emotions that require object cues (amusement, disgust) are expressed less clearly than those dominated by atmosphere (sadness, awe). Third, when the source emotion is strong and opposite to the target (e.g., fear to contentment), the object edit must change expressions

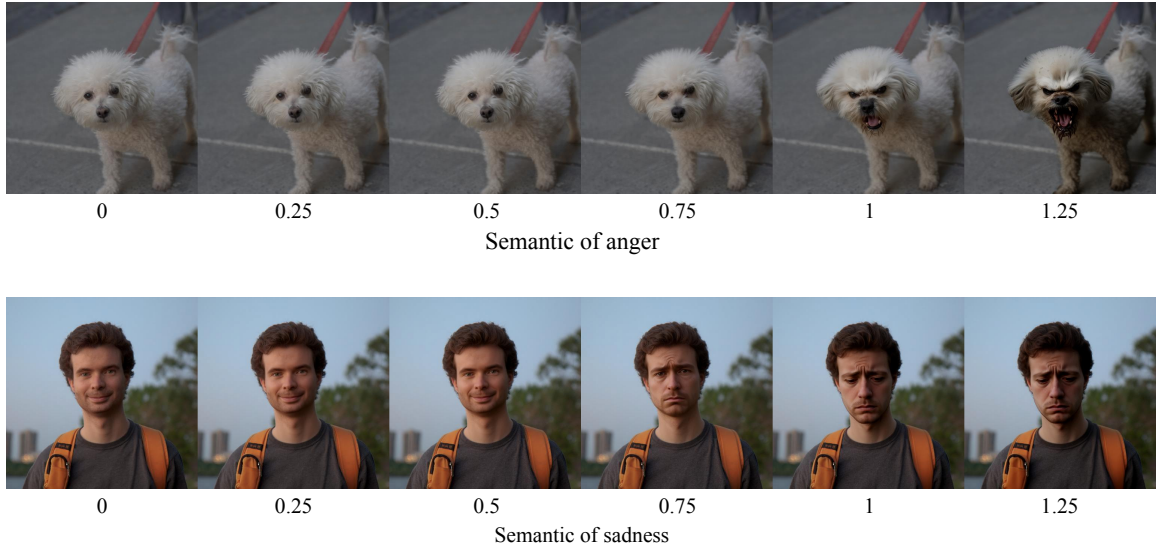


Figure 8: Continuous control of the object slider.

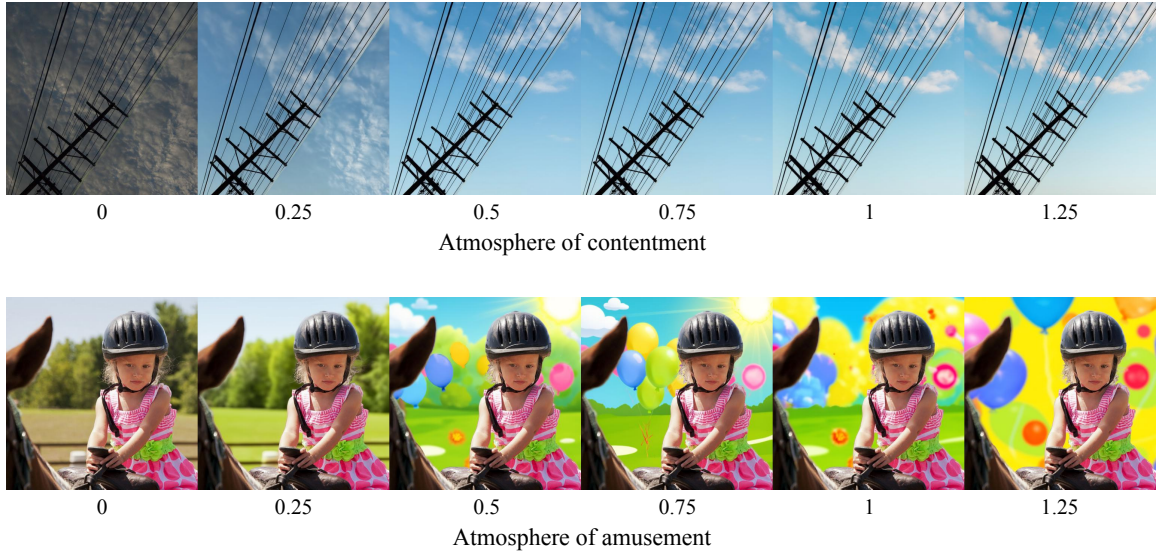


Figure 9: Continuous control of the atmospheric slider.



Figure 10: OB-AT joint sweep for contentment.

and poses substantially, which increases LPIPS and occasionally alters fine identity details. Examples of the first two cases are shown in Figure 15(c) and

(a), respectively.

C.5 Failure Case Analysis

Figure 15 shows three representative failures of ACE, which we group into the categories below.

Missing object factor. When the source image has no salient subject, as in the rocky shore of Figure 15(a), the object mask M_s degenerates to a small or empty region and OB-LoRA has nothing to act on. The edit is then carried almost entirely by AT-LoRA, which can only darken or recolor the scene; the result is a low-contrast image with blurred surface texture in which the target emotion (anger) is weakly expressed. Emotions that rely on object cues are therefore harder to evoke on such

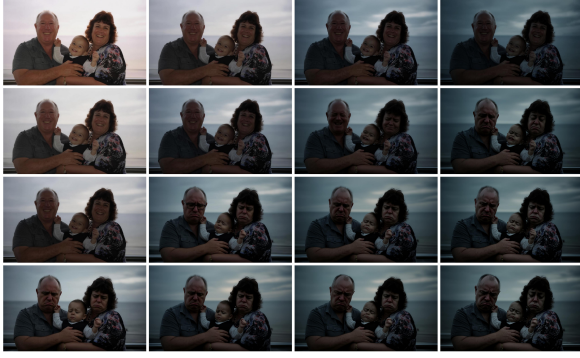


Figure 11: OB-AT joint sweep for sadness.

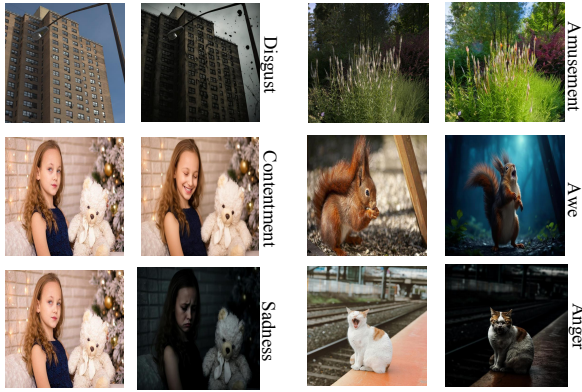


Figure 12: Additional qualitative examples from neutral source images to specified target emotions.



Figure 13: Additional qualitative examples from source images with existing emotions to related or opposite target emotions.

inputs than atmosphere-dominated ones such as sadness or awe.

Over-editing and inappropriate cues. At high object strength, the object branch occasionally exaggerates the affective cues far beyond what the target requires. In Figure 15(b), editing an excited source toward disgust turns a nail file into a blade and adds wounds and stains to the arm and clothing, which is both an unsuitable depiction and a change of subject identity, and is reflected in a lower ID score. This behavior inherits the stereo-

typed associations between negative emotions and harmful objects present in the foundation models used by ADA and in the emotion classifier used for filtering; reducing λ_s recovers identity at the cost of weaker expression, and we consider stronger content constraints on the generated targets an important direction for future work.

Crowded scenes and text. When many small subjects are present, as in the protest scene of Figure 15(c) edited from anger toward contentment, the object mask covers many disconnected regions and OB-LoRA modifies only the largest faces, which are re-synthesized rather than re-expressed, so the identities of the foreground participants are lost while the remaining crowd stays unchanged. Text regions inside the mask are also corrupted, since the backbone does not preserve legible characters under strong object edits. Such scenes call for instance-level masks and explicit text preservation, which the current pipeline does not provide.

Other observed limitations. Beyond these examples, two further error sources appear in the automatic metrics. Semantically close targets, especially amusement versus excitement and contentment versus awe, account for most Emo-Acc8 errors: the edit moves the image toward the correct valence and arousal, but the classifier assigns the neighboring category, consistent with the confusion observed for prior methods in Section 1. In addition, when the perception tools mis-segment the subject during data construction (e.g., merging a person with a held object), the counterfactual targets and regional losses are defined on an incorrect region, and the trained model may leak object edits into the background for similar images; this motivates the leakage analysis in Section 4.6.

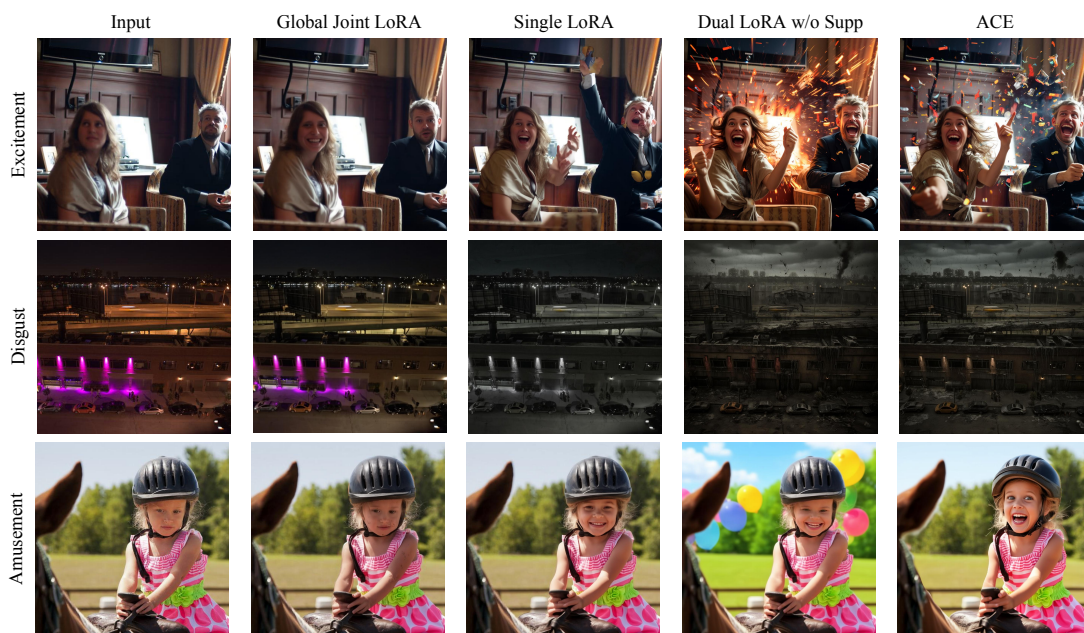


Figure 14: Qualitative ablation comparison of the variants in Table 4.

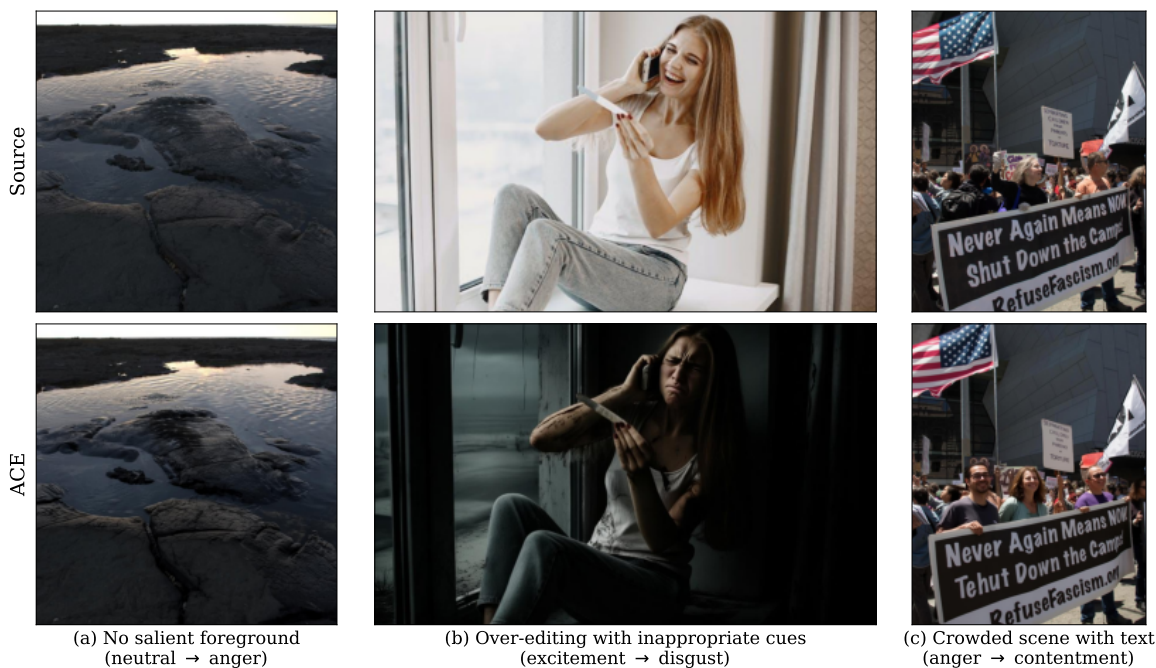


Figure 15: Representative failure cases of ACE (top: source; bottom: edited result). (a) A scene without a salient foreground edited toward anger: the object mask is nearly empty, so the edit is carried by the atmospheric branch alone and mostly darkens the scene while blurring surface texture, leaving the target emotion weakly expressed. (b) Over-editing with inappropriate cues (excitement to disgust): the object branch not only changes the expression but turns the nail file into a blade and adds wounds and stains, an exaggerated and unsuitable rendering of a negative emotion that also alters the subject’s identity. (c) A crowded scene with text (anger to contentment): the faces of the foreground participants are replaced rather than re-expressed, and the banner text is corrupted.