IECLA: INTERPRETABLE EMOTION CONTRIBUTION LEARNING FOR AFFECTIVE IMAGE EDITING

Anonymous authors
Paper under double-blind review



Figure 1: Affective image editing with IECLA across diverse scenes and target emotions. Given only a target-emotion word, IECLA produces edits that effectively evoke the requested emotion while preserving the source image's content and structure.

ABSTRACT

Affective image editing (AIE) aims to evoke a target emotion while preserving the content and structure of a source image. Existing methods improve emotional expression or visual fidelity, but the target emotion alone does not reveal which source regions should be changed and which should remain intact, especially in complex scenes. We introduce Interpretable Emotion Contribution Learning for Affective Image Editing (IECLA), a framework that separates source-emotion attribution from target-aware editing. IECLA constructs signed object-level supervision through counterfactual interventions and trains a context-conditioned contribution predictor to localize evidence that supports or opposes the source emotion. At inference, an image-level evaluator estimates the source emotion directly from the input, while the predictor identifies editable and preservation evidence without repeated interventions. A target-aware editing agent then translates this evidence and the desired emotion into localized, content-aware operations for existing editors. Extensive evaluations show that IECLA predicts counterfactual contributions more faithfully on both in-domain and cross-dataset tests, while external VLM and independent human references support the quality of its localized evidence. Across six editing backbones, IECLA consistently strengthens target-emotion expression while preserving image content and structure, and user studies further show improved perceived target expression and naturalness.[1]

[1]Code will be released in the final version of the paper.

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

1 INTRODUCTION

Affective image editing (AIE) seeks to evoke a target emotion while preserving the source image's identity and structure. This objective differs from an attribute-level instruction: an emotion word specifies a perceptual outcome, but does not indicate which visual evidence should change. The relevant cues may lie in a person's expression, surrounding objects, or the broader scene context.

Modern image editors combine textual instructions with attention and spatial controls (Meng et al., 2022; Hertz et al., 2023; Brooks et al., 2023; Avrahami et al., 2022; Zhang et al., 2023b;a). Such controls are effective when the desired attribute or region is explicit, but they leave the source-side decision implicit for affective editing (Fu et al., 2024). Existing affective editors expose a related tension: methods that produce stronger emotion shifts can introduce artifacts or structural drift, whereas more constrained methods preserve the source but make weak or misplaced changes when the relevant cues are unspecified (Yang et al., 2025; Ye & Huang, 2025; Yin et al., 2026). The ambiguity is amplified in complex scenes, where affect is distributed across multiple objects and contextual relations rather than concentrated in a face or dominant subject.

Such distributed affect makes source-emotion evidence difficult to identify and use for editing. Visual emotion datasets and explanatory annotations characterize content associated with perceived affect (Peng et al., 2015; Kosti et al., 2020; Achlioptas et al., 2021; Yang et al., 2023). Complementary attribution methods expose model evidence through gradients, local surrogates, or perturbations (Selvaraju et al., 2017; Sundararajan et al., 2017; Ribeiro et al., 2016; Petsiuk et al., 2018; Fong & Vedaldi, 2017; Goyal et al., 2019). These resources support recognition and diagnosis, but they do not directly yield reusable, signed object-level contributions that distinguish evidence supporting the source emotion from evidence opposing it. Moreover, localization alone does not specify how to introduce the target emotion while preserving unrelated content (Yang et al., 2024). However, effective AIE consistently requires two coupled decisions: identify and characterize the source evidence, then translate that evidence and the target emotion into an appropriate editing operation. Neither decision follows from the target label alone.

To address this need, we introduce **IECLA**, which learns signed, object-level source-emotion contributions from counterfactual interventions and exposes them as an editing interface. At inference, a frozen evaluator predicts the source emotion from the input, and the contribution predictor assigns regional scores without additional annotations. A target-aware agent combines these signals with the desired emotion to construct constrained instructions for existing editors.

In summary, our main contributions are summarized as follows:

- We formulate signed object-level counterfactual contributions as a supervision signal for source-emotion attribution. The learned contributions distinguish editable source-emotion evidence from regions to preserve, without requiring a source-emotion label at inference.
- We build a unified attribution-guided framework that connects the learned evidence to target-aware image editing. It produces interpretable emotion heatmaps and converts attributed evidence into guidance for image editors without retraining their parameters.
- We validate IECLA through multiple benchmarks, experiments, and backbones. The results show more faithful attribution and stronger target-emotion expression while preserving source structure and naturalness.

2 RELATED WORK

Visual emotion understanding and attribution. Visual affective computing models depict and elicit viewer's emotions with categorical labels, response distributions, scene context, and textual explanations (Peng et al., 2015; Kosti et al., 2020; Achlioptas et al., 2021; Yang et al., 2023). Vision and language approaches further organize affective knowledge into interpretable rationales (Man et al., 2025). Complementary attribution methods expose model evidence through gradients (Selvaraju et al., 2017; Jiang et al., 2021; Sundararajan et al., 2017), local surrogates (Ribeiro et al., 2016), or perturbations (Petsiuk et al., 2018; Fong & Vedaldi, 2017); EmoCAM combines class-activation maps with detected objects for corpus-scale emotion analysis (Doulfoukar et al., 2024), and counterfactual explanations identify evidence whose alteration changes a prediction (Goyal et al., 2019).

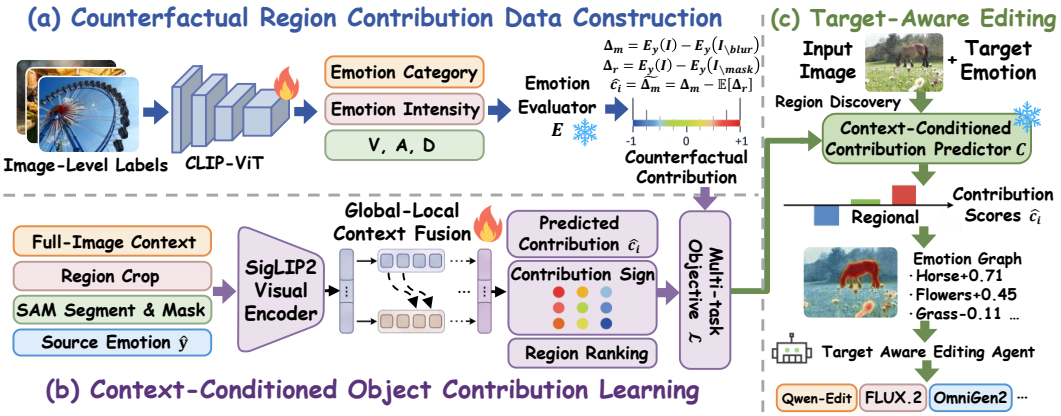


Figure 2: IECLA framework. (a) Counterfactual interventions construct signed supervision. (b) Context-conditioned learning predicts regional contributions. (c) Signed predictions form emotion heatmaps and ground target-aware editing operations.

These resources support recognition and diagnosis, but their supervision or outputs are not defined as reusable, signed object-level controls for affective editing. Perturbation-based estimates require multiple evaluator queries, whereas visually plausible maps do not need to be strictly faithful (Adebayo et al., 2018). IECLA builds on this line by learning intervention-grounded contributions with an explicit sign and an editor-facing interface.

Affective image editing and agentic control. General editors use partial noising (Meng et al., 2022), spatial and attention controls (Avrahami et al., 2022; Zhang et al., 2023b; Hertz et al., 2023), or learned instruction following (Brooks et al., 2023). Affective editing additionally requires translating emotion into image-dependent changes (Yuan et al., 2026; Wei et al., 2025). EmoEdit learns an emotion adapter from content-aware edit pairs (Yang et al., 2025), Moodifier derives contextual attributes through affect-specialized representations and multimodal reasoning (Ye & Huang, 2025), and MooD provides continuous valence–arousal control (Yin et al., 2026). EmoAgent decomposes editing into planning, execution, and critique (Mao et al., 2025), while EmoKGEdit combines affective localization, knowledge-guided cue transformation, and structure-aware rendering (Zhang & Fan, 2026). These approaches advance target-side control. IECLA instead learns signed source-emotion contributions as an explicit interface to target-aware editing.

3 Method

3.1 Task and Framework Overview

Given a source image I and a target emotion t, affective image editing seeks to evoke t while preserving unrelated content. IECLA infers the source emotion y from I, estimates how candidate regions support or oppose y, and uses this signed evidence to guide an existing editor.

Figure 2 shows the complete pipeline. During training, counterfactual interventions construct region-level contributions, and a context-conditioned predictor learns to estimate them from the original image and masks. At inference, the predicted contributions form an emotion heatmap and separate editable from preservation evidence. A target-aware agent combines this evidence with t, verifies feasible changes, and selects a budgeted editing plan. After structural review and instruction auditing, the compiled instruction is passed to the existing editor.

3.2 Counterfactual Contribution Label Construction

To obtain signed region-level supervision, we first infer the source emotion, then measure how individual regions affect its score. Image-level supervision identifies the perceived affect but does not reveal which regions produce it or how they contribute. We use evaluator E, trained with EmoVerse (Guo et al., 2025) image-level supervision, to infer the source emotion from the full image:

$$y = \arg\max_{e \in \mathcal{E}} E_e(I), \tag{1}$$

where $\mathcal{E}$ is the emotion vocabulary and $E_e(I)$ is the probability assigned to emotion e. At inference, the same frozen evaluator is evaluated only on the unmodified image to obtain y; the user specifies the target emotion t, and no source-emotion annotation is required.

To assess each candidate region relative to a comparable blur baseline, we compare its blur intervention against a matched reference-mask intervention. Grounding DINO (Liu et al., 2024) and SAM (Kirillov et al., 2023) provide candidate regions indexed by i. Let $I_{i,\setminus\text{blur}}$ denote the intervention that blurs candidate region i, and let $I_{i,\setminus\text{mask}}$ denote its matched reference-mask intervention. Their effects on the source-emotion response are defined as:

$$\Delta_m = E_y(I) - E_y(I_{i,\setminus\text{blur}}),\ \Delta_r = E_y(I) - E_y(I_{i,\setminus\text{mask}}),\ c_i = \Delta_m - \mathbb{E}[\Delta_r]. \tag{2}$$

The correction estimates the candidate's effect relative to the average response to comparable interventions. Here, $E_y(\cdot)$ denotes the frozen evaluator response for the inferred source emotion y, whereas $\mathbb{E}[\Delta_r]$ averages the effects of area- and position-matched reference interventions. We retain Δ_m because averaging it would remove the region-specific signal that c_i is intended to preserve. The resulting corrected contribution c_i is the training target: positive and negative values denote effects above and below the reference baseline, respectively.

The resulting c_i provides the direct supervision interface for the predictor in Section 3.3. Repeating the intervention procedure over candidate regions yields one signed training instance per region.

3.3 Context-Conditioned Signed Contribution Prediction

To estimate c_i without interventions, we train a context-conditioned predictor that operates on a single unmodified image. Built on the SigLIP2 vision backbone (Tschannen et al., 2025), the predictor integrates region-conditioned visual features, the inferred source emotion, and geometric metadata to recover the signed contribution target constructed in Section 3.2.

We model the contextual dependence of regional affect by constructing three visual views at both global and local scales: the candidate region, its complementary context, and the complete image. At each scale $s \in \{g, \ell\}$, the backbone encodes these views as $\mathbf{r}_i^s$, $\mathbf{k}_i^s$, and $\mathbf{a}^s$, respectively. We fuse the three representations with candidate–context difference and product features:

$$\begin{aligned} \mathbf{v}_i^s &= [\mathbf{r}_i^s; \mathbf{k}_i^s; \mathbf{a}^s; \mathbf{r}_i^s - \mathbf{k}_i^s; \mathbf{r}_i^s \odot \mathbf{k}_i^s], \\ \mathbf{h}_i &= [\mathbf{v}_i^g; \mathbf{v}_i^\ell; \mathbf{e}_i; \mathrm{Emb}(y); \mathbf{u}_i; \mathbf{g}_i]. \end{aligned} \tag{3}$$

Here, $\mathbf{e}_i$ encodes region semantics, $\mathbf{u}_i$ summarizes the evaluator state before intervention, and $\mathbf{g}_i$ contains geometric and quality metadata. From the resulting representation $\mathbf{h}_i$, the prediction head estimates a nonnegative contribution magnitude and probabilities over negative, neutral, and positive signs, which are combined into a signed prediction:

$$\widehat{c}_i = m_i \left(p_i^+ - p_i^-\right), \tag{4}$$

where $m_i = \mathrm{softplus}(z_i)$, and p_i^+ and p_i^- are the positive and negative probabilities; the neutral class contributes zero. The predictor must recover not only the value of c_i, but also its direction and the relative importance of regions within the same image. We therefore optimize

$$\mathcal{L} = \mathcal{L}_{\text{Huber}}^w + 0.25\mathcal{L}_{\text{abs}} + 0.10\mathcal{L}_{\text{signed}} + 0.20\mathcal{L}_{\text{focal}} + 0.10\mathcal{L}_{\text{aux}}. \tag{5}$$

The training objective jointly constrains contribution magnitude, within-image ranking, sign, and affective auxiliaries. The weighted Huber term learns contribution magnitude, the absolute-ranking term preserves regional importance, and the signed-ranking and focal terms supervise direction. The auxiliary term retains emotion-intensity and VAD information. Appendix J provides loss definitions and weighting details. Once trained, the predictor directly yields signed regional evidence without counterfactual interventions.

3.4 Attribution-Grounded Target Editing

Upon completing training, we apply the predictor to candidate regions identified by Region Discovery in the source image I. The resulting signed scores $\widehat{c}_i$ are mapped onto their regions to form an interpretable attribution heatmap, showing which content supports or opposes the inferred source emotion y. These predictions provide the attribution grounding for subsequent target editing.



Figure 3: Representative emotion attribution maps on diverse scenes. Existing approaches mislocalize affective evidence or omit opposing regions. IECLA produces more accurate heatmaps.

The target-aware agent translates the signed regional evidence and target emotion t into an executable, image-specific editing plan. It verifies the attribution against the visible content, identifies the primary emotional carrier, and selects the least disruptive operations. The primary local action is anchored to attributed source-emotion evidence, while optional atmosphere or minor-object actions are included only when the scene supports; negative evidence imposes preservation constraints.

Finally, a visual and structural critic checks the plan for executability and content preservation. A deterministic compiler assembles the approved operations into an editing prompt, and an integrated-instruction auditor checks it for conflicts before it is passed to the editor. Detailed routing and validation procedures are provided in Appendices I and J.

4 Experiments

4.1 Experimental Setup

The locked attribution test contains 7,278 images and 53,180 SAM masks. The training, validation, and test subsets are strictly disjoint. Training uses three NVIDIA A800 GPUs. Cross-dataset evaluation excludes the external datasets' original labels from training and teacher-emotion selection. Training details and the eight editing emotions are provided in Appendix J.

Main editing comparisons and ablations cover 680 source images paired with all eight target emotions, yielding 5,440 tasks across Qwen-Edit (Wu et al., 2025a), InstructPix2Pix (Brooks et al., 2023), FLUX.2 (Black Forest Labs, 2026), LongCat (Meituan LongCat Team et al., 2025), OmniGen2 (Wu et al., 2025b), and HiDream (Cai et al., 2025); exact editor variants are listed in Appendix J. We compare each editor with its native pipeline using matched source images, targets, seeds, samplers, and resolutions. The teacher E provides source-emotion inference and training supervision, while the contribution predictor and agent perform candidate scoring and editing planning. Target Success and Δp measure teacher-aligned responses under E, whereas Emo-A and Emo-S provide a second classifier view; none of these post-hoc scores is used for planning. Attribution ranking and localization metrics are computed per image and averaged; Sign Macro-F1 and continuous contribution errors are computed over all valid candidate regions. Confidence intervals use image-level cluster bootstrap.

4.2 Attribution Benchmark Results

To evaluate the accuracy of IECLA's source-emotion attribution, we compare it with two baseline families: conventional heatmap attribution methods and signed attribution models built on



Figure 4: Eight-emotion editing with Qwen-Edit, with and without IECLA. IECLA consistently elicits the requested emotion while preserving scene structure across all eight target categories.

Table 1: Comparison with attribution baselines on the locked test set. IECLA achieves the strongest performance across ranking, sign agreement, and contribution regression.

Method	Imp. $\rho\uparrow$	Signed $\rho\uparrow$	Top-1↑	NDCG@5↑	Macro-F1↑	MAE↓	RMSE↓	$R^2\uparrow$
Heatmap attribution methods								
Ablation-CAM (2020)	-0.0452	–	0.2279	0.5877	–	–	–	–
Grad-CAM (2017)	0.0328	–	0.2525	0.6313	–	–	–	–
Direct attention	0.0432	–	0.2376	0.6321	–	–	–	–
Grad-CAM++ (2018)	0.0690	–	0.2567	0.6443	–	–	–	–
RISE-256 (2018)	0.0861	–	0.2521	0.6428	–	–	–	–
ADAPT-TF (2026)	0.1214	–	0.2700	0.6592	–	–	–	–
EigenCAM (2020)	0.1249	–	0.2751	0.6705	–	–	–	–
HiResCAM (2020)	0.1986	–	0.2987	0.7051	–	–	–	–
LayerCAM (2021)	0.2176	–	0.2991	0.7078	–	–	–	–
Vision-encoder contribution methods								
CLIP (2021)	0.3534	0.1715	0.4325	0.7918	0.3273	0.1145	0.2009	0.2036
DINOV3 (2025)	0.3576	0.1692	0.4376	0.7940	0.3387	0.1116	0.2081	0.1458
SigLIP2 (2025)	0.3819	0.2314	0.4445	0.8050	0.3454	0.1020	0.1833	0.3370
IECLA (ours)	**0.5885**	**0.5088**	**0.5808**	**0.8915**	**0.3703**	**0.0939**	**0.1743**	**0.4005**

CLIP (Radford et al., 2021), DINOV3 (Siméoni et al., 2025), and SigLIP2 (Tschannen et al., 2025). The signed models are trained on the same counterfactual contribution targets, whereas the non-negative heatmap methods are evaluated only on unsigned ranking metrics. Table 1 reports region ranking, contribution direction, and magnitude against the frozen evaluator's counterfactual contributions; this benchmark measures teacher-defined fidelity rather than human causal attribution.

Two patterns distinguish contribution learning from generic saliency. The heatmap methods recover some relevant regions, but their nonnegative outputs cannot represent evidence that opposes the source emotion. The learned vision-encoder models improve regional ranking, while IECLA further leads them in ranking, signed agreement, and regression. This joint improvement is important because the teacher scores concentrate near zero: a near-neutral predictor can attain a low MAE without recovering candidate order or direction. IECLA instead captures which regions matter, whether they support or oppose the source emotion, and their relative contribution strength, Figure 3 illustrates the corresponding localization in complex scenes.

4.3 Comparison with Native Editors

To assess transfer across editing backbones, Table 2 compares IECLA-guided editing with native emotion-request pipelines on the 5,440 matched tasks defined in Section 4.1.

Native editors exhibit two complementary failure modes: conservative edits often preserve the source but under-express the requested emotion, whereas stronger changes can alter subject identity or surrounding structure. Table 2 shows that IECLA shifts this trade-off rather than merely increas-



Figure 5: Qualitative comparison across six editing backbones, with and without IECLA. IECLA improves target-emotion expression across diverse editors while preserving source structure.

Table 2: Comparison between native and IECLA-guided editing across six backbones. IECLA consistently improves target-emotion expression and structural preservation.

Affective outcomes	Target Success↑			Emo-A↑			Emo-S↑		
Editor	Native	IECLA-guided	Rel. Δ	Native	IECLA-guided	Rel. Δ	Native	IECLA-guided	Rel. Δ
Qwen-Edit (2025a)	0.2996	0.3579	**+19.5%**	0.2588	0.3081	**+19.0%**	0.1843	0.2173	**+17.9%**
InstructPix2Pix (2023)	0.1906	0.2127	**+11.6%**	0.2329	0.2704	**+16.1%**	0.1103	0.1508	**+36.7%**
FLUX.2 (2026)	0.2283	0.2643	**+15.8%**	0.1465	0.1522	**+3.9%**	0.0867	0.1508	**+73.9%**
LongCat (2025)	0.2774	0.3099	**+11.7%**	0.1801	0.1844	**+2.4%**	0.0976	0.1053	**+7.9%**
OmniGen2 (2025b)	0.2752	0.3182	**+15.6%**	0.2285	0.2877	**+25.9%**	0.1229	0.1811	**+47.4%**
HiDream (2025)	0.2634	0.2947	**+11.8%**	0.2259	0.2461	**+8.9%**	0.0735	0.1130	**+53.7%**
Structural preservation	PSNR↑			SSIM↑			LPIPS↓		
Editor	Native	IECLA-guided	Rel. Δ	Native	IECLA-guided	Rel. Δ	Native	IECLA-guided	Rel. Δ
Qwen-Edit (2025a)	15.4985	20.3046	**+31.0%**	0.5790	0.6137	**+6.0%**	0.3698	0.3440	**-7.0%**
InstructPix2Pix (2023)	23.4641	27.7156	**+18.1%**	0.6017	0.6682	**+11.1%**	0.3890	0.1840	**-52.7%**
FLUX.2 (2026)	30.8471	31.9981	**+3.7%**	0.4166	0.4346	**+4.3%**	0.4190	0.3249	**-22.5%**
LongCat (2025)	16.0436	16.6780	**+4.0%**	0.4555	0.4876	**+7.0%**	0.3224	0.2923	**-9.3%**
OmniGen2 (2025b)	14.3197	17.4542	**+21.9%**	0.4854	0.5966	**+22.9%**	0.2881	0.1820	**-36.8%**
HiDream (2025)	14.7940	19.9107	**+34.6%**	0.4435	0.5366	**+21.0%**	0.2893	0.1758	**-39.2%**

ing edit strength. Across all six editors, affective outcomes improve together with higher PSNR and SSIM and lower LPIPS. The common direction across editors with different native behavior supports attribution and target-aware planning as a transferable control interface, while the variation in gain magnitude reflects the execution capacity of each backbone. Appendix G examines whether the gains persist across seeds and resolutions.

Figures 1, 4, and 5 illustrate the same trend in multi-object and context-dependent scenes: IECLA directs edits toward attributed cues while constraining unrelated content, avoiding the subject drift or scene-wide changes visible in several native outputs. Appendix A provides further Qwen-Edit examples and analyzes the corresponding failure modes in Figures 8 and 9.

4.4 Cross-Dataset Evaluation

To test transfer beyond the training distribution, we evaluate IECLA on 3,370 images from five external datasets: COCO (Lin et al., 2014), EmoSet (Yang et al., 2023), FI (You et al., 2016), Emotion6 (Peng et al., 2015), and EmoArt (Zhang et al., 2025). We then test a downstream editing task on Qwen-Edit, comparing its native pipeline with IECLA-guided editing.

Across the five external datasets, IECLA remains the leading attribution predictor (Appendix C) and improves downstream editing across natural scenes, affective photographs, and artwork (Table 3). All three preservation metrics improve on every dataset. EmoArt shows smaller structural gains, consistent with artistic and abstract images distributing affective cues across palette, texture, and composition rather than a localized region. Even in this domain, the positive gains show that stronger target-emotion expression does not sacrifice structural preservation.

Table 3: Cross-dataset editing with Qwen-Edit. IECLA improves affective expression and structural preservation across all five datasets.

Affective outcomes	Target Success ↑			Emo-A ↑			Emo-S ↑		
Dataset	Native	IECLA-guided	Rel. Δ	Native	IECLA-guided	Rel. Δ	Native	IECLA-guided	Rel. Δ
COCO	0.2188	0.2939	**+34.3%**	0.1687	0.2375	**+40.8%**	0.0394	0.1015	**+157.6%**
EmoArt	0.2757	0.3664	**+32.9%**	0.1627	0.2397	**+47.3%**	0.0425	0.1120	**+163.5%**
EmoSet	0.2642	0.3396	**+28.5%**	0.1949	0.2642	**+35.6%**	0.0685	0.1114	**+62.6%**
Emotion6	0.1527	0.2611	**+71.0%**	0.1587	0.2168	**+36.6%**	0.0313	0.1065	**+240.3%**
FI	0.2704	0.3082	**+14.0%**	0.1761	0.2390	**+35.7%**	0.0593	0.0937	**+58.0%**
Structural preservation	PSNR ↑			SSIM ↑			LPIPS ↓		
Dataset	Native	IECLA-guided	Rel. Δ	Native	IECLA-guided	Rel. Δ	Native	IECLA-guided	Rel. Δ
COCO	12.8522	16.3243	**+27.0%**	0.2761	0.3918	**+41.9%**	0.4932	0.2427	**-50.8%**
EmoArt	14.4890	15.7233	**+8.5%**	0.3429	0.3659	**+6.7%**	0.2339	0.2219	**-5.1%**
EmoSet	12.7428	17.0118	**+33.5%**	0.3237	0.4595	**+42.0%**	0.4201	0.2181	**-48.1%**
Emotion6	13.5218	17.4566	**+29.1%**	0.3944	0.5366	**+36.1%**	0.3534	0.1719	**-51.4%**
FI	14.2610	17.6349	**+23.7%**	0.4066	0.5093	**+25.3%**	0.3389	0.1675	**-50.6%**

Table 4: Ablation of IECLA's attribution, learning, and agent modules.

Condition	*Target-emotion outcomes*				*Structural preservation*					
	Success ↑	Δp ↑	Emo-A ↑	Emo-S ↑	CLIP-I ↑	G. LPIPS ↓	G. SSIM ↑	O. LPIPS ↓	O. SSIM ↑	O. MAE ↓
Qwen-Edit	0.2996	0.2745	0.2588	0.1843	0.8492	0.3698	0.5790	0.4106	0.3049	0.1919
Qwen-Edit + IECLA (ours)	**0.3579**	**0.3344**	**0.3081**	**0.2173**	**0.8527**	**0.3440**	**0.6137**	**0.3767**	**0.3230**	**0.1733**
-w/o attribution	0.2915	0.2735	0.2454	0.1737	0.8394	0.3513	0.6035	0.3851	0.3184	0.1783
-w/o learning	0.3046	0.2889	0.2724	0.2017	0.8412	0.3454	0.5877	0.4107	0.3147	0.1918
-w/o agent	0.3540	0.3298	0.3024	0.2160	0.8400	0.3543	0.5990	0.3874	0.3171	0.1800

4.5 Ablation Study

Table 4 reports systematic ablations of each core component of the IECLA framework to assess its contribution, with qualitative examples in Figure 6. Appendix F provides additional ablation of editing controls and loss ablations, with the latter summarized in Figure 7(b).

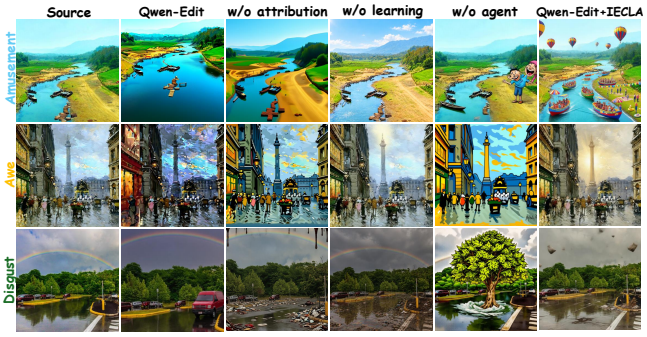


Figure 6: Qualitative ablation on Qwen-Image-Edit. IECLA better expresses target emotions while preserving structure than the baseline and component-removed variants.

Qwen-Edit serves as the native baseline. The ablations follow the three modules in Fig. 2: attribution in (a), contribution learning in (b), and target-aware execution in (c). IECLA achieves the strongest overall profile, leading affective metrics and structural measures. Removing attribution produces the largest affective degradation, confirming that source-emotion evidence is the primary signal for target-oriented editing. When the learning module is removed, the attribution output is passed directly from (a) to (c); the resulting decrease across all affective metrics shows that the learned contribution representation provides more useful editing evidence than direct transfer. Removing the agent while retaining attribution yields the strongest affective scores among the ablations, but still lowers every affective metric, indicating that the evidence must be converted into a target-aware plan to realize the full benefit of the framework.

The loss-level ablation further clarifies how the training objective maintains this balance. Focal-sign supervision has the largest affective impact, highlighting the importance of distinguishing supportive from opposing evidence. The ranking terms preserve regional importance and direction, while auxiliary supervision reinforces affective semantics. Although all ablations achieve slightly lower O. LPIPS, they degrade the other preservation measures and affective outcomes. Overall, the complete objective provides the best balance between target-emotion control and structural fidelity.

4.6 Attribution Validation

Attribution validation assesses whether our predicted regional contributions align with independent external annotations. We evaluate three external VLM references on the same 289 images and 1,032

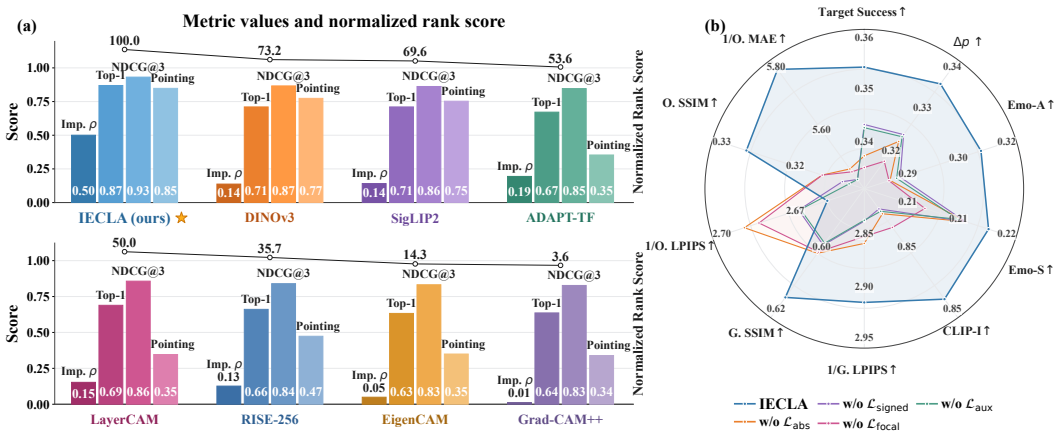


Figure 7: Attribution validation and loss-function ablation. (a) Metric values and normalized rank scores for representative attribution methods. (b) Ablation of the training objective, where IECLA provides the strongest overall affective and structural profile.

aggregated SAM regions, with method outputs and aggregation fixed while varying only the reference labels. An independent Human Review provides a complementary perceptual evaluation that is excluded from training and model selection. Figure 7(a) visualizes the Human Review comparison, and Appendix D reports the complete protocol and results.

The same profile appears under all three VLM references: IECLA leads in regional ordering, signed agreement, top-region selection, and localization despite varying competitors. This stability indicates that the result is not an isolated match to one external evaluator. Human Review provides a complementary perceptual test, where IECLA also aligns more closely with both the location and direction of affective evidence. In particular, agreement on direction supports negative contributions as meaningful opposing evidence rather than a by-product of the shared counterfactual teacher.

4.7 User Study

A blinded User Study examines whether attribution-guided planning produces perceptually stronger edits without sacrificing visual naturalness. Forty independent reviewers score 120 paired editing tasks across six editors for target expression and naturalness, while 20 additional heatmap tasks assess the perceptual quality of the underlying attribution maps.

Across all six editors, IECLA-guided outputs receive higher mean ratings for both target expression and naturalness. This consistent joint preference indicates that stronger affect does not come at the expense of perceptual naturalness. Together with the comparison in Table 2, the common direction across editors supports attribution-guided planning as a transferable editing signal, while differences in gain magnitude reflect how effectively each backbone realizes the localized guidance. The accompanying heatmap evaluation provides complementary perceptual evidence for the readability and focus of the underlying attribution. Overall, the User Study primarily supports the downstream editing utility of IECLA; the complete protocol and detailed results are reported in Appendix E.

5 Conclusion

We presented IECLA, a framework for affective image editing that separates source-emotion attribution from target-specific execution. Its counterfactual contribution predictor localizes signed object-level evidence: positive regions provide an editable focus, while negative regions provide preservation evidence. A target-aware agent then translates this evidence and the desired emotion into localized, content-aware editing operations with explicit preservation constraints. Experiments on the locked attribution benchmark and five external datasets, together with reference validation, show consistent attribution quality across evaluation sources and domains. Across six editing backbones, matched comparisons, ablations, qualitative results, and the User Study further demonstrate improved target-emotion expression with stronger structural preservation. These results establish affective attribution as both an explanation of perceived emotion and an actionable interface for controlling general image editors.

AI USE STATEMENT

Generative AI tools supported the development of the conceptual framework, refinement of the experimental methodology, and implementation of experimental code. They were also used to synthesize and analyze related literature, organize references, and improve the clarity and readability of the manuscript. The authors reviewed the resulting text, code, and references and retain full responsibility for all claims, experimental results, and AI-assisted material in the final manuscript.

ETHICS STATEMENT

This work uses existing image datasets and pretrained models for affective image editing. Human annotation and the user study were conducted with participants' informed consent; their protocols are described in Appendices D.2 and E. Emotion labels describe perceived affect rather than a depicted individual's internal state. The framework supports creative expression but shares the potential risks of generative image editing, including misleading portrayals and emotional manipulation. Responsible use should respect the privacy of depicted individuals and applicable dataset and model licenses, and disclose image manipulation where authenticity matters.

REPRODUCIBILITY STATEMENT

Section 3 describes the counterfactual supervision, contribution predictor, and target-aware editing framework, while Section 4.1 summarizes the experimental setup. Appendix J documents data provenance and disjoint splits, evaluation metrics, reference calibration, training objectives, baseline implementations, and agent and editor configurations. Appendix I provides the structured editing-agent template. Appendices D and E describe the external-reference and human-evaluation protocols, while Appendices F and G provide additional ablations and stability analyses. Code will be released with the final version of the paper. Exact numerical results may vary with stochastic generation and differences in the software and model-serving environment.

REFERENCES

Panos Achlioptas, Maks Ovsjanikov, Kilichbek Haydarov, Mohamed Elhoseiny, and Leonidas Guibas. Artemis: Affective language for visual art. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11564–11574, 2021. doi: 10.1109/CVPR46437.2021.01140.

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. *arXiv preprint arXiv:1810.03292*, 2018. doi: 10.48550/arXiv.1810.03292.

Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18187–18197, 2022. doi: 10.1109/CVPR52688.2022.01767.

Black Forest Labs. FLUX.2 [klein]: Fast, efficient image generation. `https://bfl.ai/models/flux-2-klein`, 2026.

Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18392–18403, 2023. doi: 10.1109/CVPR52729.2023.01764.

Qi Cai, Jingwen Chen, Yang Chen, Yehao Li, Fuchen Long, Yingwei Pan, Zhaofan Qiu, Yiheng Zhang, Fengbin Gao, Peihan Xu, et al. HiDream-I1: A high-efficient image generative foundation model with sparse diffusion transformer. *arXiv preprint arXiv:2505.22705*, 2025. doi: 10.48550/arXiv.2505.22705.

Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N. Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE Winter Conference on Applications of Computer Vision*, pp. 839–847, 2018. doi: 10.1109/WACV.2018.00097.

Saurabh Desai and Harish G. Ramaswamy. Ablation-cam: Visual explanations for deep convolutional network via gradient-free localization. In *2020 IEEE Winter Conference on Applications of Computer Vision*, pp. 972–980, 2020. doi: 10.1109/WACV45572.2020.9093360.

Youssef Doulfoukar, Laurent Mertens, and Joost Vennekens. Emocam: Toward understanding what drives CNN-based emotion recognition. *arXiv preprint arXiv:2407.14314*, 2024. doi: 10.48550/arXiv.2407.14314.

Rachel Lea Draelos and Lawrence Carin. Use hirescam instead of grad-cam for faithful explanations of convolutional neural networks. *arXiv preprint arXiv:2011.08891*, 2020.

Ruth C. Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3449–3457, 2017. doi: 10.1109/ICCV.2017.371.

Tsu-Jui Fu, Xin Eric Wang, and William Yang Wang. Language-driven artistic style transfer. In *European Conference on Computer Vision*, pp. 717–734, 2022.

Tsu-Jui Fu, Wenze Hu, Xianzhi Du, William Wang, Yinfei Yang, and Zhe Gan. Guiding instruction-based image editing via multimodal large language models. In *International Conference on Learning Representations*, volume 2024, pp. 54820–54833, 2024.

Zigang Geng, Binxin Yang, Tiankai Hang, Chen Li, Shuyang Gu, Ting Zhang, Jianmin Bao, Zheng Zhang, Han Hu, Dong Chen, and Baining Guo. Instructdiffusion: A generalist modeling interface for vision tasks. *arXiv preprint arXiv:2309.03895*, 2023. doi: 10.48550/arXiv.2309.03895.

Yash Goyal, Ziyan Wu, Jan Ernst, Dhruv Batra, Devi Parikh, and Stefan Lee. Counterfactual visual explanations. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2376–2384. PMLR, 2019.

Yijie Guo, Dexiang Hong, Weidong Chen, Zihan She, Cheng Ye, Xiaojun Chang, and Zhendong Mao. Emoverse: A mllms-driven emotion representation dataset for interpretable visual emotion analysis. *arXiv preprint arXiv:2511.12554*, 2025.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2208.01626.

Peng-Tao Jiang, Chang-Bin Zhang, Qibin Hou, Ming-Ming Cheng, and Yunchao Wei. Layercam: Exploring hierarchical class activation maps for localization. *IEEE Transactions on Image Processing*, 30:5875–5888, 2021. doi: 10.1109/TIP.2021.3089943.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4015–4026, 2023. doi: 10.1109/ICCV51070.2023.00371.

Ronak Kosti, Jose M. Alvarez, Adria Recasens, and Agata Lapedriza. Context based emotion recognition using EMOTIC dataset. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(11):2755–2766, 2020. doi: 10.1109/TPAMI.2019.2916866.

Dongxu Li, Junnan Li, and Steven C. H. Hoi. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing, 2023. URL https://arxiv.org/abs/2305.14720.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision*, pp. 740–755, 2014. doi: 10.1007/978-3-319-10602-1_48.

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. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, pp. 38–55, 2024. doi: 10.1007/978-3-031-72970-6_3.

Fanhang Man, Xiaoyue Chen, Huandong Wang, Baining Zhao, Han Li, and Xinlei Chen. VAEER: Visual attention-inspired emotion elicitation reasoning. *arXiv preprint arXiv:2505.24342*, 2025. doi: 10.48550/arXiv.2505.24342.

Qi Mao, Haobo Hu, Yujie He, Difei Gao, Haokun Chen, and Libiao Jin. Emoagent: A multi-agent framework for diverse affective image manipulation. *arXiv preprint arXiv:2503.11290*, 2025. doi: 10.48550/arXiv.2503.11290.

Meituan LongCat Team, Hanghang Ma, Haoxian Tan, Jiale Huang, Junqiang Wu, Jun-Yan He, Lishuai Gao, Songlin Xiao, Xiaoming Wei, Xiaoqi Ma, Xunliang Cai, Yayong Guan, and Jie Hu. LongCat-Image technical report. *arXiv preprint arXiv:2512.07584*, 2025. doi: 10.48550/arXiv.2512.07584.

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. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2108.01073.

Mohammed Bany Muhammad and Mohammed Yeasin. Eigen-cam: Class activation map using principal components. In *2020 International Joint Conference on Neural Networks*, pp. 1–7, 2020. doi: 10.1109/IJCNN48605.2020.9206626.

Kuan-Chuan Peng, Tsuhan Chen, Amir Sadovnik, and Andrew Gallagher. A mixed bag of emotions: Model, predict, and transfer emotion distributions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 860–868, 2015. doi: 10.1109/CVPR.2015.7298687.

Vitali Petsiuk, Abir Das, and Kate Saenko. RISE: Randomized input sampling for explanation of black-box models. In *British Machine Vision Conference 2018*, pp. 151. BMVA Press, 2018.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pp. 97–101, 2016. doi: 10.18653/v1/N16-3020.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 618–626, 2017. doi: 10.1109/ICCV.2017.74.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025. doi: 10.48550/arXiv.2508.10104.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3319–3328. PMLR, 2017.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025. doi: 10.48550/arXiv.2502.14786.

Cong Wei, Zheyang Xiong, Weiming Ren, Xeron Du, Ge Zhang, and Wenhu Chen. Omniedit: Building image editing generalist models through specialist supervision. In *International Conference on Learning Representations*, volume 2025, pp. 259–271, 2025.

Shuchen Weng, Peixuan Zhang, Zheng Chang, Xinlong Wang, Si Li, and Boxin Shi. Affective image filter: Reflecting emotions from text to images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10810–10819, 2023.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-Image technical report. *arXiv preprint arXiv:2508.02324*, 2025a. doi: 10.48550/arXiv.2508.02324.

Chenyuan Wu, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, Junjie Zhou, et al. OmniGen2: Towards instruction-aligned multimodal generation. *arXiv preprint arXiv:2506.18871*, 2025b. doi: 10.48550/arXiv.2506.18871.

Jingyuan Yang, Qirui Huang, Tingting Ding, Dani Lischinski, Daniel 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*, pp. 20326–20337, 2023. doi: 10.1109/ICCV51070.2023.01864.

Jingyuan Yang, Jiawei Feng, Weibin Luo, Dani Lischinski, Daniel Cohen-Or, and Hui Huang. Emoedit: Evoking emotions through image manipulation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24690–24699. IEEE, 2025.

Ling Yang, Zhilong Zhang, Zhaochen Yu, Jingwei Liu, Minkai Xu, Stefano Ermon, and Bin Cui. Cross-modal contextualized diffusion models for text-guided visual generation and editing. In *International Conference on Learning Representations*, volume 2024, pp. 13972–13997, 2024.

Zhiyuan Yao, Zheren Fu, Zhixiao Zheng, Jiajun Li, Yi Tu, and Zhendong Mao. ADAPT: Attention dynamics alignment with preference tuning for faithful MLLMs. *arXiv preprint arXiv:2606.31054*, 2026. doi: 10.48550/arXiv.2606.31054.

Jiarong Ye and Sharon X. Huang. Moodifier: MLLM-enhanced emotion-driven image editing. *arXiv preprint arXiv:2507.14024*, 2025. doi: 10.48550/arXiv.2507.14024.

Xinyi Yin, Yiduo Wang, Tingqi Hu, Meicong Si, Yunyun Shi, Shi Chen, Hao Wang, Junxiao Xue, and Xuecheng Wu. Mood: Perception-enhanced efficient affective image editing via continuous valence–arousal modeling. *arXiv preprint arXiv:2605.02521*, 2026. doi: 10.48550/arXiv.2605.02521.

Quanzeng You, Jiebo Luo, Hailin Jin, and Jianchao Yang. Building a large scale dataset for image emotion recognition: The fine print and the benchmark. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2016. doi: 10.1609/aaai.v30i1.9987.

Kaishen Yuan, Yuting Zhang, Shang Gao, Yijie Zhu, Wenshuo Chen, and Yutao Yue. Coemogen: Towards semantically-coherent and scalable emotional image content generation. In *International Conference on Learning Representations*, volume 2026, pp. 6341–6361, 2026.

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*, pp. 12644–12650, 2025. doi: 10.1145/3746027.3758201.

Jing Zhang and Bingjie Fan. Emokgedit: Training-free affective injection via visual cue transformation. *arXiv preprint arXiv:2601.12326*, 2026. doi: 10.48550/arXiv.2601.12326.

Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. Magicbrush: A manually annotated dataset for instruction-guided image editing. In *Advances in Neural Information Processing Systems 36*, pp. 31428–31449, 2023a. doi: 10.52202/075280-1365.

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*, pp. 3813–3824, 2023b. doi: 10.1109/ICCV51070.2023.00355.

Supplementary Material

CONTENTS

A Qualitative Results **15**

B Editing Baselines **17**

C Cross-Dataset Results **17**

D Reference Validation **20**

D.1 VLM Reference 20

D.2 Human Reference 21

E User Study **22**

F Additional Ablations **23**

G Stability Analysis **24**

H Teacher Replay and Predictor Comparison **25**

I Agent Template **26**

J Implementation Details **27**

A QUALITATIVE RESULTS

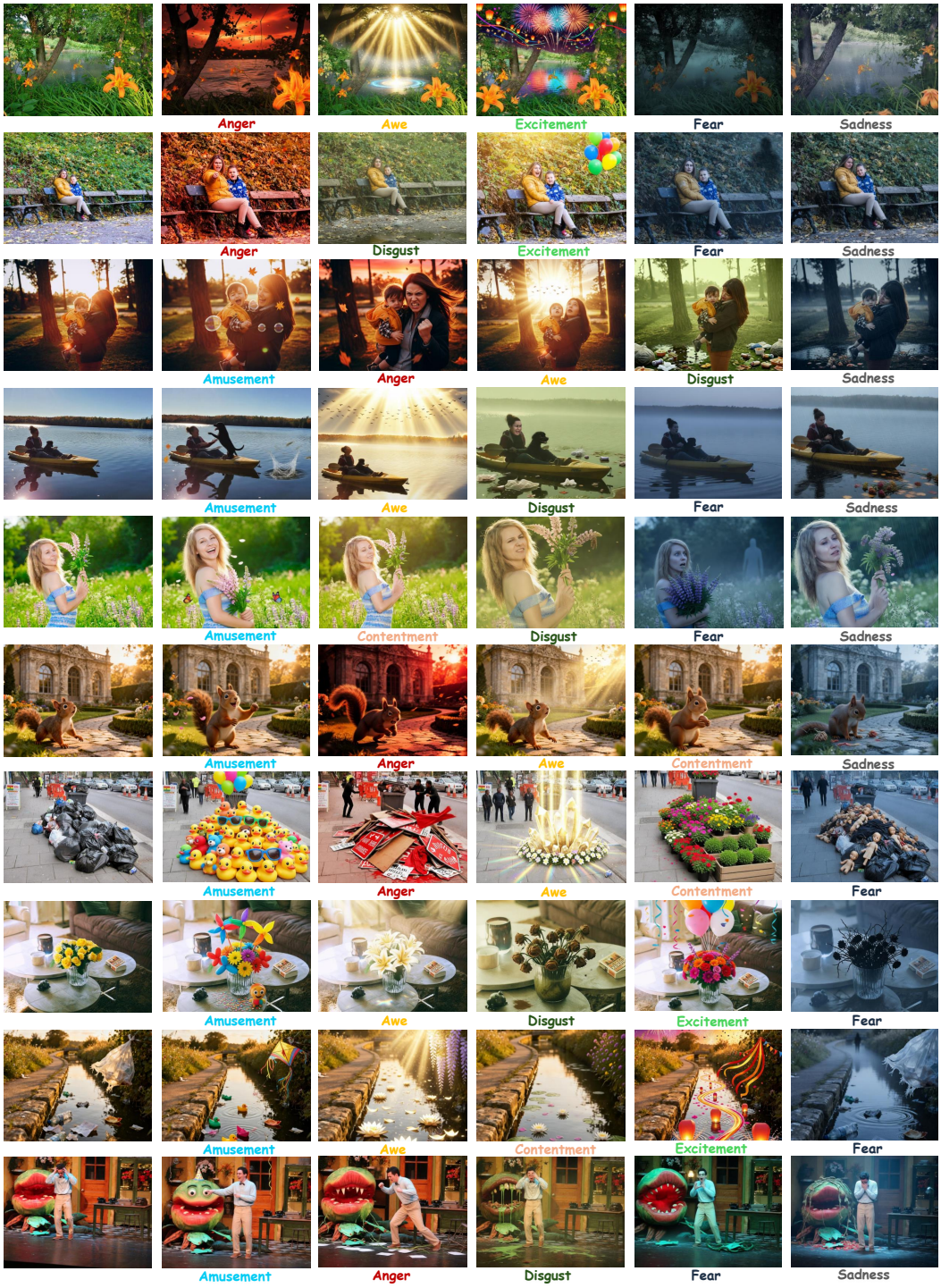


Figure 8: Additional qualitative results with IECLA-guided Qwen-Edit. Each row presents a source image followed by edits toward different target emotions, showing diverse affective changes with consistent source content and composition.

Figure 8 complements the examples in Figure 1 with additional IECLA-guided Qwen-Edit outputs. The examples span people, animals, objects, and scene-centric images. Across these settings, the attributed guidance produces target-specific changes through localized semantic cues and scene appearance while retaining recognizable source content and composition.

Figure 9 further compares native Qwen-Edit with the same editor guided by IECLA. Across varied source content and target emotions, attribution-guided editing introduces clearer emotion-specific cues while better preserving the original subject and scene organization.



Figure 9: Additional comparisons between native Qwen-Edit and IECLA-guided editing. IECLA strengthens target-emotion expression while preserving scene structure across diverse images.

Taken together, the examples expose two recurring failure modes of native editing. When a source image already carries cues close to contentment, the native editor may introduce text or decorative objects even when a small chromatic or illumination adjustment would suffice. IECLA instead uses positive signed counterfactual contributions as editable evidence and negative contributions as preservation evidence, allowing the agent to make a restrained change while retaining the original composition. Several contentment examples follow this pattern: the guided outputs adjust color and light without adding unrelated content.

For targets such as fear, the native editor often darkens the background or adds red illumination as isolated operations. These cues do not necessarily form a coherent relation with the foreground and can leave the perceived emotion weak or ambiguous. IECLA can route evidence across the subject and its surrounding context, so expression, lighting, and background cues are changed jointly when supported by the image. The same pattern appears in face-dominant scenes, where editing only the expression misses contextual evidence, and in artistic images, where IECLA changes palette or atmosphere while preserving composition. These cases illustrate how signed attribution supplies relational editing evidence rather than a generic instruction to modify the most salient region.

B Editing Baselines

This appendix compares IECLA with representative affective and general-purpose image editing methods on the same test set used for the main editing tables. The comparison includes global diffusion editing (SDEdit (Meng et al., 2022)), instruction-based editing (InsDiff, i.e., InstructDiffusion (Geng et al., 2023)), spatial control (ControlNet (Zhang et al., 2023b)), subject-preserving editing (BLIP-Diffusion (Li et al., 2023)), and affective style transfer (CLVA (Fu et al., 2022) and AIF (Weng et al., 2023)), together with EmoEdit (Yang et al., 2025). All methods are evaluated on the 680 source images and 5,440 source–target tasks used in the main editing comparison; the Qwen-Edit + IECLA row uses the same IECLA-guided Qwen-Edit configuration as the main tables. Since absolute scores depend on the editing backbone, this single-backbone comparison is reported here, while Table 2 presents matched native-versus-guided results across six editors.

Table 5 shows the tension between emotion control and structural preservation across prior approaches. Qwen-Edit + IECLA leads in five of six evaluations, indicating stronger target-emotion expression alongside close agreement with the source image. EmoEdit ranks second on LPIPS, while AIF leads in CLIP-I; the latter metric is also high for CLVA despite its low emotion scores. Thus, high image similarity alone does not imply effective emotion control, whereas IECLA performs well on both affective and preservation measures.

Table 5: Comparison with affective and general-purpose editing baselines on the main editing test set. IECLA provides the strongest joint emotion improvement and structure retention.

Method	PSNR↑	SSIM↑	LPIPS↓	CLIP-I↑	Emo-A↑	Emo-S↑
SDEdit (Meng et al., 2022)	14.1547	0.3820	0.3841	0.7481	0.2050	0.0929
InsDiff (Geng et al., 2023)	9.8607	0.3870	0.3999	0.7388	0.0551	0.0537
ControlNet (Zhang et al., 2023b)	12.2646	0.2370	0.4893	0.6768	0.1860	0.1659
BLIP-Diffusion (Li et al., 2023)	8.7242	0.3260	0.5514	0.8169	0.0676	0.0190
CLVA (Fu et al., 2022)	12.1975	0.4960	0.3963	<u>0.8572</u>	0.0379	0.0138
AIF (Weng et al., 2023)	16.1320	0.6080	0.4147	**0.9932**	0.0812	0.0345
EmoEdit (Yang et al., 2025)	<u>18.1564</u>	<u>0.6090</u>	<u>0.3459</u>	0.8413	<u>0.2647</u>	<u>0.2149</u>
Qwen-Edit + IECLA (ours)	**20.3046**	**0.6137**	**0.3440**	0.8527	**0.3081**	**0.2173**

C Cross-Dataset Results

The five external datasets are examined separately under the attribution protocol summarized in Table 11. All tables use the same frozen counterfactual teacher and Grounding DINO and SAM candidate regions. Ranking and localization columns use image-macro aggregation, whereas Sign Macro-F1 and regression columns pool all valid candidate regions; images without valid regions are removed independently within each dataset. Bold and underlined entries mark the best and second-best results, while "–" denotes metrics undefined for nonnegative saliency outputs.

Table 6: Cross-dataset attribution results on COCO (612 images, 5,008 masks).

Method	Images	Masks	Imp. $\rho \uparrow$	Signed $\rho \uparrow$	Top-1↑	NDCG@3↑	Macro-F1↑	MAE↓	RMSE↓	$R^2 \uparrow$
Heatmap attribution methods										
Ablation-CAM	612	5008	-0.0632	–	0.1765	0.5035	–	–	–	–
Grad-CAM	612	5008	0.0740	–	0.2320	0.5814	–	–	–	–
Direct attention	612	5008	0.0507	–	0.2141	0.5541	–	–	–	–
Grad-CAM++	612	5008	0.0697	–	0.2190	0.5737	–	–	–	–
RISE-256	612	5008	0.1688	–	0.2467	0.5986	–	–	–	–
ADAPT-TF	612	5008	-0.0061	–	0.1912	0.5125	–	–	–	–
EigenCAM	612	5008	0.1102	–	0.2533	0.5950	–	–	–	–
HiResCAM	612	5008	0.1840	–	0.2565	0.6413	–	–	–	–
LayerCAM	612	5008	0.2162	–	0.2843	0.6554	–	–	–	–
Vision-encoder contribution methods										
CLIP	612	5008	0.3549	0.1995	0.4444	0.7530	0.4669	0.1617	0.2458	0.0102
DINOV3	612	5008	0.3853	0.1977	0.4248	0.7588	0.4950	0.1459	0.2394	0.0610
SigLIP2	612	5008	0.4038	0.3008	0.4641	0.7800	0.5442	0.1306	0.2078	0.2924
IECLA (ours)	612	5008	**0.5833**	**0.5558**	**0.5735**	**0.8656**	**0.7674**	**0.0757**	**0.1545**	**0.6087**

Table 7: Cross-dataset attribution results on EmoSet (915 images, 6,304 masks).

Method	Images	Masks	Imp. $\rho \uparrow$	Signed $\rho \uparrow$	Top-1↑	NDCG@3↑	Macro-F1↑	MAE↓	RMSE↓	$R^2 \uparrow$
Heatmap attribution methods										
Ablation-CAM	915	6304	-0.0195	–	0.2634	0.5877	–	–	–	–
Grad-CAM	915	6304	0.0089	–	0.2710	0.6188	–	–	–	–
Direct attention	915	6304	0.0395	–	0.2667	0.6307	–	–	–	–
Grad-CAM++	915	6304	0.0572	–	0.2831	0.6345	–	–	–	–
RISE-256	915	6304	0.1582	–	0.2962	0.6701	–	–	–	–
ADAPT-TF	915	6304	0.0659	–	0.2842	0.6260	–	–	–	–
EigenCAM	915	6304	0.1197	–	0.3213	0.6761	–	–	–	–
HiResCAM	915	6304	0.1766	–	0.3126	0.6955	–	–	–	–
LayerCAM	915	6304	0.2472	–	0.3344	0.7296	–	–	–	–
Vision-encoder contribution methods										
DINOV3	915	6304	0.3088	0.2190	0.4667	0.7931	0.5131	0.1536	0.2440	0.1647
CLIP	915	6304	0.3214	0.2159	0.4710	0.8022	0.4856	0.1625	0.2482	0.1358
SigLIP2	915	6304	0.3406	0.2920	0.5104	0.8155	0.5512	0.1394	0.2179	0.3340
IECLA (ours)	915	6304	**0.4804**	**0.4519**	**0.5530**	**0.8636**	**0.7197**	**0.1034**	**0.2025**	**0.4247**

Table 8: Cross-dataset attribution results on FI (1,544 images, 9,966 masks).

Method	Images	Masks	Imp. $\rho \uparrow$	Signed $\rho \uparrow$	Top-1↑	NDCG@3↑	Macro-F1↑	MAE↓	RMSE↓	$R^2 \uparrow$
Heatmap attribution methods										
Ablation-CAM	1544	9966	0.0103	–	0.3394	0.6174	–	–	–	–
Grad-CAM	1544	9966	0.0023	–	0.3342	0.6281	–	–	–	–
Direct attention	1544	9966	0.0418	–	0.3154	0.6378	–	–	–	–
Grad-CAM++	1544	9966	0.0599	–	0.3549	0.6463	–	–	–	–
RISE-256	1544	9966	0.0925	–	0.3472	0.6604	–	–	–	–
ADAPT-TF	1544	9966	0.0182	–	0.3187	0.6253	–	–	–	–
EigenCAM	1544	9966	0.0972	–	0.3659	0.6806	–	–	–	–
HiResCAM	1544	9966	0.1519	–	0.3763	0.6963	–	–	–	–
LayerCAM	1544	9966	0.1874	–	0.3789	0.6995	–	–	–	–
Vision-encoder contribution methods										
DINOV3	1544	9966	0.2733	0.1345	0.4644	0.7826	0.5017	0.1505	0.2486	0.1187
CLIP	1544	9966	0.2875	0.1157	0.4903	0.7887	0.4662	0.1588	0.2514	0.0983
SigLIP2	1544	9966	0.3002	0.1869	0.4961	0.7956	0.5193	0.1373	0.2191	0.3155
IECLA (ours)	1544	9966	**0.4770**	**0.4715**	**0.6101**	**0.8825**	**0.7884**	**0.0746**	**0.1675**	**0.5999**

Table 9: Cross-dataset attribution results on Emotion6 (226 images, 1,335 masks).

Method	Images	Masks	Imp. $\rho \uparrow$	Signed $\rho \uparrow$	Top-1↑	NDCG@3↑	Macro-F1↑	MAE↓	RMSE↓	$R^2 \uparrow$
Heatmap attribution methods										
Ablation-CAM	226	1335	0.0104	–	0.3584	0.6734	–	–	–	–
Grad-CAM	226	1335	0.0531	–	0.3496	0.6925	–	–	–	–
Direct attention	226	1335	-0.1231	–	0.2788	0.6421	–	–	–	–
Grad-CAM++	226	1335	0.1207	–	0.3274	0.6914	–	–	–	–
RISE-256	226	1335	0.0549	–	0.3673	0.6960	–	–	–	–
ADAPT-TF	226	1335	0.0520	–	0.3496	0.6353	–	–	–	–
EigenCAM	226	1335	0.1091	–	0.3894	0.7102	–	–	–	–
HiResCAM	226	1335	0.1617	–	0.3496	0.7149	–	–	–	–
LayerCAM	226	1335	0.1476	–	0.3673	0.7192	–	–	–	–
Vision-encoder contribution methods										
CLIP	226	1335	0.2581	0.1617	0.4735	0.8059	0.4561	0.2005	0.3051	0.0892
DINOV3	226	1335	0.3458	0.2180	0.4779	0.8366	0.4731	0.1950	0.3122	0.0463
SigLIP2	226	1335	0.3488	0.2283	0.5398	0.8387	0.4982	0.1775	0.2764	0.2523
IECLA (ours)	226	1335	**0.5188**	**0.4822**	**0.6106**	**0.9002**	**0.7400**	**0.1107**	**0.2324**	**0.4716**

Table 10: Cross-dataset attribution results on EmoArt (73 images, 533 masks).

Method	Images	Masks	Imp. $\rho \uparrow$	Signed $\rho \uparrow$	Top-1↑	NDCG@3↑	Macro-F1↑	MAE↓	RMSE↓	$R^2 \uparrow$
Heatmap attribution methods										
Ablation-CAM	73	533	-0.1059	–	0.1781	0.5470	–	–	–	–
Grad-CAM	73	533	0.0319	–	0.2055	0.6264	–	–	–	–
Direct attention	73	533	-0.0193	–	0.1507	0.5811	–	–	–	–
Grad-CAM++	73	533	0.0793	–	0.1918	0.6302	–	–	–	–
RISE-256	73	533	0.1492	–	0.2192	0.6854	–	–	–	–
ADAPT-TF	73	533	-0.0326	–	0.1370	0.5921	–	–	–	–
EigenCAM	73	533	0.0586	–	0.2055	0.6684	–	–	–	–
HiResCAM	73	533	0.1681	–	0.2740	0.6730	–	–	–	–
LayerCAM	73	533	0.0541	–	0.1781	0.6315	–	–	–	–
Vision-encoder contribution methods										
CLIP	73	533	0.3520	0.3301	0.3562	0.7994	0.4815	0.1946	0.2780	0.1517
DINOV3	73	533	0.3616	0.2625	0.3699	0.8284	0.5296	0.1798	0.2729	0.1827
SigLIP2	73	533	0.3858	0.3425	**0.4384**	0.7920	0.5253	0.1768	0.2597	0.2600
IECLA (ours)	73	533	**0.5028**	**0.5103**	**0.4384**	**0.8488**	**0.7069**	**0.1208**	**0.2176**	**0.4801**

Table 11: Pooled cross-dataset attribution results under the shared counterfactual teacher. Ranking metrics are image-weighted, whereas Sign Macro-F1 and regression metrics are pooled over valid regions. IECLA leads all reported metrics.

Method	Imp. $\rho \uparrow$	Signed $\rho \uparrow$	Top-1↑	NDCG@3↑	Macro-F1↑	MAE↓	RMSE↓	$R^2 \uparrow$
Heatmap attribution methods								
Ablation-CAM (2020)	-0.0137	–	0.2869	0.5909	–	–	–	–
Grad-CAM (2017)	0.0212	–	0.2967	0.6214	–	–	–	–
Direct attention	0.0304	–	0.2777	0.6198	–	–	–	–
Grad-CAM++ (2018)	0.0655	–	0.3053	0.6326	–	–	–	–
RISE-256 (2018)	0.1229	–	0.3136	0.6548	–	–	–	–
ADAPT-TF (2026)	0.0279	–	0.2843	0.6050	–	–	–	–
EigenCAM (2020)	0.1056	–	0.3315	0.6656	–	–	–	–
HiResCAM (2020)	0.1654	–	0.3332	0.6869	–	–	–	–
LayerCAM (2021)	0.2033	–	0.3445	0.6995	–	–	–	–
Vision-encoder contribution methods								
CLIP (2021)	0.3083	0.1659	0.4727	0.7873	0.4728	0.1637	0.2534	0.0987
DINOV3 (2025)	0.3101	0.1773	0.4567	0.7858	0.5031	0.1536	0.2501	0.1219
SigLIP2 (2025)	0.3351	0.2423	0.4958	0.8010	0.5331	0.1396	0.2211	0.3137
IECLA (ours)	**0.5006**	**0.4830**	**0.5843**	**0.8748**	**0.7600**	**0.0858**	**0.1806**	**0.5422**

The pooled table is not an equal-weight average of the five datasets: its ranking columns weight datasets by the number of valid images, while its sign and regression columns pool all valid regions.

Across all five datasets, IECLA maintains Importance and Signed Spearman correlations above 0.47 and 0.45, respectively, and ranks first on every signed, list-level, and regression metric. The pooled advantage in Table 11 is therefore not driven by a single source. It persists on FI, the largest external set, and on EmoArt, the smallest and most stylistically distinct set, where IECLA ties SigLIP2 on Top-1 but retains the lead in region ordering, signed agreement, and contribution regression.

The method families exhibit a consistent hierarchy across datasets. Vision-encoder contribution predictors generally improve the available ranking metrics over conventional heatmap attribution

methods, indicating that semantic representations better match the teacher-defined regional evidence. IECLA further improves upon the strongest vision-encoder baseline in both signed criteria and contribution-error measures. This joint gain shows that its transfer extends beyond selecting the most relevant region to estimating whether each region supports or opposes the source emotion and by how much. Together, the results establish consistent agreement with the frozen counterfactual teacher across natural, affective, and artistic image distributions.

D Reference Validation

This appendix collects the complete attribution validation associated with Section 4.6. The three VLM references are evaluated on the same 289 images and the same 1,032 aggregated SAM regions. Method prediction maps, heatmaps, and aggregation rules remain fixed across references, and only the external contribution labels vary. No VLM output is used for training. The two Qwen references provide within-family checks, while GPT Terra supplies a reference from a different model family. Tables 12–14 report image-macro results for each reference, and Table 15 gives their unweighted mean before the independent Human Review comparison.

D.1 VLM Reference

Qwen Chat and Qwen Reasoner both use Qwen3.8 through the official Qwen API, under direct-response and reasoning configurations, respectively. GPT Terra uses GPT-5.6-Terra. These models serve only as external attribution references and do not participate in contribution-predictor training, checkpoint selection, or editing-agent planning.

Table 12: Attribution agreement with the Qwen Chat reference. IECLA leads all five metrics.

Method	Imp. $\rho\uparrow$	Signed $\rho\uparrow$	Top-1↑	NDCG@3↑	Pointing↑
Heatmap attribution methods					
Ablation-CAM	-0.2294	–	0.4014	0.7097	0.2388
Grad-CAM	-0.0161	–	0.4775	0.7731	0.2768
Direct attention	0.0315	–	0.4810	0.7790	0.2111
Grad-CAM++	0.0843	–	0.5363	0.7933	0.2630
EigenCAM	0.0538	–	0.5156	0.7862	0.2699
HiResCAM	0.1158	–	0.5571	0.7929	0.2837
RISE-256	0.0748	–	0.5017	0.7738	0.4394
LayerCAM	0.1835	–	0.5675	0.8144	0.2699
ADAPT-TF	0.1890	–	0.5606	0.8081	0.2941
Vision-encoder contribution methods					
CLIP	0.1107	0.0407	0.5744	0.7936	0.6678
SigLIP2	0.1527	0.1280	0.6159	0.8083	0.6851
DINOV3	0.1507	0.1688	0.6125	0.8083	0.6782
IECLA (ours)	**0.6187**	**0.4868**	**0.8720**	**0.9331**	**0.8685**

Table 13: Qwen Reasoner reference results. IECLA retains its lead across all five metrics.

Method	Imp. $\rho\uparrow$	Signed $\rho\uparrow$	Top-1↑	NDCG@3↑	Pointing↑
Heatmap attribution methods					
Ablation-CAM	-0.2647	–	0.3772	0.6826	0.2249
Grad-CAM	-0.0222	–	0.4913	0.7578	0.2630
Direct attention	0.0340	–	0.4844	0.7722	0.2007
Grad-CAM++	0.0988	–	0.5467	0.7746	0.2491
EigenCAM	0.0193	–	0.5052	0.7712	0.2526
HiResCAM	0.1695	–	0.5848	0.7942	0.2699
RISE-256	0.1335	–	0.5225	0.7734	0.4187
LayerCAM	0.2348	–	0.5813	0.8158	0.2457
ADAPT-TF	0.2291	–	0.5779	0.8021	0.2907
Vision-encoder contribution methods					
CLIP	0.1088	0.0501	0.5640	0.7827	0.6540
SigLIP2	0.1142	0.1370	0.5952	0.7873	0.6574
DINOV3	0.1163	0.1741	0.5882	0.7831	0.6574
IECLA (ours)	**0.5998**	**0.5108**	**0.8789**	**0.9313**	**0.8616**

Table 14: GPT Terra reference results. IECLA retains its lead under a different VLM family.

Method	Imp. $\rho\uparrow$	Signed $\rho\uparrow$	Top-1↑	NDCG@3↑	Pointing↑
Heatmap attribution methods					
Ablation-CAM	-0.1633	–	0.4048	0.7231	0.2388
Grad-CAM	0.0698	–	0.5121	0.7959	0.2803
Direct attention	0.0077	–	0.4221	0.7664	0.2215
Grad-CAM++	0.0863	–	0.5121	0.7971	0.2803
EigenCAM	0.0460	–	0.4706	0.7851	0.2803
HiResCAM	0.1044	–	0.5363	0.7960	0.2941
RISE-256	-0.0277	–	0.4256	0.7524	0.4394
LayerCAM	0.1517	–	0.5225	0.8119	0.2837
ADAPT-TF	0.0418	–	0.4844	0.7832	0.2699
Vision-encoder contribution methods					
CLIP	0.1850	0.0231	0.5882	0.8121	0.6609
SigLIP2	0.2138	0.0705	0.6055	0.8275	0.6644
DINOV3	0.2252	0.0626	0.6159	0.8261	0.6574
IECLA (ours)	**0.6165**	**0.5324**	**0.8789**	**0.9323**	**0.8547**

Table 15: Mean agreement across VLM references. IECLA leads every metric.

Method	Imp. $\rho\uparrow$	Signed $\rho\uparrow$	Top-1↑	NDCG@3↑	Pointing↑
Heatmap attribution methods					
Ablation-CAM	-0.2191	–	0.3945	0.7051	0.2342
Grad-CAM	0.0105	–	0.4936	0.7756	0.2734
Direct attention	0.0244	–	0.4625	0.7725	0.2111
Grad-CAM++	0.0898	–	0.5317	0.7883	0.2641
EigenCAM	0.0397	–	0.4971	0.7808	0.2676
HiResCAM	0.1299	–	0.5594	0.7944	0.2826
RISE-256	0.0602	–	0.4833	0.7665	0.4325
LayerCAM	0.1900	–	0.5571	0.8140	0.2664
ADAPT-TF	0.1533	–	0.5410	0.7978	0.2849
Vision-encoder contribution methods					
CLIP	0.1348	0.0380	0.5755	0.7961	0.6609
SigLIP2	0.1602	0.1118	0.6055	0.8077	0.6690
DINOV3	0.1641	0.1352	0.6055	0.8058	0.6643
IECLA (ours)	**0.6117**	**0.5100**	**0.8766**	**0.9322**	**0.8616**

The reference-specific results expose a stability pattern that is hidden by the average alone. Across Qwen Chat, Qwen Reasoner, and GPT Terra, IECLA's Top-1, NDCG@3, and Pointing scores vary

by at most 0.014, although the strongest competing method changes with the metric and reference. Importance and Signed Spearman show larger reference-dependent shifts, yet IECLA remains ahead of the competing methods. The result is therefore supported by agreement across references rather than an isolated match to one evaluator.

The baseline profiles also separate localization from signed attribution. Heatmap methods can obtain competitive NDCG@3 or Pointing on individual references, but their nonnegative outputs cannot represent opposing evidence. Vision-encoder contribution predictors add signed estimates and improve region localization, while their signed agreement remains substantially below IECLA. The simultaneous gain in ranking, sign, and Pointing indicates that the contribution predictor aligns these complementary properties instead of optimizing a single localization criterion.

D.2 Human Reference

Human Review starts from 320 emotion-balanced images; fixed usability filtering and human-region merging define the reported evaluation set. Forty annotators complete disjoint batches, with each item assigned to one annotator. All methods use the same SAM candidate masks and region mapping. We compute the ranking and localization metrics in Table 16 per image and average over images. Signed metrics are reported only for methods that produce signed contributions.

Table 16: Attribution agreement with independent Human Review annotations. IECLA leads in region ranking, signed agreement, and localization.

Method	Imp. $\rho \uparrow$	Signed $\rho \uparrow$	Top-1↑	NDCG@3↑	Pointing↑
Heatmap attribution methods					
Ablation-CAM	-0.2207	–	0.5502	0.7809	0.3218
Grad-CAM	-0.0450	–	0.5952	0.8159	0.3426
Direct attention	-0.0294	–	0.5744	0.8214	0.2595
Grad-CAM++	0.0115	–	0.6332	0.8272	0.3391
EigenCAM	0.0491	–	0.6298	0.8321	0.3495
HiResCAM	0.0707	–	0.6678	0.8474	0.3495
RISE-256	0.1264	–	0.6574	0.8397	0.4706
LayerCAM	0.1526	–	0.6886	0.8572	0.3460
ADAPT-TF	0.1942	–	0.6713	0.8464	0.3529
Vision-encoder contribution methods					
CLIP	0.0726	0.0547	0.6851	0.8492	0.7543
DINOV3	0.1373	0.1164	0.7093	0.8662	0.7716
SigLIP2	0.1412	0.0921	0.7093	0.8619	0.7509
IECLA (ours)	**0.5001**	**0.4525**	**0.8685**	**0.9315**	**0.8478**

Comparing Table 16 with the VLM average in Table 15 reveals which properties transfer most closely to human judgments. IECLA's NDCG@3 is nearly unchanged, while Top-1 and Pointing decrease only modestly; the larger reductions occur in Importance and Signed Spearman correlation. Thus, the leading region and top-weighted ordering are more stable across reference types than the full ordering of regional contributions. IECLA remains first throughout, whereas the strongest alternative changes between ADAPT-TF, DINOV3, and SigLIP2 depending on the criterion.

Signed contribution classification. Signed direction is evaluated independently on the same Human Review annotations. The protocol starts from 320 images and 1,319 annotated regions; filtering leaves 289 images and 1,032 human regions. When a human region overlaps several SAM masks, source-mask scores are aggregated with source-mask-area weights. The Sign Macro-F1 values in Table 17 use image-macro aggregation with 10,000 image-level bootstrap resamples, and signed methods use the fixed threshold $\tau = 0.1$. Nonnegative attention and saliency methods are omitted because they do not produce signed outputs.

IECLA reaches a Sign Macro-F1 of 0.3601, more than twice the strongest competing score of 0.1765. Because this classification uses independent Human Review annotations and a fixed sign threshold, the result complements the continuous signed-correlation analysis. IECLA aligns more closely with both regional ordering and the discrete distinction between evidence that supports or opposes the perceived source emotion.

This comparison tests the usefulness of the calibrated sign for distinguishing supporting and opposing perceptual evidence; it does not establish an absolute causal polarity for every candidate region.

E USER STUDY

This appendix provides the complete protocol and detailed results for the blinded User Study in Section 4.7. The study has two complementary components: a heatmap evaluation of whether the attributed regions are accurate and easy to interpret, and a paired editing evaluation of whether the same localized evidence improves target-emotion expression while preserving naturalness. We randomize candidate order independently within each task.

Protocol and scoring. Forty independent reviewers score the anonymized candidates on a five-level scale from -2 to 2. Heatmap candidates are evaluated by source-evidence accuracy and focus clarity; editing pairs are evaluated by target expression and naturalness. After decoding the blinded assignments, we aggregate reviewer scores. Table 18 summarizes the evaluation units and the corresponding ordinal reliability coefficients.

Table 17: Signed contribution classification against Human Review.

Method	Sign Macro-F1↑
DINOV3	0.1530
CLIP	0.1711
SigLIP2	0.1765
IECLA (ours)	**0.3601**

Table 18: Ordinal inter-rater reliability in the blinded User Study. Each evaluation unit is rated by all 40 reviewers.

Component	Tasks	Evaluation units	Krippendorff's α
Heatmap	20	520	0.8392
Editing	120 paired	480	0.7364
Overall	–	1,000	–

The two components exhibit different levels of ordinal inter-rater reliability. Heatmap ratings reach a Krippendorff's α of 0.8392, while the paired editing ratings reach 0.7364. These coefficients are computed from the unaggregated ratings of all 40 reviewers and measure reviewer consistency rather than method superiority. Reviewer means are used only for the subsequent task-level comparisons, so the reliability results support the interpretation of the study without replacing the task-level tests.

Heatmap evaluation. Each heatmap task fixes the source image and target emotion, then presents anonymized candidate maps generated under the same candidate-region protocol. The two criteria are averaged per candidate and then across the 20 tasks. The complete comparison is reproduced in Table 19. Its Friedman statistic and Kendall's W indicate that the observed ordering is descriptive because the repeated-measures test does not reach significance.

The heatmap scores are tightly clustered: most methods lie between 0.925 and 1.025, and task-level ties are frequent. Within this compressed range, IECLA attains the highest mean and appears among the top-scoring candidates on 18 of 20 tasks.

Table 19: Detailed heatmap User Study results. Mean score averages source-evidence accuracy and focus clarity; top-score counts include task-level ties among 20 tasks.

Method	Mean score↑	Top score↑
LayerCAM	0.775	10
CLIP	0.925	12
HiResCAM	0.950	13
ADAPT-TF	0.950	13
DINOV3	0.975	13
Grad-CAM++	0.975	14
EigenCAM	0.975	13
RISE-256	1.025	15
SigLIP2	1.025	14
Direct attention	1.025	14
Grad-CAM	1.025	15
Ablation-CAM	1.025	16
IECLA (ours)	**1.075**	**18**

Editing evaluation. For each editor, the same 20 source–target jobs are rendered with the native configuration and with IECLA-guided localized evidence. Each pair fixes the source image, target emotion, editor, and corresponding generation configuration, while the A/B order is randomized independently within each task. The table reports reviewer means and within-editor differences. Mean ratings for both target expression and naturalness increase across all six editors, consistent with the main-paper comparison in Section 4.7. Reviewer means are computed within each evaluation unit before the paired task-level comparison, while ordinal reliability is reported separately for the heatmap and editing components in Table 18.

The sign of the effect is stable across editors, whereas its magnitude is editor dependent. Qwen-Edit and HiDream show the largest target-expression gains, while Qwen-Edit and FLUX.2 show the largest naturalness gains. OmniGen2 already has the highest native target-expression score yet still improves by 0.800, so the benefit is not limited to correcting editors with weak native emotion con-

Table 20: Paired editing ratings across six backbones. IECLA guidance improves both target expression and naturalness for every editor.

	Target expression ↑			Naturalness ↑		
Editor	Native	IECLA-guided	Δ	Native	IECLA-guided	Δ
Qwen-Edit	0.0625	1.325	**+1.2625**	0.375	1.025	**+0.650**
HiDream	0.200	1.400	**+1.200**	0.425	0.675	**+0.250**
InstructPix2Pix	0.225	0.400	**+0.175**	0.500	0.575	**+0.075**
FLUX.2	-0.350	0.150	**+0.500**	1.050	1.425	**+0.375**
OmniGen2	0.650	1.450	**+0.800**	0.350	0.600	**+0.250**
LongCat	-0.075	0.075	**+0.150**	0.125	0.200	**+0.075**

trol. InstructPix2Pix and LongCat exhibit smaller but positive changes on both criteria, indicating that the response to localized evidence varies across editing architectures.

The paired pattern also clarifies the relation between the two subjective objectives. Stronger target expression does not coincide with a naturalness decrease for any of the six editors in this evaluation; all naturalness differences remain positive. Together with the variation in gain magnitude, these point estimates suggest that localized IECLA evidence is useful across editing architectures, while the realized benefit remains dependent on the underlying editor.

F Additional Ablations

Table 21 isolates training and editing decisions that are not separated in the main end-to-end ablation. All variants use the same 680-image, 5,440-task evaluation pool, predictor architecture, editor, and metric protocol; only the named component or supervision term is changed.

Table 21: Additional editing and loss-function ablations. IECLA provides the strongest overall balance between target-emotion control and structural preservation.

Variant	Target Success ↑	Δp ↑	Emo-A ↑	Emo-S ↑	CLIP-I ↑	G. LPIPS ↓	G. SSIM ↑	O. LPIPS ↓	O. SSIM ↑	O. MAE ↓
IECLA	**0.3579**	**0.3344**	**0.3081**	**0.2173**	**0.8527**	**0.3440**	**0.6137**	0.3767	**0.3230**	0.1733
-w/o $\mathcal{L}_{\text{abs}}$	0.3412	0.3210	0.2901	0.2150	0.8421	0.3507	0.6000	**0.3721**	0.3185	0.1831
-w/o $\mathcal{L}_{\text{signed}}$	0.3471	0.3226	0.2921	0.2150	0.8415	0.3533	0.5975	0.3751	0.3173	0.1842
-w/o $\mathcal{L}_{\text{focal}}$	0.3390	0.3164	0.2899	0.2110	0.8438	0.3515	0.5996	0.3729	0.3185	0.1834
-w/o $\mathcal{L}_{\text{aux}}$	0.3465	0.3221	0.2914	0.2143	0.8418	0.3534	0.5969	0.3752	0.3169	0.1842
-w/o fusion	0.3542	0.3318	0.2921	0.2139	0.8451	0.3476	0.6073	0.3851	0.3214	0.1746
-w/o template	0.3518	0.3280	0.2969	0.2131	0.8397	0.3524	0.6035	0.3862	0.3180	0.1793
-w/o negative protection	0.3504	0.3335	0.2820	0.1980	0.8454	**0.3440**	0.6028	0.3839	0.3167	0.1774
Random matched	0.2739	0.2507	0.2706	0.1658	0.8481	0.3442	0.6051	0.3772	0.3143	**0.1716**

The complete IECLA configuration is the reference condition. The "w/o fusion" variant removes global–local context fusion from the contribution predictor, testing whether regional evidence benefits from joint local and scene-level representations. The "w/o template" variant retains attribution and agent planning but omits the structured editing template, testing whether the template converts evidence into consistent editing instructions. The "w/o learning" variant bypasses the contribution predictor and passes the corrected counterfactual contributions directly from attribution to target-aware editing, testing whether learned contribution representations improve over direct evidence transfer. These settings complement the main-paper "w/o attribution", "w/o learning", and "w/o agent" conditions rather than duplicating them.

Two controls target the use of signed evidence during editing. In *w/o negative protection*, positive contribution regions and the affective instruction are unchanged, but the prompt omits preservation constraints derived from negatively contributing regions. This tests whether negative contributions act as preservation evidence rather than merely being unused scores. In *Random matched*, the attributed positive region is replaced by a deterministically sampled candidate satisfying the same matching criteria, while prompt structure and editing strength are retained. This control tests whether the gain comes from selecting emotionally relevant evidence instead of from structured prompting alone. The fixed seed and image identifier make the sampled candidate reproducible.

The remaining rows remove one loss from Equation 5 at a time while keeping the predictor, data, and downstream editing protocol fixed. "w/o $\mathcal{L}_{\text{abs}}$" removes absolute ranking supervision and tests contribution-magnitude ordering; "w/o $\mathcal{L}_{\text{signed}}$" removes signed ranking supervision and tests

direction-aware ordering; "w/o $\mathcal{L}_{\text{focal}}$" removes focal-sign supervision and tests emphasis on difficult or imbalanced sign decisions; and "w/o $\mathcal{L}_{\text{aux}}$" removes auxiliary supervision and tests the contribution of the auxiliary affective signals. Bold and underlined entries denote the best and second-best point estimates in each column.

The editing controls separate evidence selection from instruction structure. Removing negative protection lowers all four affective measures and reduces outside-region SSIM, while G. LPIPS remains tied with IECLA at 0.3440. Negative contributions therefore constrain unnecessary changes without simply suppressing the edit. Random matched causes the largest affective drop, reaching 0.2739 Target Success and 0.1658 Emo-S, even though it obtains the lowest O. MAE (0.1716). The combination of weak emotion control and conservative pixel change shows that the selected region, rather than structured prompting alone, determines whether the edit is useful. Removing fusion or the template produces smaller but consistent losses in affective and structural measures, with the template ablation particularly lowering CLIP-I to 0.8397 and increasing O. LPIPS to 0.3862.

The main-paper component ablations expose three distinct roles in the editing pipeline. Removing attribution leaves the agent without region-specific affective evidence, so its planning cannot reliably identify where the target emotion should be expressed or which regions should be preserved. Removing learning retains the attribution signal but bypasses its learned calibration, context integration, and stable magnitude and sign ordering; direct evidence transfer therefore provides a noisier control signal and permits more unintended changes outside the selected region. Removing the agent retains the learned contribution representation, allowing the editor to use localized evidence directly, but loses the structured decomposition of target-emotion actions and preservation constraints that makes the evidence consistently executable. Together, the three conditions separate the roles of evidence construction, evidence calibration, and evidence-to-instruction planning.

The loss ablations reveal a complementary training effect. Removing focal-sign supervision gives the largest affective degradation among the four loss variants, whereas removing either ranking term or auxiliary supervision produces smaller declines across the affective metrics. Individual variants can improve one preservation distance, such as O. LPIPS after removing absolute ranking, but none matches IECLA's joint profile: the complete model is best on all four affective measures, CLIP-I, G. SSIM, and O. SSIM, while tying the best G. LPIPS value and remaining second-best on O. MAE. The results link the full objective to three functions of the framework: ordering relevant evidence, preserving the direction of signed contributions, and maintaining structural fidelity during editing.

G Stability Analysis

Robustness to sampling variation is assessed with Qwen-Edit, using the configuration in Appendix J, on 16 routable images stratified by the eight source emotions, with two images per class. Each image is edited toward the other seven emotions, yielding 112 matched tasks. The source image, target emotion, agent configuration, routing protocol, and SAM regions are fixed. Native receives only the global target-emotion instruction; IECLA-guided editing augments it with at most three human-related SAM regions whose predicted positive contribution exceeds 0.05. Each pair shares its source image, target emotion, seed, and resolution.

Seeds 48141, 48142, and 48143 are evaluated at the main resolution 832×480 and the adjacent resolution 1024×576. Each resolution contains $112 \times 2 \times 3 = 672$ outputs, or 1,344 in total. Target Success is the frozen EmoVerse classifier's target decision, and Target probability Δ measures change relative to the source image. Preservation is measured outside the hard union of routed SAM masks. Affective IECLA-guided–Native differences use 10,000 source-level paired bootstrap resamples. Preservation values remain point differences because verified paired intervals were not retained; resolution-level standard deviations are computed across the three seed-level means.

IECLA-guided editing improves Target Success and target probability for every seed and resolution. At 1024×576, all affective intervals exclude zero; the three-seed mean gains are $+0.101 \pm 0.005$ for Target Success and $+0.095 \pm 0.015$ for target probability. At 832×480, the corresponding gains are smaller at $+0.054 \pm 0.009$ and $+0.053 \pm 0.010$; the seed-48141 intervals cross zero, while the other two exclude zero. Outside-mask LPIPS decreases and SSIM increases in all six settings. The repeated directions support sampling stability within each resolution, while their magnitudes indicate sensitivity to output resolution.

Table 22: Stability across seeds and resolutions. IECLA shows favorable affective and outside-mask preservation differences in all six settings.

	Affective response		Outside-mask preservation	
Seed	Δ Target Success [95% CI]	Δ Target prob. [95% CI]	Outside-mask Δ LPIPS	Outside-mask Δ SSIM
Resolution: 1024×576				
48141	+0.098 [+0.027, +0.179]	+0.098 [+0.034, +0.169]	-0.010	+0.001
48142	+0.107 [+0.018, +0.214]	+0.108 [+0.035, +0.199]	-0.018	+0.002
48143	+0.098 [+0.036, +0.170]	+0.078 [+0.036, +0.128]	-0.005	+0.010
Resolution: 832×480				
48141	+0.045 [-0.009, +0.107]	+0.044 [-0.013, +0.107]	-0.011	+0.005
48142	+0.054 [+0.009, +0.116]	+0.051 [+0.004, +0.108]	-0.008	+0.001
48143	+0.062 [+0.018, +0.107]	+0.063 [+0.019, +0.104]	-0.005	+0.008

The stability evidence is limited to 16 source images, three seeds, two resolutions, and one editor. It supports consistency of the observed direction under this protocol, rather than universal robustness across editors, samplers, resolutions, or image populations.

H Teacher Replay and Predictor Comparison

This appendix isolates whether the learned contribution predictor improves on direct replay of the intervention-based supervision that defines its target. The comparison is useful because the replay is an exact counterfactual measurement, whereas the predictor can denoise that signal and generalize it to the reference-validation protocol.

Experimental setup. The replay uses the same frozen Phase-1 emotion evaluator E that constructs the corrected counterfactual contributions in Section 3.2. For each of the 1,032 aggregated candidate regions on the 289-image reference-validation set, we infer the source emotion with E, apply the candidate blur intervention, apply its matched reference intervention, and compute the signed contribution with Equation 2. No predictor is trained or selected for this condition. Candidate masks, reference-region matching, prediction-map rendering, region aggregation, sign threshold, and image-macro metric computation are identical to the reference-validation protocol. The replay is compared with IECLA's learned predictor against Human Review, Qwen Chat, Qwen Reasoner, GPT Terra, and their unweighted VLM mean.

Table 23: Teacher replay versus the IECLA predictor under the reference-validation protocol.

	Imp. ρ ↑			Signed ρ ↑			Top-1↑			Top-3↑			Top-5↑		
Reference	Replay	IECLA	Δ	Replay	IECLA	Δ	Replay	IECLA	Δ	Replay	IECLA	Δ	Replay	IECLA	Δ
Human Review	0.3730	**0.5001**	**+0.1271**	0.4094	**0.4525**	**+0.0431**	0.8097	**0.8685**	**+0.0588**	0.9758	**0.9896**	**+0.0138**	0.9965	**0.9965**	**+0.0000**
Qwen Chat	0.4522	**0.6187**	**+0.1665**	0.4522	**0.4868**	**+0.0346**	0.7751	**0.8720**	**+0.0969**	0.9792	**0.9896**	**+0.0104**	0.9931	**1.0000**	**+0.0069**
Qwen Reasoner	0.4330	**0.5998**	**+0.1668**	0.4195	**0.5108**	**+0.0913**	0.7785	**0.8789**	**+0.1004**	0.9723	**0.9792**	**+0.0069**	0.9931	**1.0000**	**+0.0069**
GPT Terra	0.4477	**0.6165**	**+0.1688**	0.4275	**0.5324**	**+0.1049**	0.7612	**0.8789**	**+0.1177**	0.9654	**0.9827**	**+0.0173**	0.9931	**1.0000**	**+0.0069**
VLM mean	0.4443	**0.6117**	**+0.1674**	0.4330	**0.5100**	**+0.0770**	0.7716	**0.8766**	**+0.1050**	0.9723	**0.9838**	**+0.0115**	0.9931	**1.0000**	**+0.0069**
	NDCG@3↑			NDCG@5↑			Pointing↑			Sign Accuracy↑			Sign F1↑		
Reference	Replay	IECLA	Δ	Replay	IECLA	Δ	Replay	IECLA	Δ	Replay	IECLA	Δ	Replay	IECLA	Δ
Human Review	0.9071	**0.9315**	**+0.0244**	0.9361	**0.9537**	**+0.0176**	0.7993	**0.8478**	**+0.0485**	0.2055	**0.4320**	**+0.2265**	0.1522	**0.3601**	**+0.2079**
Qwen Chat	0.8943	**0.9331**	**+0.0388**	0.9229	**0.9517**	**+0.0288**	0.8131	**0.8685**	**+0.0554**	0.3334	**0.5164**	**+0.1830**	0.2630	**0.4500**	**+0.1870**
Qwen Reasoner	0.8914	**0.9313**	**+0.0399**	0.9147	**0.9489**	**+0.0342**	0.7993	**0.8616**	**+0.0623**	0.3660	**0.5440**	**+0.1780**	0.2880	**0.4842**	**+0.1962**
GPT Terra	0.8851	**0.9323**	**+0.0472**	0.9187	**0.9521**	**+0.0334**	0.7924	**0.8547**	**+0.0623**	0.3074	**0.5395**	**+0.2321**	0.2446	**0.4745**	**+0.2299**
VLM mean	0.8903	**0.9322**	**+0.0419**	0.9188	**0.9509**	**+0.0321**	0.8016	**0.8616**	**+0.0600**	0.3356	**0.5333**	**+0.1977**	0.2652	**0.4695**	**+0.2043**

The replay results in Table 23 show that the learned predictor is not merely a compressed copy of the teacher's ranking. IECLA is higher on every metric for the three individual VLM references and on the VLM mean, while it also improves Human Review agreement on all metrics except the tied Top-5 score. The largest gaps occur in signed measures: on the VLM mean, Signed Spearman rises from 0.4330 to 0.5100, Sign Accuracy from 0.3356 to 0.5333, and Sign F1 from 0.2652 to 0.4695. This pattern is consistent with the predictor learning a more stable signed representation from the intervention supervision rather than reproducing a single noisy replay.

The result also clarifies the role of direct replay. The teacher replay remains the exact intervention-defined source of supervision, but its direct maps are weaker than the learned predictor under the independent VLM and Human Review references. At deployment, replay would require repeated evaluator calls for candidate and matched reference interventions, whereas IECLA reuses one source-emotion inference and applies the learned predictor once per region. Although latency was not measured in this experiment, this replacement reduces counterfactual computation and is expected to lower inference time.

I AGENT TEMPLATE

The extended agent retains attribution-grounded local editing while admitting two optional, independently controlled channels for scene atmosphere and minor-object operations. Editable and preservation evidence remain distinct throughout planning. A fourth round audits the complete assembled instruction after the structured plan has passed visual and structural review. New-object proposals are treated as context-conditioned editing hypotheses rather than observed attribution evidence. Figure 10 summarizes the paper-level template.

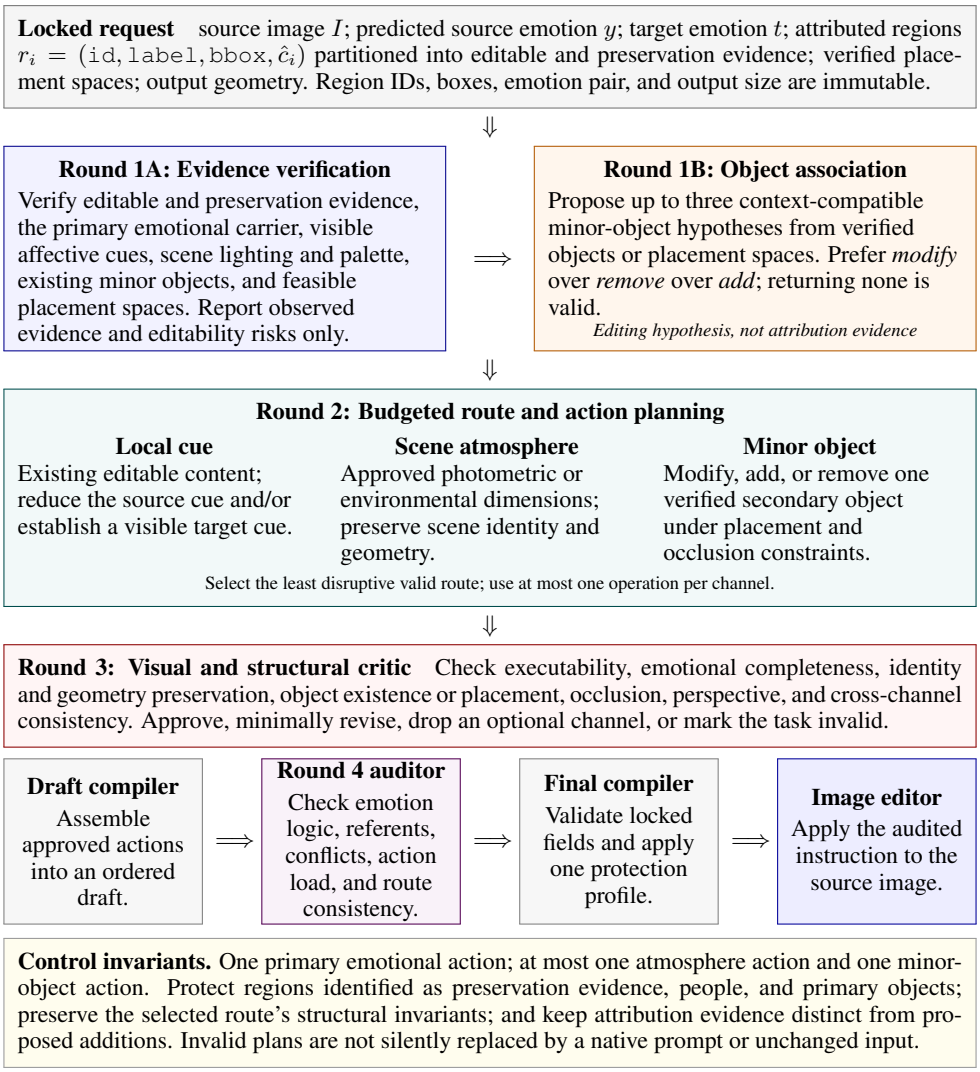


Figure 10: Extended attribution-grounded editing agent template. Verified evidence is routed through local, atmosphere, and optional minor-object operations; structural review and whole-instruction auditing constrain the final editor instruction.

The route-specific compiler emits only approved clauses and one compatible protection profile. This avoids contradictory constraints such as simultaneously adding an object and requiring object count to remain unchanged.

J Implementation Details

This appendix specifies the training configuration and editing controls used in Sections 3 and 4.1.

J.1 Data Provenance and Splits

The contribution-training, locked attribution-test, and editing-evaluation pools are constructed from EmoVerse. Images belonging to the external evaluation datasets, including EmoSet and EmoArt, are removed before contribution-model training. The training pool supplies 1,056,949 candidate masks, whereas the image-disjoint locked attribution test contains 7,278 images and 53,180 masks. The editing benchmark uses a separate set of 680 challenging images selected to include multi-subject scenes, scenes without a dominant subject, and scenes containing many objects. Deduplication and overlap checks are applied before the final splits, leaving the training, locked attribution-test, and editing-evaluation pools mutually disjoint.

J.2 Editing Tasks and Evaluation Metrics

Task construction. The main editing pool contains 680 source images, each paired with all eight target emotions: amusement, anger, awe, contentment, disgust, excitement, fear, and sadness. This construction yields $680 \times 8 = 5{,}440$ source–target task keys, including 680 same-emotion tasks ($t = y$) and 4,760 cross-emotion tasks ($t \neq y$), where the frozen image-level evaluator supplies the source emotion y. The six editors use the same task keys, and Native and IECLA-guided conditions share the source image, target emotion, seed, sampler, and resolution rule. Tables 2, 4, and 21 use the same evaluation set for their corresponding Qwen-Edit conditions, which accounts for their identical shared entries.

Cross-dataset editing applies every target emotion to every image used in the attribution evaluation. The 3,370 external images therefore produce 26,960 editing tasks, comprising 3,370 same-emotion and 23,590 cross-emotion tasks. Table 3 averages over 4,896 COCO tasks, 7,320 EmoSet tasks, 12,352 FI tasks, 1,808 Emotion6 tasks, and 584 EmoArt tasks. The stability analysis in Appendix G is a separate diagnostic: its 16 images are edited only toward the other seven emotions, giving 112 tasks for each seed and resolution.

Table 24: Editing-task accounting. The main and cross-dataset evaluations include all eight targets, whereas the controlled cross-emotion diagnostics exclude $t = y$.

Evaluation	Images	Targets/image	Same-emotion	Cross-emotion	Tasks
Main editing pool	680	8	680	4,760	5,440
COCO	612	8	612	4,284	4,896
EmoSet	915	8	915	6,405	7,320
FI	1,544	8	1,544	10,808	12,352
Emotion6	226	8	226	1,582	1,808
EmoArt	73	8	73	511	584
Cross-dataset total	**3,370**	**8**	**3,370**	**23,590**	**26,960**
Stability, per seed/resolution	16	7	0	112	112

Affective metrics. Let I_s and I_e denote the source and edited images, and let t be the requested emotion. The frozen EmoVerse-supervised evaluator E defines

$$\begin{aligned} \mathrm{Success}(I_e, t) &= \mathbb{I}\Big[\arg\max_e E_e(I_e) = t\Big], \\ \Delta p(I_s, I_e, t) &= E_t(I_e) - E_t(I_s). \end{aligned} \tag{6}$$

An independent CLIP-based emotion classifier C provides a second evaluation:

$$\begin{aligned} \text{Emo-A}(I_e, t) &= \mathbb{I}\Big[\arg\max_e C_e(I_e) = t\Big], \\ \text{Emo-S}(I_s, I_e, t) &= C_t(I_e) - C_t(I_s). \end{aligned} \tag{7}$$

The classifier C follows the EmoEdit evaluation model: a frozen CLIP ViT-B/32 image encoder feeds a two-layer classifier with dimensions $512 \to 256 \to 8$ and a softmax output. Its released checkpoint was pretrained on EmoSet. We performed image-level overlap checks between its EmoSet training data and the EmoSet subset used in our cross-dataset evaluation, confirming that the two sets are disjoint. Thus, Target Success and Emo-A are target-class accuracies from different held-out evaluators, while Δp and Emo-S measure target-probability changes from the corresponding evaluators. Reported entries are arithmetic means over the task set specified for each experiment; source-image cluster resampling is used when confidence intervals are reported.

Evaluator roles and metric provenance. The frozen evaluator E provides the source-emotion label and counterfactual supervision, while the contribution predictor scores candidate regions and the target-aware agent plans edits from those scores. The classifier C is used only for post-hoc Emo-A and Emo-S evaluation; its scores are never fed back to the predictor or agent. VLM references, Human Review, and the User Study provide complementary external and perceptual evaluations.

Global preservation metrics. All source and edited images are converted to RGB and resized to 512×512 with Lanczos interpolation. PSNR, SSIM, and LPIPS compare the complete images, while CLIP-I is the cosine similarity between their CLIP image embeddings. In Tables 4 and 21, the prefix G. denotes the same complete-image evaluation. G. SSIM is computed on uint8 images with data range 255. G. LPIPS uses the AlexNet LPIPS backbone on float32 images scaled to $[0, 1]$. We call the official LPIPS implementation with `normalize=True`, which internally maps these inputs to the expected $[-1, 1]$ range:

$$\text{G. LPIPS} = D_{\text{LPIPS}}(I_s, I_e). \tag{8}$$

All compared conditions use this identical normalized implementation. The reported values therefore follow the standard LPIPS preprocessing while retaining the common $[0, 1]$ image representation used by the rest of our evaluation code.

Common outside-region protocol. Outside-region metrics measure changes beyond a common allowed editing area. For each task, the selected binary SAM masks from all compared conditions are resized to 512×512 by nearest-neighbor interpolation and combined by a hard union,

$$A = \bigcup_{v \in \mathcal{V}} M_v, \qquad O = \neg A, \tag{9}$$

where $\mathcal{V}$ is the set of variants in the comparison. Fixing A across variants prevents a condition from improving its preservation score by declaring a larger editable region. In the loss-function ablation, $\mathcal{V}$ contains the complete objective and the four loss-removed variants. Their common allowed region covers 22.7521% of the image on average, leaving 77.2479% for outside-region evaluation. This experiment uses the same 680 source images and all eight target emotions, yielding $680 \times 8 = 5{,}440$ tasks.

O. MAE averages the RGB absolute error over O:

$$\text{O. MAE} = \frac{1}{|O|} \sum_{p \in O} \frac{\|I_e(p) - I_s(p)\|_1}{3}. \tag{10}$$

O. SSIM first computes the full-resolution SSIM map and then averages its values over O. For O. LPIPS, the allowed region is restored to the source image,

$$I_{\text{outside}} = A \odot I_s + (1 - A) \odot I_e, \qquad \text{O. LPIPS} = D_{\text{LPIPS}}(I_s, I_{\text{outside}}). \tag{11}$$

O. LPIPS is therefore an outside-focused perceptual distance rather than a pixelwise regional average; convolutional receptive fields may still cross the mask boundary.

The affective and preservation metrics are interpreted jointly: minimal image change is not beneficial when the requested emotion is not expressed, while improved affective scores accompanied by poor outside-region preservation indicate unnecessary rewriting.

J.3 Counterfactual Reference Calibration

The reference correction in Equation 2 uses dataset-level calibration rather than a fixed number k of reference masks for each candidate. We sample $N_{\text{ref}} = 10{,}000$ training records and generate one reference intervention per record. Reference regions are matched to candidate regions in nominal area, spatial position, and shape. The resulting reference effects are averaged within source-emotion and area bins, using five area ranges: $[0, 0.005)$, $[0.005, 0.02)$, $[0.02, 0.10)$, $[0.10, 0.30)$, and $[0.30, 1]$. Each emotion–area cell contains at least 30 references; otherwise, the area-level and then global mean provide the fallback. Thus, $\mathbb{E}[\Delta_r]$ denotes the corresponding calibration mean, not a per-region Monte Carlo average.

Both candidate and reference interventions replace the matched region with a radius-24 Gaussian-blurred version. Matching area, position, and shape ensures comparable intervention geometry, while averaging over the calibration set estimates the blur-induced nuisance effect. Subtracting this expectation leaves the candidate-specific residual as the corrected contribution.

We interpret c_i on this calibrated relative scale: $c_i < 0$ indicates an effect below the matched-reference mean, while its role as an opposing editing cue is validated against Human Review.

J.4 Predictor Training and Objective

Predictor training. The contribution predictor is trained on 1,056,949 candidate masks and validated on 58,591 disjoint masks. Every loss term uses the corrected counterfactual contribution c_i from Equation 2 as its supervision target, and checkpoint selection minimizes validation MAE on these corrected contributions. Training uses three NVIDIA A800 GPUs.

Supervision and weighted regression. Let $\widehat{c}_i$ denote the predicted contribution and $\tau = 0.1$ the fixed sign threshold. The target sign is negative for $c_i < -\tau$, neutral for $|c_i| \le \tau$, and positive for $c_i > \tau$. We define the unit-transition Huber penalty as

$$\rho(e) = \begin{cases} \frac{1}{2}e^2, & |e| \le 1, \\ |e| - \frac{1}{2}, & |e| > 1. \end{cases} \tag{12}$$

Let b_i be the evaluator's pre-intervention class and $\widetilde{y}_i$ the image-level training label. The sample weight

$$w_i = 3^{\mathbb{I}[c_i < -\tau]} 1.5^{\mathbb{I}[|c_i| > \tau]} 0.5^{\mathbb{I}[b_i \neq \widetilde{y}_i]} \tag{13}$$

combines the three multipliers rather than selecting only one of them. The normalized regression term is

$$\mathcal{L}^{w}_{\text{Huber}} = \frac{\sum_i w_i \rho(\widehat{c}_i - c_i)}{\sum_i w_i}. \tag{14}$$

Within-image ranking. Let $g(i)$ identify the source image of region i. Absolute and signed ranking use, respectively, the valid pair sets

$$\begin{aligned} \mathcal{P}_{\text{abs}} &= \{(i,j) : g(i) = g(j),\ i < j,\ |c_i| \neq |c_j|\}, \\ \mathcal{P}_{\text{signed}} &= \{(i,j) : g(i) = g(j),\ i < j,\ c_i \neq c_j\}. \end{aligned} \tag{15}$$

With $\sigma(\cdot)$ denoting the sigmoid function, the two losses are

$$\begin{aligned} \mathcal{L}_{\text{abs}} &= -\frac{1}{|\mathcal{P}_{\text{abs}}|} \sum_{(i,j)\in\mathcal{P}_{\text{abs}}} \log \sigma\big(s^a_{ij}(|\widehat{c}_i| - |\widehat{c}_j|)\big),\ s^a_{ij} = \text{sign}(|c_i| - |c_j|), \\ \mathcal{L}_{\text{signed}} &= -\frac{1}{|\mathcal{P}_{\text{signed}}|} \sum_{(i,j)\in\mathcal{P}_{\text{signed}}} \log \sigma\big(s^s_{ij}(\widehat{c}_i - \widehat{c}_j)\big),\ s^s_{ij} = \text{sign}(c_i - c_j). \end{aligned} \tag{16}$$

The absolute term orders regions by contribution strength, whereas the signed term preserves their ordering on the corrected contribution scale.

Sign and affective auxiliary supervision. For the three sign classes $q_i \in \{-, 0, +\}$, let $p_i^{q_i}$ be the predicted probability of the target class. With class weights $(\alpha_-, \alpha_0, \alpha_+) = (4.0, 0.3, 1.1)$ and focal exponent $\gamma = 2$, the sign-classification loss is

$$\mathcal{L}_{\text{focal}} = -\frac{\sum_i \alpha_{q_i}(1 - p_i^{q_i})^2 \log p_i^{q_i}}{\sum_i \alpha_{q_i}}. \tag{17}$$

The auxiliary heads predict the intensity difference d_i^{int} and the three-dimensional VAD difference $\mathbf{d}_i^{\text{VAD}}$ associated with the same intervention. For a batch of B regions, their loss is

$$\mathcal{L}_{\text{aux}} = \frac{1}{B}\sum_i \rho(\widehat{d}_i^{\text{int}} - d_i^{\text{int}}) + \frac{0.2}{3B}\sum_i \sum_{k=1}^{3} \rho(\widehat{d}_{ik}^{\text{VAD}} - d_{ik}^{\text{VAD}}). \tag{18}$$

The coefficient 0.2 applies within $\mathcal{L}_{\text{aux}}$, Equation 5 further weights the full auxiliary term by 0.10.

Grouped sampling and pair construction. Candidate regions are grouped by source image, image groups are shuffled each epoch, and complete groups are packed without splitting into batches of up to 128 regions per process. With three training processes, this gives the effective batch size of 384 reported above. Both ranking terms enumerate every valid unordered pair within each image group; they use neither random pair subsampling nor a ranking margin. Tied targets are omitted, and image groups with fewer than two regions contribute zero to the ranking terms.

J.5 Attribution Baseline Implementations

The vision-encoder contribution baselines are trained comparison versions used to produce signed heatmaps. They share the candidate masks, counterfactual supervision, and mask-to-pixel rendering protocol of IECLA, but use simpler backbone-specific predictors. The CLIP version combines frozen ViT-L/14 region-crop and full-image features with emotion conditioning, object labels, and region metadata. The SigLIP2 version uses a frozen SigLIP2-base encoder with region and full-image context, emotion and object conditions, and the same metadata. The DINOv3 version pools frozen ViT-7B/16 patch tokens within each SAM mask and combines the region representation with full-image context, emotion and object conditions, and metadata. Each predictor outputs a signed regional contribution that is projected onto the fixed masks; these baselines are trained specifically for heatmap comparison and are not native attention maps.

Grad-CAM, Grad-CAM++, HiResCAM, LayerCAM, EigenCAM, and Ablation-CAM use the frozen eight-class emotion evaluator E as their target network and explain its inferred source-emotion class. RISE-256 uses the same evaluator with 256 random masks sampled on a 7×7 grid. Each resulting saliency map is resized to the input resolution, and its score for a candidate region is the mean saliency within the corresponding fixed SAM mask. Direct attention instead uses the CLIP visual encoder from the CLIP contribution baseline: multi-head self-attention is augmented with the identity connection and rolled out across transformer layers to obtain the CLS-to-patch map. ADAPT-TF likewise follows its own Qwen image-token attention path rather than the evaluator E. These heatmap methods produce nonnegative region scores and are therefore evaluated only with unsigned ranking and localization metrics.

Agent constraints and candidate selection. The predictor supplies up to three attributed regions with $\widehat{c}_i > 0.05$ as editable evidence and retains strongly negative regions as preservation evidence. The agent verifies a primary emotional carrier, and the local action remains anchored to the editable evidence. Scene-atmosphere changes and minor-object operations are optional, independently budgeted channels that require visible scene support; a proposed added object is recorded as an editing hypothesis rather than attributed evidence. At most one primary, one atmosphere, and one minor-object operation are retained. A visual and structural critic checks feasibility and preservation, after which a deterministic draft compiler and integrated-instruction auditor instantiate the template in Appendix I. Reporting evaluators remain held out from planning and candidate selection.

Agent VLM configuration. All agent rounds use the same locally served vision-language model under the API identifier `qwen36-vision`. The model is loaded from `Qwen3.6-35B-A3B-Uncensored-HauhauCS-Aggressive-Q8_K_P.gguf` together with

`mmproj-Qwen3.6-35B-A3B-Uncensored-HauhauCS-Aggressive-f16.gguf`. It uses the `qwen35moe` architecture with 34.66B total parameters and approximately 3B active parameters, and the language checkpoint uses `Q8_K_P` quantization. Thinking is enabled for every round; decoding uses temperature 0.1, the server default top-p of 0.95, at most 1,024 output tokens per round, and an 8,192-token service context.

The original image is transmitted as base64 data without client-side resizing. The vision encoder has a nominal input size of 768 and a patch size of 16, accepts dynamic inputs from 8,192 to 4,194,304 pixels, and uses a `qwen3vl_merger` projector with spatial merge size 2.

Editor configurations. The editing experiments use Qwen-Edit, InstructPix2Pix, FLUX.2, LongCat, OmniGen2, and HiDream. Qwen-Edit denotes Qwen-Image-Edit-2511, FLUX.2 denotes FLUX.2 Klein 4B, LongCat denotes LongCat-Turbo, and HiDream denotes HiDream-E1.1; InstructPix2Pix and OmniGen2 use their official released checkpoints. The stability analysis in Appendix G runs Qwen-Edit with Lightning four-step sampling.

Editing targets. The eight target emotions are amusement, anger, awe, contentment, disgust, excitement, fear, and sadness. Native and IECLA-guided outputs share the source image, target emotion, seed, sampler, and resolution rule within each editor. Absolute scores across different backbones are not interpreted as a ranking of editors.