AFFECTEDIT: DIAGNOSING AND FIXING EVALUATION FOR AFFECTIVE IMAGE EDITING

Anonymous authors
Paper under double-blind review

ABSTRACT

Image editors can now change the emotion an image conveys, yet it is unclear whether the metrics used to evaluate such affective edits agree with human judgment. To test them, we collect **AffectEdit**, a benchmark of $1{,}406$ edits, each rated by about five annotators for whether it conveys the target emotion while preserving the scene. We find that the metrics in current use for affective editing fall short because they assess these two requirements separately. Emotion-only metrics can reward an edit that replaces the entire scene, scene-only metrics reward an edit that keeps the scene but leaves the emotion unchanged, and the one metric in current use that scores both adds the two terms, so either can mask a failure of the other. Replacing its addition with a multiplication, on the metric's own terms, makes it reject unchanged images and significantly raises its correlation with human ratings over the whole benchmark. We therefore propose PAEF (Perceived Affective Edit Fidelity), a parameter-free metric that multiplies an edit's emotion shift by its scene preservation, so that an edit scores highly only when both requirements hold. PAEF agrees with human ratings better than every metric in current use for affective edits, and holds that level on a held-out set of edits that informed none of its design choices. The strongest multimodal large language models (MLLMs), prompted or fine-tuned as judges, agree better still, but PAEF runs 3.7 to 61 times faster than an MLLM judge on the same GPU and needs no rated data. We release the benchmark, the ratings and our code (supplementary material).

1 INTRODUCTION

Generative image editors have recently developed the capability to operate directly on *perceived emotional content*: instructions like *"make this image feel sadder"* or *"make the room more inviting,"* and explicit target emotions such as *fear*, *awe*, *contentment* or *excitement*, are now within the scope of foundation editing models (Brooks et al., 2023; Lin et al., 2025; Mao et al., 2026; Yang et al., 2025). Whether such systems are improving, and whether one model is better than another, depends on the metric used to score them.

Affective edit evaluation is fragmented across separate axes. Existing practice scores edits along two independently reported dimensions. For *emotional quality*, papers report classifier accuracy on the target emotion (Emo-A; Li et al., 2026; Mao et al., 2026; Yang et al., 2024; 2025), the increase in the classifier's target-class probability from source to edit (Emo-S; Mao et al., 2026; Yang et al., 2025), CLIP similarity to the target emotion as text (Hessel et al., 2021; Ye & Huang, 2025), or the KL divergence between predicted and target emotion distributions (Lin et al., 2025; Zhang et al., 2026). For *scene preservation*, editing papers separately report SSIM (Wang et al., 2004), PSNR, LPIPS (Zhang et al., 2018) or CLIP-I between the original and edited image (Li et al., 2026; Mao et al., 2026; Yang et al., 2025; Yin et al., 2026). ESMI (Lin et al., 2025) unifies the two streams into a single number, a weighted sum of an emotion score and a structure-preservation score.

What does success actually look like? We argue that a successful affective edit achieves *maximal emotion impact with minimal scene change*: a slight change of expression, a shift in colour or lighting, an added or removed atmospheric detail might be just enough to move the perceived emotion toward the target. The two requirements are not separable: an edit is valid only when both hold at once, and a metric that lets one mask the other has no claim to validity. This frames the central empirical question of our work: *which of the metrics in current use satisfy this requirement?* To answer it, we

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

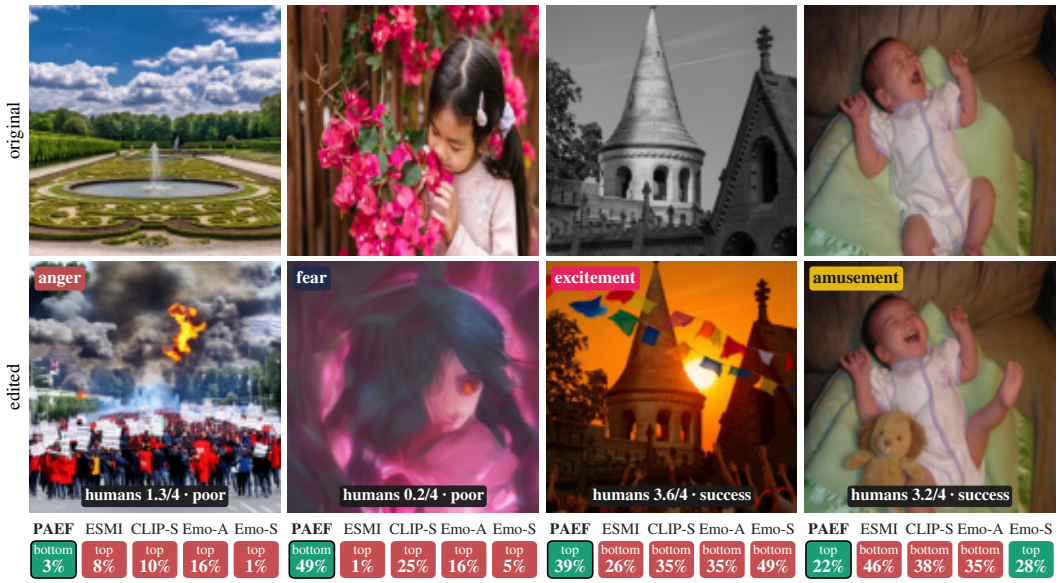

Figure 1: **Metrics in current use reward edits that replace the scene; PAEF's scene gate does not.** Each column shows an original (top) and its edit toward the named emotion (bottom), with the raters' verdict. Each cell gives where a metric ranks the edit among all 1,406 edits (green: on the raters' side of the middle; red: not). The two poorly rated edits (1.3 and 0.2 of 4) hit the emotion but replace the scene: the emotion-only metrics and ESMI rank them in their top quarter, PAEF does not. The two successes keep the scene: PAEF ranks them in its top 39% and 22%, ESMI and CLIPScore (CLIP-S) in their bottom half. The edits were chosen to illustrate these patterns; Fig. 3 shows all pairs.

collect **AffectEdit**: 1,406 edit pairs, each rated by about five qualified crowd workers on a single 0–4 scale on which 0 marks a failed edit, one that *changes nothing* (an identity edit) *or turns the image into a different photo* (a content swap). Averaged over raters, the scores are highly reliable (intraclass correlation coefficient $\mathrm{ICC}(2,5)$=0.824; Shrout & Fleiss, 1979; Sec. D).

We observe that current metrics fall into three groups, and each fails in its own way. **(1) Scene-only metrics** cannot see the emotion at all, and CLIP-I even anti-correlates with human judgment, penalizing exactly the changes a successful edit must make. **(2) Emotion-only metrics** such as Emo-A and Emo-S reach only modest agreement (Kendall's $\tau_b \approx 0.17$ (Kendall, 1945)) and ignore image content, so a content swap that fools the classifier can score as a success. **(3) ESMI**, the one metric in current use that scores both, is not significant on the benchmark as a whole (τ_b=0.02). That last result is the puzzling one: ESMI's emotion term alone does agree with people, yet adding its preservation term erases that signal, which points at the composition rule and not only at the components.

As our first contribution, we isolate the composition rule from the components that feed it. An additive composition lets one axis carry the score alone: an identity edit and a content swap both score positively, though the raters' scale counts both as failures. Figure 1 shows two such content swaps that the emotion-only metrics and ESMI all rank in their top quarter. Holding ESMI's own two components fixed and switching only the operator from additive to multiplicative makes it reject identity edits; with the components we propose below, adding instead of multiplying costs about a third of the agreement on both collection waves. Preservation must *gate*, not *add*.

We therefore propose PAEF (Perceived Affective Edit Fidelity), a simple parameter-free composition that meets both requirements: $\mathrm{PAEF}(I_o, c, I_e) = \frac{1}{2}\big(\mathrm{DA} + \mathrm{ES}\big) \cdot C$. It fuses a geometric and a categorical view of the edit's emotional shift, the *direction alignment* DA along the target emotion's valence–arousal (V/A) direction and the change ES in an emotion classifier's probability for the target, and gates their average by the DINOv2 scene preservation C between the original and the

edit. PAEF agrees with human ratings better than each of the thirteen metrics we audit, with τ_b=0.22 where the best metric in current use reaches 0.17, and holds that level on a held-out collection wave.

We also benchmark PAEF ($\sim$0.24B parameters) against nine MLLM judges from 1B to 38B, a 7B judge fine-tuned on our ratings and four learned editing evaluators. Six of the nine judges, the fine-tuned judge and two of the four evaluators are more accurate than PAEF. PAEF's case is therefore cost, not accuracy: it needs no rated data and runs 3.7 to 61 times faster than the MLLM judges on the same GPU, and its top fifth holds no more human-rated failures than any judge's or evaluator's.

Finally, we release **AffectEdit** itself, a reliability-validated benchmark for affective-edit metrics spanning five editing models, three image sources and eight Mikels emotions, together with its individual ratings. It lets a new metric or judge be checked against people before it is used to rank editors.

2 Related work

Representations of emotion. PAEF draws on two views of emotion. Russell's circumplex (Russell, 1980) places affect on continuous V/A axes, a structure replicated across self-report studies and populations (Posner et al., 2005; Watson & Tellegen, 1985) and adopted for visual emotion distributions (Yang et al., 2021); its continuity is what makes the direction of an edit definable. Editors and their classifiers instead work with the eight Mikels categories (Mikels et al., 2005). The NRC-VAD lexicon (Mohammad, 2025) links the two for words by giving each category a V/A coordinate: DA measures movement toward it, while ES keeps the categorical view.

Affective image editing. Emotion-conditioned generators and editors (Dang et al., 2025; Mao et al., 2026; Yang et al., 2024; 2025; Yin et al., 2026; Zhang et al., 2026) score emotion with classifiers, distribution divergences or V/A distances, and report preservation, if at all, in a separate column. Make Me Happier (Lin et al., 2025) is the exception: its ESMI adds emotion improvement to structural preservation, a composition that Sec. 4.2 finds non-significant on AffectEdit as a whole. Two ingredients of PAEF have precedents of their own. Shifts are scored in categorical space by EmoEdit's Emo-S, EmoAgent's shift ratio and the concurrent AffectDelta (Zhan et al., 2026), and as signed V/A change toward one fixed reference by regressor-guided editing (Gebhardt et al., 2025); EmotiCrafter and MooD work in continuous V/A space but score the distance to a fixed target, which, applied to an edit, depends on where the source started (Sec. A.1). Multiplicative gating appears in FED-Bench (Xue et al., 2026), which multiplies fidelity, alignment and expression gain so that a "lazy" facial-expression edit cannot score well. PAEF combines the two for scene-level edits: a direction in continuous V/A space, measured from the source and gated by scene preservation.

Human judgments of affect. Attributed affect has been annotated at scale for single images, as normative V/A ratings (Kurdi et al., 2017; Lang et al., 2005) and as emotions with free-text explanations (Achlioptas et al., 2021; Haydarov et al., 2024; Mohamed et al., 2022b), but not for the change between an image and its edit. For edits, AIM-Bench (Chen et al., 2026) ranks 13 editors on 800 affective tasks by scoring the edited image against a fixed target and averaging it with consistency and aesthetic scores, without human ratings of the edits; TIEdit (Gao et al., 2026) collects expert ratings of general-purpose edits, one of whose eight tasks modifies emotion; and the editors' own human studies are forced-choice comparisons, of which AIF-D (Zhang et al., 2025) is the only one we found to report rater agreement. Concurrently, Li et al. (2026) find that Emo-A and Emo-S track pairwise preferences for their editor only weakly. We are not aware of earlier absolute ratings of affective edits with reported inter-rater reliability; AffectEdit provides them and extends the audit to thirteen metrics and five editors.

Image editing evaluation. Low-level metrics such as SSIM (Wang et al., 2004) and LPIPS (Zhang et al., 2018) penalize large changes regardless of intent, which is backwards for affective editing, where a successful edit has to change the image. CLIPScore (Hessel et al., 2021) measures text-image alignment without psychological grounding, and learned preference and reward models (Kirstain et al., 2023; Luo et al., 2026; Ma et al., 2025; Wu et al., 2025b; 2023; Xu et al., 2023) and MLLM-based scores such as VIEScore (Ku et al., 2024) predict general preference or instruction fidelity rather than emotional direction. Recent benchmarks decompose editing quality into interpretable factors (Baraldi et al., 2025; Ma et al., 2024; Ryu et al., 2025; Wu et al., 2026); the most recent (Liu et al., 2026) defines twelve, spanning preservation, edit quality and instruction fidelity, none of which

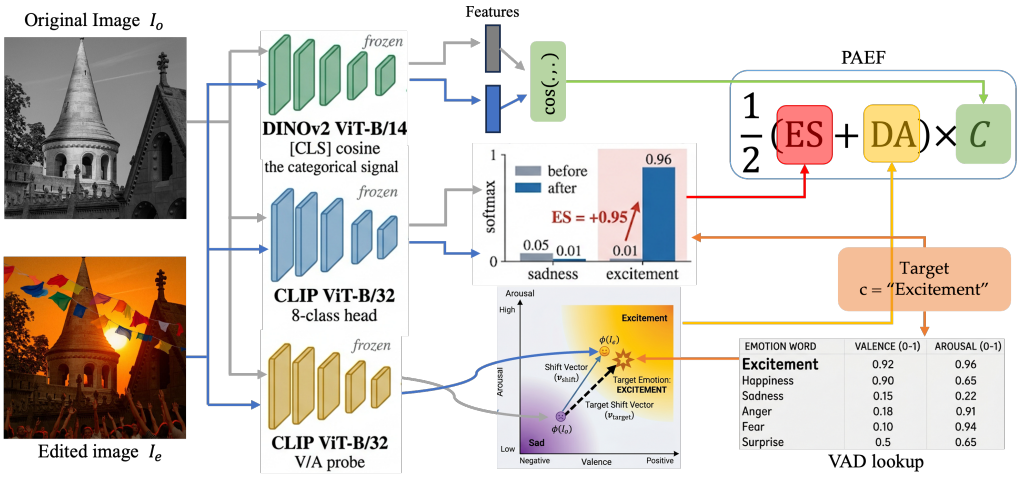

Figure 2: **How PAEF scores an edit**, shown schematically on one AffectEdit pair; the numbers and lookup rows are illustrative (Tab. 4 lists the coordinates), and DA projects onto the target's fixed unit direction, not toward a point. Two frozen backbones run on (I_o, I_e), CLIP ViT-B/32 and DINOv2-B/14, and nothing is trained. **ES** is the change in the EmoEdit classifier's softmax on the target class (Eq. (3)). **DA** projects the shift between the two predicted V/A positions onto the target's NRC-VAD direction (Eq. (4)). C is the cosine of the two DINOv2 `[CLS]` embeddings (Eq. (2)). C multiplies rather than adds: $C \to 0$ drives the score to zero however well the emotion fits, and DA$\approx$ES$\approx$0 drives it to zero however well the scene is preserved.

captures whether an edit moved perceived valence or arousal in the instructed direction. MLLMs are increasingly used as judges (Chen et al., 2024), and affect-tuned MLLMs such as EmoVIT (Xie et al., 2024) recognise emotion in a single image; we evaluate both as baselines, since neither is designed to judge a change between two images.

3 Preservation-gated equal-weight composition

PAEF asks two questions of an edit and multiplies the answers: did the edit move the image toward the target emotion, and did it keep the scene? Our key design choice is the multiplication. Preservation gates the emotion score instead of being added to it, so an edit scores well only when both answers are yes: an unchanged copy of the original scores zero, and a new picture that happens to carry the right emotion is pulled toward zero. Figure 2 shows the pieces on one AffectEdit pair; this section defines them.

The composition. Let I_o and I_e be the original and edited images and c the target, one of the eight Mikels emotions. PAEF multiplies an emotion shift E, the average of two signals defined below, by a scene-preservation gate C:

$$\mathrm{PAEF}(I_o, c, I_e) \;=\; \underbrace{\tfrac{1}{2}\big(\mathrm{DA}(I_o, c, I_e) + \mathrm{ES}(I_o, c, I_e)\big)}_{E\text{: emotion shift}} \cdot \underbrace{C(I_o, I_e)}_{\text{scene preservation}} \tag{1}$$

Every term comes from frozen, off-the-shelf backbones: no weight is fitted to AffectEdit's ratings and nothing is sampled, so the same pair always receives the same score, and every design choice below was fixed on the first of AffectEdit's two collection waves (Sec. 4).

Why multiplication, not addition? ESMI (Lin et al., 2025) adds its two terms, $\alpha S_{\text{str}} + (1-\alpha) S_{\text{emo}}$ with α=0.5, and so lets either compensate for the other: an identity edit with perfect preservation scores α, and a content swap with a strong emotion scores $1-\alpha$, though both are failures. In Eq. (1) an identity edit drives E to zero and a content swap drives C toward zero, and either one zeroes the score. Section 4.2 isolates the operator: on ESMI's own terms, multiplying instead of adding makes the metric reject identity edits and turns it significant, a gain carried by the held-out wave, and only

a gate that recognises a new scene also rejects content swaps. The operator rejects an unchanged image; the gate rejects a new one.

Scene preservation C. The gate needs an encoder that tracks scene identity, layout and objects, while staying largely blind to colour, lighting and atmosphere, the very changes an affective edit is meant to make. We use the cosine similarity of DINOv2 `[CLS]` embeddings,

$$C(I_o, I_e) \;=\; \frac{\phi(I_o)^\top \phi(I_e)}{\|\phi(I_o)\|\,\|\phi(I_e)\|}, \tag{2}$$

where ϕ is DINOv2-B/14 (Oquab et al., 2023). DINOv2's global features have exactly that blindness, which is a liability when the task is to detect appearance edits (Eldesokey et al., 2026) and an asset here.

Emotion shift E. An affective edit is a movement, not a destination, and the two are scored differently: the distance between the edited image and the target rewards pairs that began near the target, however the edit performed, while the displacement toward it does not depend on where the source started (Sec. A.1). We measure that displacement with two instruments. The *classifier shift* ES is the change in the EmoEdit classifier's (Yang et al., 2025) softmax probability for the target class, the Emo-S of the editing literature:

$$\mathrm{ES}(I_o, c, I_e) \;=\; s_c(I_e) - s_c(I_o). \tag{3}$$

The *direction alignment* DA projects the shift in predicted valence and arousal onto the target's direction in the circumplex:

$$\mathrm{DA}(I_o, c, I_e) \;=\; \mathrm{clip}\big(\Delta v \cdot u_v \;+\; \Delta a \cdot u_a,\; 0,\; 1\big), \quad (\Delta v, \Delta a) \;=\; \hat{\boldsymbol{v}}(I_e) - \hat{\boldsymbol{v}}(I_o), \tag{4}$$

where $\hat{\boldsymbol{v}}{=}(\hat{V}, \hat{A})$ are the valence and arousal heads released with FindingEmo (Mertens et al., 2024), applied to frozen CLIP ViT-B/32 features, and (u_v, u_a) is the unit direction of the target's NRC-VAD coordinate (Mohammad, 2025), listed in Sec. A. The lower clip scores a wrong-direction edit exactly zero instead of a negative value that the average could cancel against ES; the upper clip never binds on AffectEdit, where DA stays below 0.3. The two instruments, a categorical classifier and a continuous V/A regressor, agree only moderately ($\tau_b{=}0.31$), so their average is not one signal counted twice, and it dilutes either instrument's biases, such as the fire shortcut that Sec. 4.6 traces to ES. We weight them equally because nothing in the construct favours either, and a fitted weight would need rated data. The weights are equal in name only, since ES varies about six times as much as DA, yet Sec. 4.3 shows that the two do different jobs. Implementation details and costs are in Sec. A.

4 Evaluation on AffectEdit

AffectEdit comprises $N{=}1{,}406$ edit pairs from 3 image sources (OASIS-80, EmoPair-504, EmoEdit-416) and 5 editing models: AIF-Word ($n{=}100$; Affective Image Filter (Weng et al., 2023) driven by a one-word emotion template, Sec. C), EmoEdit (Yang et al., 2025) ($n{=}260$), EmoEditor (Lin et al., 2025) ($n{=}155$), FLUX.1 Kontext (Labs et al., 2025) ($n{=}446$) and Qwen-Image-Edit (Wu et al., 2025a) ($n{=}445$). Targets span the 8 Mikels emotions, 124 to 244 pairs each, and the sources show scenes, people, objects and animals (Fig. 4).

One joint rating. About five crowd workers (five for 1,271 of the pairs) rated each pair on one 0–4 scale that asks how well the edit made the image feel more like the target *while keeping the same scene*; the mean rating has $\mathrm{ICC}(2, 5){=}0.824$ (Sec. D). We did not collect separate emotion and preservation ratings, on purpose: scoring the two axes apart and combining them afterwards is the construction this paper diagnoses, and separate ratings would leave no joint judgment to validate a metric against. Raters applied both criteria: with either the emotion shift E or the gate C held fixed, the human score still rises with the other (Sec. D).

Two collection waves. Every design choice in Sec. 3, from the operator and the gate to the two signals and their equal weights, was made on the first of two collection waves (462 pairs) without looking at how any variant correlated with the second wave's ratings, so the second (944 pairs), the *held-out wave*, informed none of them. It shares the first wave's rating protocol, image sources and four of its five editors (AIF-Word's 100 pairs were collected only in the second wave), so it tests PAEF on new pairs rather than on a new distribution; its raters agree somewhat less ($\mathrm{ICC}(2, 5)$ 0.785 against 0.880).

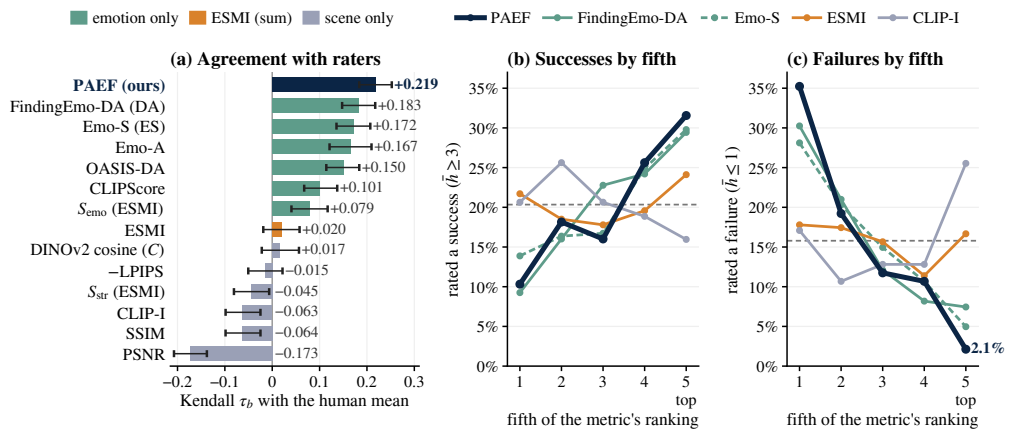

Figure 3: **Scene-only metrics do not track raters, emotion-only metrics track them modestly, and PAEF tracks them best** (N=1,406). **(a)** Kendall τ_b with the per-pair human mean and 95% cluster-bootstrap intervals (Sec. E); DA, ES and C in parentheses mark PAEF's own signals scored alone, OASIS-DA is DA with a V/A probe we fit on OASIS, and LPIPS is negated so that, like every similarity here, higher means closer to the original (Sec. A). **(b, c)** Share of successes and failures in each fifth of a metric's ranking (dashed: base rate). Judges: Tab. 3; every metric: Tab. 13.

Protocol. We measure a metric's agreement with people as Kendall's τ_b (Kendall, 1945) between its scores and the per-pair human mean. Intervals come from a cluster bootstrap over the 731 source images (B=2,000); one draw with a fixed seed serves every table and figure in this section except the probes and the fine-tuned judge (Sec. E), so the comparisons are paired, and P is the share of resamples in which PAEF scores higher; PAEF is deterministic, so its intervals reflect the sampled images, not a seed. A metric's top fifth is its 282 best-scored pairs; a failure is a pair rated $\bar{h}\leq 1$ (15.8% of all pairs) and a success one rated $\bar{h}\geq 3$ (20.3%).

4.1 Do the metrics in current use agree with raters?

Figure 3 ranks the thirteen metrics we audit with PAEF; three are its own signals scored alone: DA (FindingEmo-DA), ES (Emo-S) and C (DINOv2 cosine). **Scene-only metrics** carry no signal or a negative one, and their top fifths hold failures at or above the base rate; **emotion-only metrics** agree with people only modestly (Sec. H). **ESMI**, which scores both, is not significant on the benchmark as a whole, although its emotion term alone is, and its top fifth holds both kinds of failure. Two probes kept out of the benchmark show the same patterns without PAEF's signals (Tab. 2): at the bar where each metric accepts as many genuine edits as the raters did, every emotion-only metric accepts 26 to 39% of 62 content swaps (an original beside another image's edit) that raters rejected, every scene-only metric all 200 identity edits, and ESMI both; PAEF rejects both by construction. No affective-editing metric in current use rejects both kinds of failure, and the best of them, Emo-S, reaches only τ_b=0.17.

4.2 Is the composition rule to blame?

Table 1 builds PAEF from ESMI one change at a time. ESMI adds its emotion and structure terms (config A), so an unchanged image passes on structure alone. Multiplying them (config B) scores such an edit zero, rejects identity edits and makes the correlation with raters significant on the benchmark as a whole, a gain carried by the held-out wave; on the first wave both forms are near zero. Config B still accepts content swaps, because neither of its terms recognises a new scene (Tab. 19); replacing the two terms by the classifier shift ES and the DINOv2 gate (config C) rejects the swaps, cuts the failures in the top fifth from 16.7% to 2.1%, and raises the agreement by 0.25 on the first wave but only 0.02 on the held-out wave. Averaging DA into the emotion term (config D, PAEF) raises the agreement again, most on the held-out wave, and keeps the top fifth as clean.

Which multiplicative operator is used matters less than avoiding addition. On PAEF's terms, adding lowers τ_b by about a third and scores identity edits above most genuine edits (Tab. 1, last row), and

Table 1: **From ESMI to PAEF, one change at a time**; the bottom rows change one part of PAEF. τ_b with the human mean on each wave (grey: 95% interval includes zero); P: share of resamples in which PAEF scores higher on all pairs; Fail.: % of failures in the top fifth; Swaps, Identity: % of 62 content swaps and 200 identity edits accepted at the matched bar (✓/×: passes/fails, Sec. I; [a]as in Tab. 2).

Config	First	Held-out	All	P	Fail.	Swaps	Identity
A ESMI, $\frac{1}{2}(S_{emo}+S_{str})$	−0.038	+0.081	+0.020	>0.99	16.7	21 ×	100 ×
B × for +: $S_{emo}\cdot S_{str}$	−0.051	+0.166	+0.066	>0.99	16.7	24 ×	0 ✓
C PAEF's terms: ES·C	+0.198	+0.185	+0.184	>0.99	2.1	0 ✓	0 ✓
D DA averaged in: PAEF	+0.213	+0.229	+0.219	—	2.1	0 ✓	0 ✓
PAEF *with one part changed*							
no gate: E	+0.194	+0.215	+0.203	>0.99	4.3	34 ×	0 ✓
no ES: DA·C	+0.193	+0.213	+0.204	0.82	5.3	0 ✓	0 ✓
+ for ×: $\frac{1}{2}(E+C)$	+0.140	+0.156	+0.148	>0.99	3.5	0 ✓	0[a] ×

Table 2: **Probe items accepted** (%) at each metric's matched bar (Sec. I): 62 content swaps that raters rejected, 200 identity edits. ✓/×: passes/fails. [a]Scores identity edits almost as high as genuine ones. [b]All but ESMI's structure term pass.

Metric	Swaps	Identity
PAEF	0 ✓	0 ✓
Emotion-only (6)	26–39 ×	0–8 ✓
Scene-only (6)[b]	0–16	100 ×
ESMI	21 ×	100 ×
EditReward[a]	3 ✓	0 ×
HPSv3	19 ×	14 ✓
EditScore	0 ✓	0 ✓

across ten operators and five rescalings addition is always among the bottom three (Sec. H). The sweep does not single out the product: after min–max rescaling the minimum, harmonic mean and Hamacher product score significantly higher than the product on that scale (Holm-corrected), though none above 0.220; we keep the product because it needs no common scale for two signals from different instruments, not because it wins. Multiplication is what rejects an unchanged image, and a gate that recognises a new scene is what rejects a content swap; on ESMI's terms, however, the product helps only on the held-out wave.

4.3 What does each part of PAEF contribute?

The bottom rows of Tab. 1 remove one part of PAEF at a time. **The gate rejects new scenes.** Without it, PAEF accepts a third of the content swaps, and both its agreement with raters and the cleanness of its top fifth fall, each in over 99% of resamples. The gate acts as a veto rather than a fine adjustment: for the 56% of pairs with $C \geq 0.8$ PAEF ranks edits as E alone does (τ_b 0.215 vs. 0.216), and its gain comes from the 14% whose scene is replaced ($C < 0.5$; 0.235 vs. 0.185); every content swap has $C \leq 0.25$. Weakening the gate to C^γ with $\gamma < 1$ lowers τ_b steadily, which it would not if the gate over-penalised the scene change an edit needs, and strengthening it changes τ_b by less than 0.003 (Sec. H). **DA carries the ranking.** Without it (ES·C), τ_b drops from 0.219 to 0.184 overall and from 0.229 to 0.185 on the held-out wave, in over 99% of resamples, while the top fifth stays as clean. **ES keeps the top clean.** Without it, the agreement is level (PAEF higher in only 82% of resamples), but the share of failures in the top fifth rises from 2.1% to 5.3% (in over 99%). Pooling matters less. Of the combinations tried after the design was frozen, the minimum, product and geometric mean lower the held-out agreement, and the maximum raises it from 0.229 to 0.239, beating the average in 99% of resamples (95% on all pairs); we keep the average, fixed before the held-out wave was examined (Sec. H). Each part does a different job, but ES adds nothing to the overall agreement.

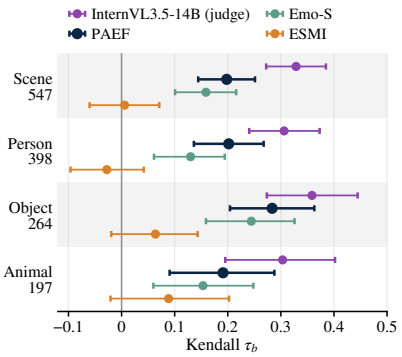

Figure 4: **Agreement by content category** (pairs under each name), with 95% intervals; raters agree about equally well in all four (ICC(2, 5) 0.81–0.84).

4.4 Does PAEF hold on edits it was not designed on?

On the held-out wave, which informed no design choice, PAEF agrees with raters at least as well as on the first (0.229 against 0.213; Tab. 1), although the raters agree with each other less there. Over all pairs, it is positive on all 16 editor, source and emotion slices, with an interval above zero on 15, but within three of the five editors, FLUX.1 Kontext and Qwen-Image-Edit among them, it reaches less than half its overall agreement (Sec. B), and it holds in each content category (Fig. 4). From

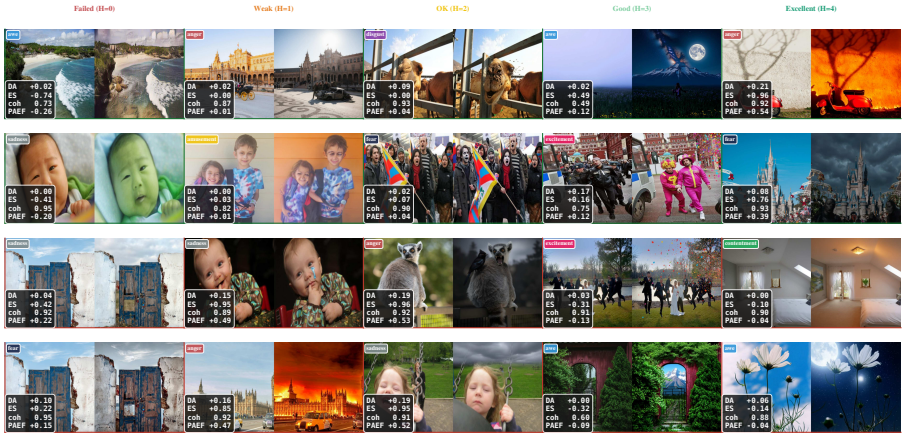

Figure 5: **PAEF across the human-rating range.** Columns: rounded human mean H; cells: original, edit, target and PAEF's components (coh is C). Per column, the two pairs whose PAEF percentile is closest to $25H$ (top rows) and the two farthest from it (bottom rows): PAEF's errors, false positives at $H \leq 1$ with a high ES and false negatives at $H \geq 3$ with a negative one (Sec. 4.6).

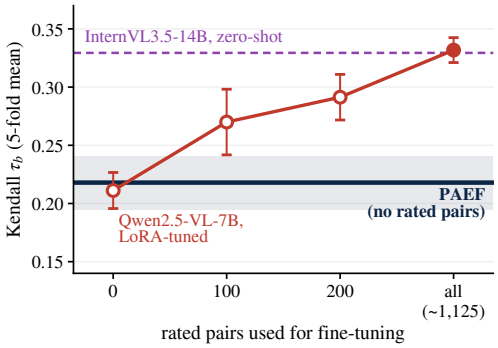

Figure 6: **A judge fine-tuned on our ratings** (Qwen2.5-VL-7B, LoRA) by rated pairs used, mean and s.d. over five folds, against PAEF (band: ± 1 s.d.) and zero-shot InternVL3.5-14B. Filled: lead over PAEF significant.

Table 3: **Cost and agreement of learned judges and evaluators.** Seconds per pair and peak GPU memory on one A100-80GB, batch 1 (Sec. A.5); succ., fail.: % of successes and failures among the 282 top-scored pairs (VIEScore: of its 1,209 parsed). Untimed judges: Tab. 5.

Metric	s/pair	GB	τ_b	succ.	fail.
PAEF (0.24B)	0.059	2.1	+0.219	31.6	2.1
Zero-shot MLLM judges					
InternVL3.5-2B	1.3	5.1	+0.058	21.4	15.4
InternVL3.5-4B	1.5	9.7	+0.297	28.9	8.1
InternVL3.5-14B	1.8	29	+0.329	35.9	5.2
InternVL3.5-38B	3.6	73	+0.318	32.2	6.5
Qwen2.5-VL-7B	1.3	16	+0.200	26.8	10.4
Qwen2.5-VL-32B	3.2	64	+0.299	38.7	3.8
Qwen3-VL-32B	3.3	63	+0.339	30.0	5.2
EmoVIT recipe on Qwen2.5-VL-7B, one runner					
affect-tuned	1.2	18	+0.241	28.6	8.8
untuned	—	—	+0.214	27.8	9.2
General-purpose learned evaluators					
EditReward	0.37	16	+0.283	36.2	4.3
HPSv3	0.31	18	+0.253	33.3	7.1
EditScore	4.7	17	+0.227	30.1	8.4
VIEScore	5.8	18	+0.163	25.5	3.8

the bottom fifth of its ranking to the top, the share of human-rated failures falls monotonically from 35% to 2.1% (Fig. 3c), and Fig. 5 shows pairs from every rating level. PAEF holds its level on new pairs from the benchmark's own distribution; whether it holds for other editors, raters and cultures is untested.

4.5 PAEF OR A LEARNED JUDGE?

We compare PAEF with nine zero-shot MLLM judges, a judge fine-tuned on our ratings, an affect-tuned judge and four learned evaluators (Tab. 3). Every model gets the target emotion, as the raters did, but not the edit instruction. The judges share the raters' rubric with no prompt search and ran once each (a re-run moves τ_b by about 0.025); the evaluators use their official settings (Sec. A.5).
On accuracy PAEF loses. Six of the nine judges agree with raters better, each in at least 97% of resamples, the strongest, Qwen3-VL-32B (Bai et al., 2025b), by about 0.12 in τ_b; PAEF is level with Qwen2.5-VL-7B (Bai et al., 2025a) and above the 1B and 2B InternVL3.5 judges (Wang et al., 2025). Fine-tuned on our ratings with LoRA (Hu et al., 2021), Qwen2.5-VL-7B pulls ahead from

100 rated pairs on, significantly only with all $\sim$1,125 training pairs of each fold, where it reaches 0.332 ± 0.011 against PAEF's 0.218 on the same folds (mean and s.d. over five folds, one training run each; Fig. 6). Tuning a judge for affect does not reliably help (Tab. 3 and Sec. A.5). Among the evaluators, EditReward (Wu et al., 2025b) agrees with raters better than PAEF in over 99% of resamples and HPSv3 (Ma et al., 2025), not significantly, in 95%; EditScore (Luo et al., 2026) is level and VIEScore (Ku et al., 2024) behind. HPSv3's and EditScore's agreement comes largely from ranking the five editors as people do, where PAEF favours EmoEdit's edits (Secs. A.5 and 4.6). **What PAEF offers instead is cost.** It has about 0.24B parameters and needs no rated edits; on the same A100 at batch size 1 it takes 59 ms per pair and 2 GB of memory, against 1.3 to 3.6 s and 5 to 73 GB for the timed zero-shot judges and 0.3 to 5.8 s and 16 to 18 GB for the evaluators. Read from one forward pass, the untrained 7B judge is level with PAEF at 0.22 s per pair, and the fine-tuned one takes 0.23 s after a 1.9 GPU-hour training sweep. No model was optimised for speed; batched, PAEF takes 31 ms per pair. On the probes, EditReward and HPSv3 each fail one, and the two evaluators that pass both, EditScore and VIEScore, are 79 to 98 times slower than PAEF and at best level with it (Tabs. 2 and 19). PAEF's top fifth holds no more human-rated failures (2.1%, six pairs) than that of any audited metric, judge or evaluator, one fewer than the fine-tuned judge's, but against the next three it is lower in only 92 to 96% of resamples, and the best judges' top fifths hold more successes (Tabs. 3 and 16). Where accuracy matters and a GPU budget exists, a judge is the better choice; PAEF suits scoring in bulk or without rated data.

4.6 Where does PAEF fail?

Where it is weak. Its agreement with humans is lowest on the largest third of edits and on edits whose colour change already agrees with the target, where it partly follows the colour cue; on a small set of edits that add balloons for amusement it carries no signal while a judge does (Sec. J). We detect no difference on images with faces, and a global V/A probe cannot localise one. **Biases it inherits.** PAEF over-rewards added fire when the target is anger or fear, an excess that comes from ES, not DA, and concentrates in EmoEdit edits; no other cue we tested shows an excess that survives correction for multiple comparisons (Sec. J). The same classifier, which selected EmoEdit's training pairs (Yang et al., 2025), makes PAEF rank EmoEdit first among the five editors, where people, $\mathrm{DA}{\cdot}C$ and every learned evaluator rank it third, so to compare editors built with and without it we recommend $\mathrm{DA}{\cdot}C$, which agrees with raters about as well and ranks the editors as they do (Tab. 12). **Its errors.** The six failures in PAEF's top fifth have a large ES (0.42–0.97) and a DA like its successes' (median 0.10 vs. 0.11), and 22 of the 30 successes in its bottom fifth have a negative ES, on sources the classifier already rated high on the target (median 0.43 vs. 0.002 overall; Figs. 5 and 9). Both kinds of error trace to the classifier term ES, not to the composition.

5 Conclusions

On AffectEdit, $1{,}406$ edits with reliable joint ratings, each of the thirteen metrics we audited that is in current use misses at least one of the two ways an affective edit fails: emotion-only metrics accept content swaps, scene-only metrics accept unchanged images, and ESMI, which adds the two, accepts both. Multiplying ESMI's terms makes it reject unchanged images, and a scene gate makes it reject content swaps too. PAEF, the gated average of a classifier shift and a valence–arousal direction signal, agrees with raters better than every metric we audit, holds that level on a held-out wave and by construction rejects both failures. The strongest MLLM judges, a judge fine-tuned on our ratings and the strongest editing reward model agree with raters better still; PAEF is cheaper than all of them and, unlike the fine-tuned judge, needs no rated data.

Limitations and future work. All evidence comes from one benchmark: its held-out wave shares the first wave's protocol, sources and four of its five editors, its $1{,}406$ pairs cover a small part of the space of scenes, edits, intensities and cultures, and newer editors (Team et al., 2025; 2026) are not in it. PAEF inherits its classifier's biases, and edits that change only a facial expression are untested (Sec. 4.6). Three steps would test or remove these limits: a benchmark collected independently of PAEF, with new editors, raters and cultures; an emotion classifier that no editor was trained against, in place of ES; and PAEF's gate as a cheap veto in front of an MLLM judge, which is untested.

ETHICS STATEMENT

Human subjects. The $1{,}406$ edit pairs in AffectEdit carry ratings from crowdworkers recruited on Amazon Mechanical Turk. Raters judged ordinary photographs and their edits toward eight emotions, four of them negative; no personal information was collected beyond the platform-assigned identifiers needed to pay and de-duplicate raters. Workers were paid $\$0.65$ per HIT, $\sim\$13$/hour at the median completion time, above the U.S. federal minimum wage. The study was conducted under a standing institutional ethics approval covering crowdsourced collection of human responses to visual stimuli; the issuing committee and protocol number are withheld for anonymity and will be named in the camera-ready. Rater identifiers are replaced by salted hashes in the public release, so no platform account can be recovered from the distributed data.

What PAEF measures. PAEF scores the affect a typical viewer reads from an edited image. It does not measure the felt affect of any individual viewer or the mental state of any depicted person, and using it as either is a misuse.

Whose emotions. The Mikels-8 taxonomy was developed with American undergraduate participants, the NRC-VAD lexicon is English-language, and the CLIP and DINOv2 backbones are trained on web corpora with documented over-representation of Western imagery. Our raters were drawn from English-speaking countries, and affective responses to the same artwork are known to differ across languages and cultures (Mohamed et al., 2022a; 2024). PAEF therefore measures perceived affect *as read by this population*, and we make no claim that its judgments transfer across cultures without revalidation.

Inherited biases and a known shortcut. CLIP, DINOv2 and the EmoEdit classifier each carry representational biases that propagate into the metric, and we add no safeguards beyond those of the released backbones. Our audit of the metric against the human ratings found one shortcut that survives correction for multiple comparisons: PAEF over-rewards added fire when the target is anger or fear, an artifact inherited from the EmoSet-trained classifier that supplies the ES signal. We report the effect size and the analysis that isolates it rather than leaving it for readers to discover.

Dual use. PAEF and AffectEdit are evaluation infrastructure and produce no images, but a parameter-free, deterministic affective-edit score lowers the cost of optimising generators toward emotionally manipulative imagery: propaganda, deceptive advertising, or content built to elicit fear or outrage. We release no generative weights and judge the risk low relative to the benefit of a reliable evaluation standard, but we state it.

Release. Our ratings and scores are released under CC BY 4.0, the images under their sources' and editors' terms (originals whose redistribution is unclear are restored from their source rather than redistributed), with salted-hash rater identifiers and a datasheet (Sec. C). Section G treats beneficial uses, foreseeable harms, worker welfare and open questions at length.

REPRODUCIBILITY STATEMENT

Every number in this paper is computed by released scripts from released outputs. Section 3 defines PAEF; Sec. A gives the backbones, checkpoints, preprocessing and NRC-VAD coordinates, and Sec. A.5 the judge prompts, decoding settings and fine-tuning protocol. Sections C and D document the data, the annotation protocol, the rating interface, the restored second-round ratings and the reliability computation; first-round raters are identified only by their position within each pair, which is all the reported reliability statistics need. Sections H and J record the analyses, whose protocols were fixed before they were run, except those marked post hoc or exploratory. The analysis scripts named in the appendix will be released with the final version. The supplementary material contains the PAEF package with a demo notebook, the per-pair scores of every metric and judge, the individual ratings with anonymised rater identifiers, and a script that recomputes the point estimates of every table and the reliability values from them.

AI USE STATEMENT

In this work, we used generative AI tools for the following tasks with required disclosure. **Generating synthetic data:** the edited images in AffectEdit are produced by the five generative image editors

we evaluate, and the edit instructions for two of them were written by InternVL3.5-14B (Sec. A.4). **Implementing methods** and **cleaning and reformatting data:** LLM coding assistants wrote much of the analysis code, including the scripts that compute every reported statistic and generate the tables and figures, and merged per-pair outputs into the released tables. **Designing experiments, proposing hypotheses and interpreting results:** LLM assistants proposed hypotheses and drafted analysis protocols, which the authors fixed before the analyses ran, and drafted interpretations, which the authors checked against the numbers. We did not use generative AI tools to develop theoretical models, to formulate or prove mathematical claims, or for translation, and qualitative or thematic data analysis is not applicable to this work. Additionally, we used generative AI tools for tasks with recommended disclosure: drafting and editing parts of the paper, creating figures, editing software code, identifying and summarising related work, and formatting and verifying references. We have reviewed all AI-assisted work and re-derived every reported number from released outputs with scripts we inspected. No generative AI component is part of PAEF itself; the MLLM judges of Sec. 4 are baselines. We take responsibility for the final content of this work, including text, claims and artifacts produced with the aid of generative AI.

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. 11569–11579, 2021.

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

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

Lorenzo Baraldi, Davide Bucciarelli, Federico Betti, Marcella Cornia, Nicu Sebe, and Rita Cucchiara. What changed? detecting and evaluating instruction-guided image edits with multimodal large language models. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16217–16226. IEEE, 2025.

Lisa Feldman Barrett. The theory of constructed emotion: an active inference account of interoception and categorization. *Social cognitive and affective neuroscience*, 12(1):1–23, 2017.

Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18392–18402. IEEE, 2023.

Dongping Chen, Ruoxi Chen, Shilin Zhang, Yinuo Liu, Yaochen Wang, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. Mllm-as-a-judge: Assessing multimodal llm-as-a-judge with vision-language benchmark. *arXiv preprint arXiv:2402.04788*, 2024.

Shi Chen, Xuecheng Wu, Heli Sun, Yunyun Shi, Xinyi Yin, Fengjian Xue, Jinheng Xie, Dingkang Yang, Hao Wang, Junxiao Xue, et al. Aim-bench: Benchmarking and improving affective image manipulation via fine-grained hierarchical control. *arXiv preprint arXiv:2604.10454*, 2026.

Domenic V Cicchetti. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological assessment*, 6(4):284, 1994.

Shengqi Dang, Yi He, Long Ling, Ziqing Qian, Nanxuan Zhao, and Nan Cao. Emoticrafter: Text-to-emotional-image generation based on valence-arousal model. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15218–15228. IEEE, 2025.

Abdelrahman Eldesokey, Aleksandar Cvejić, Bernard Ghanem, and Peter Wonka. Mind-the-glitch: Visual correspondence for detecting inconsistencies in subject-driven generation. *Advances in Neural Information Processing Systems*, 38:3226–3245, 2026.

Shiqi Gao, Zitong Xu, Kang Fu, Huiyu Duan, Xiongkuo Min, and Jia Wang. Evaluating image editing with llms: A comprehensive benchmark and intermediate-layer probing approach. *Displays*, pp. 103494, 2026.

Christoph Gebhardt, Robin Willardt, Seyedmorteza Sadat, Chih-Wei Ning, Andreas Brombach, Jie Song, Otmar Hilliges, and Christian Holz. Regressor-guided image editing shifts emotion and disengagement timing in social media. *arXiv preprint arXiv:2501.12289*, 2025.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12): 86–92, 2021.

Kotaro Hara, Abigail Adams, Kristy Milland, Saiph Savage, Chris Callison-Burch, and Jeffrey P Bigham. A data-driven analysis of workers' earnings on amazon mechanical turk. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, pp. 1–14, 2018.

Kilichbek Haydarov, Xiaoqian Shen, Avinash Madasu, Mahmoud Salem, Li-Jia Li, Gamaleldin Elsayed, and Mohamed Elhoseiny. Affective visual dialog: A large-scale benchmark for emotional reasoning based on visually grounded conversations. In *ECCV*, 2024.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pp. 7514–7528, 2021.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pp. 65–70, 1979.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Maurice G Kendall. The treatment of ties in ranking problems. *Biometrika*, 33(3):239–251, 1945.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in neural information processing systems*, 36:36652–36663, 2023.

Max Ku, Dongfu Jiang, Cong Wei, Xiang Yue, and Wenhu Chen. Viescore: Towards explainable metrics for conditional image synthesis evaluation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12268–12290, 2024.

Benedek Kurdi, Shayn Lozano, and Mahzarin R Banaji. Introducing the open affective standardized image set (OASIS). *Behavior Research Methods*, 49(2):457–470, 2017.

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

Peter J Lang, Margaret M Bradley, Bruce N Cuthbert, et al. *International affective picture system (IAPS): Affective ratings of pictures and instruction manual*. NIMH, Center for the Study of Emotion & Attention Gainesville, FL, 2005.

Qing Li, Zeyu Dong, Yin Cui, Chuan Yan, and Xiaojiang Peng. What can i edit? open-ended strategy discovery and the emotion editability landscape. *arXiv preprint arXiv:2607.23920*, 2026.

Qing Lin, Jingfeng Zhang, Yew-Soon Ong, and Mengmi Zhang. Make me happier: Evoking emotions through image diffusion models. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16367–16376. IEEE, 2025.

Runzhou Liu, Hailey Weingord, Sejal Mittal, Prakhar Dungarwal, Anusha Nandula, Bo Ni, Samyadeep Basu, Hongjie Chen, Nesreen K Ahmed, Li Li, et al. Human-aligned mllm judges for fine-grained image editing evaluation: A benchmark, framework, and analysis. *arXiv preprint arXiv:2602.13028*, 2026.

Xin Luo, Jiahao Wang, Chenyuan Wu, Shitao Xiao, Xiyan Jiang, Defu Lian, Jiajun Zhang, Dong Liu, and Zheng Liu. Editscore: Unlocking online rl for image editing via high-fidelity reward modeling. In *International Conference on Learning Representations*, volume 2026, pp. 33027–33056, 2026.

Yiwei Ma, Jiayi Ji, Ke Ye, Weihuang Lin, Zhibin Wang, Yonghan Zheng, Qiang Zhou, Xiaoshuai Sun, and Rongrong Ji. I2ebench: A comprehensive benchmark for instruction-based image editing. *Advances in Neural Information Processing Systems*, 37:41494–41516, 2024.

Yuhang Ma, Xiaoshi Wu, Keqiang Sun, and Hongsheng Li. Hpsv3: Towards wide-spectrum human preference score. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15086–15095. IEEE, 2025.

Qi Mao, Haobo Hu, Yujie He, Difei Gao, Haokun Chen, and Libiao Jin. Emoagent: A multi-agent framework for diverse affective image manipulation. *IEEE Transactions on Affective Computing*, 2026.

Laurent Mertens, Elahe' Yargholi, Hans O de Beeck, Jan Van den Stock, and Joost Vennekens. Findingemo: An image dataset for emotion recognition in the wild. *Advances in neural information processing systems*, 37:4956–4996, 2024.

Joseph A Mikels, Barbara L Fredrickson, Gregory R Larkin, Casey M Lindberg, Sam J Maglio, and Patricia A Reuter-Lorenz. Emotional category data on images from the International Affective Picture System. *Behavior Research Methods*, 37(4):626–630, 2005.

Youssef Mohamed, Mohamed Abdelfattah, Shyma Alhuwaider, Feifan Li, Xiangliang Zhang, Kenneth Church, and Mohamed Elhoseiny. ArtELingo: A million emotion annotations of WikiArt with emphasis on diversity over language and culture. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 8770–8785, Abu Dhabi, United Arab Emirates, December 2022a. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.600. URL https://aclanthology.org/2022.emnlp-main.600/.

Youssef Mohamed, Faizan Farooq Khan, Kilichbek Haydarov, and Mohamed Elhoseiny. It is okay to not be okay: Overcoming emotional bias in affective image captioning by contrastive data collection. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21231–21240. IEEE, 2022b.

Youssef Mohamed, Runjia Li, Ibrahim Said Ahmad, Kilichbek Haydarov, Philip Torr, Kenneth Church, and Mohamed Elhoseiny. No culture left behind: ArtELingo-28, a benchmark of WikiArt with captions in 28 languages. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 20939–20962, 2024.

Saif M Mohammad. Nrc vad lexicon v2: Norms for valence, arousal, and dominance for over 55k english terms. *arXiv preprint arXiv:2503.23547*, 2025.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Jonathan Posner, James A Russell, and Bradley S Peterson. The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and psychopathology*, 17(3):715–734, 2005.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

James A Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6): 1161, 1980.

Suho Ryu, Kihyun Kim, Eugene Baek, Dongsoo Shin, and Joonseok Lee. Towards scalable human-aligned benchmark for text-guided image editing. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18292–18301. IEEE, 2025.

Patrick E Shrout and Joseph L Fleiss. Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin*, 86(2):420, 1979.

Meituan LongCat Team, Hanghang Ma, Haoxian Tan, Jiale Huang, Junqiang Wu, Jun-Yan He, Lishuai Gao, Songlin Xiao, Xiaoming Wei, Xiaoqi Ma, et al. Longcat-image technical report. *arXiv preprint arXiv:2512.07584*, 2025.

Super Intelligence Team, Changhao Qiao, Chao Hui, Chen Li, Cunzheng Wang, Dejia Song, Jiale Zhang, Jing Li, Qiang Xiang, Runqi Wang, et al. Firered-image-edit-1.0 technical report. *arXiv preprint arXiv:2602.13344*, 2026.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

David Watson and Auke Tellegen. Toward a consensual structure of mood. *Psychological Bulletin*, 98(2):219, 1985.

Shuchen Weng, Peixuan Zhang, Zheng Chang, Xinlong Wang, Si Li, and Boxin Shi. Affective image filter: Reflecting emotions from text to images. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10776–10785. IEEE, 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.

Keming Wu, Sicong Jiang, Max Ku, Ping Nie, Minghao Liu, and Wenhu Chen. Editreward: A human-aligned reward model for instruction-guided image editing. *arXiv preprint arXiv:2509.26346*, 2025b.

Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023.

Yongliang Wu, Zonghui Li, Xinting Hu, Xinyu Ye, Xianfang Zeng, Gang Yu, Wenbo Zhu, Bernt Schiele, Ming-Hsuan Yang, and Xu Yang. Kris-bench: Benchmarking next-level intelligent image editing models. *Advances in Neural Information Processing Systems*, 38, 2026.

Hongxia Xie, Chu-Jun Peng, Yu-Wen Tseng, Hung-Jen Chen, Chan-Feng Hsu, Hong-Han Shuai, and Wen-Huang Cheng. Emovit: Revolutionizing emotion insights with visual instruction tuning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26586–26595. IEEE, 2024.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:15903–15935, 2023.

Fengjian Xue, Xuecheng Wu, Heli Sun, Yunyun Shi, Shi Chen, Liangyu Fu, Jinheng Xie, Dingkang Yang, Hao Wang, Junxiao Xue, et al. Fed-bench: A cross-granular benchmark for disentangled evaluation of facial expression editing. *arXiv preprint arXiv:2603.29697*, 2026.

Jingyuan Yang, Jie Li, Leida Li, Xiumei Wang, and Xinbo Gao. A circular-structured representation for visual emotion distribution learning. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4235–4244. IEEE, 2021.

Jingyuan Yang, Qirui Huang, Tingting Ding, Dani Lischinski, Danny Cohen-Or, and Hui Huang. Emoset: A large-scale visual emotion dataset with rich attributes. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 20383–20394, 2023.

Jingyuan Yang, Jiawei Feng, and Hui Huang. EmoGen: Emotional image content generation with text-to-image diffusion models. In *CVPR*, 2024.

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.

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

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.

Xingzu Zhan, Lin Gu, and Ruogu Fang. Affectdelta: Beyond emotion labels for image editing. *arXiv preprint arXiv:2609.02616*, 2026.

Peixuan Zhang, Shuchen Weng, Jiajun Tang, Si Li, and Boxin Shi. Towards deeper emotional reflection: Crafting affective image filters with generative priors. *IEEE transactions on pattern analysis and machine intelligence*, 2025.

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

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE/CVF conference on computer vision and pattern recognition*, pp. 586–595. IEEE, 2018.

A IMPLEMENTATION DETAILS

A.1 DIRECTION SIGNALS

Target-class softmax shift (ES). We use the EmoEdit (Yang et al., 2025) affective classifier (checkpoint `2024-02-23-best.pth` from the authors' repository): a 2-layer MLP head ($512 \to 256 \to 8$) on top of frozen CLIP-ViT-B/32 (Radford et al., 2021) `[CLS]` embeddings, trained on the EmoSet 8-class taxonomy (406,964 images). For pair (I_o, I_e) with target Mikels emotion c, we run both images through the classifier and compute $\mathrm{ES}(I_o, c, I_e) = s_c(I_e) - s_c(I_o)$, where $s_c(\cdot) \in [0, 1]$ is the softmax probability of class c.

V/A direction projection (DA). We use the valence and arousal heads released by the FindingEmo authors (Mertens et al., 2024), whose dataset annotates about $25{,}000$ in-the-wild images for valence and arousal: two linear layers ($512 \to 1$) on frozen CLIP-ViT-B/32 image features. We did not train or tune them. As an alternative we fit our own Ridge probe on OASIS (Kurdi et al., 2017) with the same features; as a direction signal it agrees less with human judgment on AffectEdit ($\tau_b = +0.150$ against $+0.183$ for the FindingEmo heads), plausibly because OASIS's lab-controlled images are less diverse than FindingEmo's.

The Mikels target c is mapped to a V/A coordinate $\boldsymbol{d}_c^*$ via the NRC-VAD lexicon (**?**), and DA projects onto the unit direction $(u_v, u_a) = \boldsymbol{d}_c^* / \|\boldsymbol{d}_c^*\|$. The full mapping is given in Tab. 4.

Table 4: **NRC-VAD coordinates and unit-direction vectors** for the 8 Mikels target emotions. $\boldsymbol{d}_c^*$ is the stored V/A coordinate and (u_v, u_a) the unit direction DA projects onto. Values are those in `paef/directions.py`, the table the released implementation uses.

Emotion	d_v^*	d_a^*	u_v	u_a
amusement	0.858	0.674	0.790	0.620
awe	–0.062	0.480	–0.130	0.990
contentment	0.750	0.220	0.960	0.280
excitement	0.792	0.368	0.910	0.420
anger	–0.666	0.730	–0.670	0.740
disgust	–0.896	0.550	–0.850	0.520
fear	–0.854	0.680	–0.780	0.620
sadness	–0.896	–0.424	–0.900	–0.430

Why displacement rather than distance to the target. An affective edit is a movement, not a destination, and the two are scored differently. Measuring how close the edited image lands to the target's V/A coordinate, as EmotiCrafter does (Dang et al., 2025), rewards pairs that began near the target and penalizes pairs that began far from it, however well or badly the edit itself performed. Displacement along the intended axis does not have that dependence on the starting point, and it is what a directional instruction actually asks for. Projecting the observed V/A shift onto the target's unit direction measures exactly that: zero when nothing moved, positive when the movement is toward the target, negative when it is away.

DINOv2 coherence (C). We use the publicly released `dinov2_vitb14` model (ViT-B/14, 86.6M parameters) (Oquab et al., 2023). Each image is resized on its shorter side and center-cropped to 224×224, passed through DINOv2, and the `[CLS]` token embedding (768-dim) is L2-normalized. The stored C behind every reported number used a shorter side of 256 for the first collection wave and 224 for the second. Recomputing every pair with one setting changes PAEF's τ_b from $+0.2185$ to $+0.2190$ (224) or $+0.2208$ (256); by wave it is $+0.214$ / $+0.229$ with 224 and $+0.213$ / $+0.232$ with 256, so the held-out wave is not favoured by the mix. The released code defaults to 224 and can reproduce the published column exactly. $C(I_o, I_e)$ is the cosine similarity between the two normalized embeddings, in $[-1, 1]$; on AffectEdit it is positive on all but 2 of the $1{,}406$ pairs.

ESMI and its multiplicative variant. ESMI (Lin et al., 2025) scores an edit as $\alpha\, S_{\text{str}} + (1-\alpha)\, S_{\text{emo}}$ with $\alpha = 0.5$, where S_{emo} is a KL-divergence-based emotion shift toward the target and S_{str} is an L_1

term over Grad-CAM-selected emotion-neutral regions plus an edge-map term. The multiplicative variant in Tab. 1 keeps both terms and replaces the weighted sum by the product $S_{\text{emo}} \cdot S_{\text{str}}$.

Pixel-level preservation metrics. SSIM (Wang et al., 2004), PSNR and LPIPS (Zhang et al., 2018) compare the original and the edit after both are resized to 512×512 (bicubic): SSIM with a Gaussian window (σ=1.5) averaged over the RGB channels, PSNR over RGB, and LPIPS with its AlexNet variant (v0.1). We negate LPIPS so that, like the other similarities, a higher score means the edit keeps more of the original. The protocol was fixed in `experiments/iclr_05_pixel_metrics.py` before the first run, and the three share the global bootstrap draw of Sec. E.

A.2 Computational cost

PAEF consists of two CLIP-ViT-B/32 forward passes (for ES and DA), one DINOv2-B/14 forward pass (for C), and arithmetic combination. Total parameter budget:

- CLIP-ViT-B/32 (shared across ES and DA): 151.3M parameters
- DINOv2-B/14: 86.6M parameters
- Classifier head (ES): 0.1M parameters
- V/A linear probes ($\times 2$): $<$0.001M parameters
- **Total: $\sim$238M parameters**

On one NVIDIA A100-80GB, PAEF takes 59 ms per pair at batch size 1 with image loading, the setting of Tab. 3; batched (32 pairs), it takes 31 ms per pair end to end and 6.1 ms of GPU forward time, so the whole benchmark needs under a minute. On the same GPU and pairs, the MLLM judges take 1.2 to 3.6 s per pair and the learned evaluators 0.3 to 5.8 s (Tab. 3), and the judges' sampled decoding makes repeated runs differ.

A.3 Software and dependencies

PAEF is implemented in PyTorch 2.2 with the OpenAI CLIP package and `torch.hub` for DINOv2. NRC-VAD lexicon v2.1 is loaded from the official release. EmoEdit classifier weights are loaded from the authors' public repository. Bootstrap and permutation tests use `scipy.stats` 1.11. We release PAEF as a PyTorch package with a command-line interface and a demo notebook, together with the per-pair scores behind every table and a script that recomputes the tables' point estimates from them (supplementary material).

A.4 Edit instruction generation

For each (source image, target emotion) pair in our dataset, we generate one image-specific natural-language editing instruction that downstream zero-shot image editors (Flux-Kontext, Qwen-Image-Edit) execute to shift the image toward the target affect. We use a vision-language model (VLM) rather than a text-only LLM because instructions must reference content actually present in each image to avoid generic edits.

Backbone and decoding. We use `OpenGVLab/InternVL3_5-14B` in bfloat16 on a single A100 with flash-attention 2. Generation parameters: `max_new_tokens=512`, `do_sample=True`, `temperature=0.7`, `top_p=0.9`. Sampling (rather than greedy decoding) was used because greedy decoding collapsed onto a few generic wordings across images; sampling at T=0.7 produces diverse but coherent suggestions.

Image preprocessing. We follow InternVL's standard dynamic-tile preprocessing: convert to RGB, choose the (i, j) tile grid with $1 \le i \cdot j \le 12$ whose aspect ratio best matches the image (preferring larger area when ties occur), resize each tile to 448×448, append a 448×448 thumbnail of the full image (`use_thumbnail=True`), and normalize with ImageNet statistics.

Prompt design. The prompt is single-image, single-turn, parameterized by the target emotion word. Four constraints are embedded to avoid pathologies we observed in early runs. **(R1) No filter-like**

Table 5: **MLLM judges across scale** (Kendall's τ_b against the human mean, all 1,406 pairs, zero-shot, identical rubric and side-by-side composites; only the checkpoint changes). Re-running a judge at the sampling settings used here moves its own τ_b by $\sim$0.025, so we order judges against PAEF but not against each other. The two affect-oriented judges are discussed below.

Judge	Scale	τ_b	Parsed	Distinct
InternVL3.5-1B	1 B	0.089	1,405	5
InternVL3.5-2B	2 B	0.058	1,406	7
PAEF (ours)	0.24 B	0.219	1,406	1,406
Qwen2.5-VL-7B	8 B	0.200	1,406	5
InternVL3.5-4B	4 B	0.297	1,406	5
InternVL3.5-8B	8 B	0.259	1,406	9
InternVL3.5-14B	14 B	0.329	1,406	8
Qwen2.5-VL-32B	32 B	0.299	1,406	8
InternVL3.5-38B	38 B	0.318	1,406	8
Qwen3-VL-32B	32 B	0.339	1,406	8
Affect-oriented				
EmoVIT (released weights)	7 B	0.091	1,369	5
EmoVIT-recipe Qwen2.5-VL-7B	8 B	0.241	1,406	6

edits (no fog, blur, desaturation, dimming, contrast/colour grading, vignette, lighting-only changes), since these were over-represented in early runs and produced visually weak targets when handed to the editor. **(R2) Be specific, not vague** (e.g., "Add a golden retriever running toward the man with a ball in its mouth" rather than "Add a playful element such as a child or dog"). **(R3) One concrete change per edit field**, reducing ambiguity for the downstream editor. **(R4) Edits must reference content actually present in the image**, which is what motivates the VLM over a text-only LLM.

Output schema and parsing. The prompt fixes a deterministic output schema: `SCENE_ANALYSIS: ...`, three pairs of `EDIT_n: ...` / `WHY_n: ...`, and `COMBINED_INSTRUCTION: ...`. Only `COMBINED_INSTRUCTION` is fed to the downstream editor; the per-edit fields are kept for diagnostics and ablations. A line-by-line parser strips Markdown bold markers and fills four buckets (`scene_analysis`, `edits`, `combined_instruction`, `raw_response`); unmatched lines are ignored, so partially malformed completions still yield usable output. "None"/"N/A" edits are dropped.

Run loop. For each row in our edit-targets file, the script resolves the source image, preprocesses it, builds the prompt with the target emotion, generates, parses, and stores the result keyed by `{theme}_{target_emotion}`. `torch.cuda.empty_cache()` between calls keeps VRAM steady; the JSON output is checkpointed after every image, and the script is resume-friendly via a `--resume` flag for SLURM re-queues after node failure.

A.5 MLLM-AS-JUDGE METHODOLOGY

Judges. We evaluate the InternVL3.5 family at 1, 2, 4, 8, 14 and 38B (Wang et al., 2025), Qwen2.5-VL-7B and -32B (Bai et al., 2025a), and Qwen3-VL-32B (Bai et al., 2025b) as zero-shot judges (Tab. 5). Every judge sees the original and edited image as a single side-by-side composite (height 512 px, white background, aspect ratios preserved) and the same rubric, which spells out the 0–4 anchors given to our human raters and their dual criterion of emotional impact *and* content preservation; only the checkpoint changes. The prompt asks for a one-line JSON reply, `{"score": <0-4>, "reasoning": "<one sentence>"}`. Decoding uses `do_sample=True`, `temperature=0.3` and `top_p=0.9`; InternVL uses its dynamic-tile preprocessing and Qwen its native dynamic-resolution processor. The parser takes the first JSON object with a `score` key and falls back to a regular expression; an out-of-range or missing score is recorded as missing. All judges parsed every pair except InternVL3.5-1B (1 failure) and EmoVIT (37). Every judge returns 5 to 9 distinct values over the 1,406 pairs, so a fixed-size selection of a judge's top-scored pairs is partly filled by tie-breaking; we break ties by exact expectation wherever

a top band is reported. Following the MLLM-as-a-judge literature (Chen et al., 2024), we treat these judges as strong but costly baselines rather than as ground truth. The timings in Tab. 3 come from one run with a single A100-SXM4-80GB per model: batch size 1, the same 50 pairs after 3 warm-ups, image loading included and model loading excluded, each model with its own runner and decoding settings, under a protocol fixed before the run; the 32B and 38B judges fit on one GPU. The fine-tuned judge's 0.23 s per pair was measured on an A100 in its own run.

Affect specialisation. Asked whether an emotion-specialised judge would do better, we tested three forms of specialisation. **Prompting.** We made InternVL3.5-14B affect-specialised through its prompt alone, adding one element at a time in an order fixed in advance: an expert-annotator persona trained on the Mikels taxonomy, a definition asking for the emotion a typical viewer would read, a two-factor instruction (did valence and arousal move toward the target; is it the same scene), and anchors reworded in affective terms. τ_b fell from $+0.329$ (standard rubric) to $+0.315$, $+0.315$, $+0.295$ and $+0.251$; the largest drop came from rewording the anchors, so how the scale is worded matters more here than whether the prompt mentions emotion theory. **EmoVIT.** We ran the released EmoVIT weights (Xie et al., 2024) with the same composites, rubric and parser. They reach $\tau_b{=}{+}0.091$, with 1,369 of 1,406 replies parsed and 5 distinct scores. EmoVIT is trained for single-image emotion recognition, not for judging a change between two images, and its $224{\times}224$ encoder sees each half of a composite at roughly 112 px, so we do not read this as evidence against affect tuning. **The EmoVIT recipe on a stronger backbone.** A public third-party reproduction of the EmoVIT training recipe on Qwen2.5-VL-7B (`wudq/EmoViT-Qwen2.5-Reproduce` on the Hugging Face Hub), run through the same code as its base model, reaches $\tau_b{=}{+}0.241$ against $+0.214$ for the base; the gain is not significant (Tab. 13), and it comes entirely from the first wave (on the held-out wave the tuned judge scores $+0.210$, below PAEF's $+0.229$). Our Qwen2.5-VL-7B row in Tab. 5 ($+0.200$) comes from a different runner with a different pixel budget; the matched-runner base is the correct control for the recipe.

Fine-tuning a small judge on our ratings. We fine-tuned Qwen2.5-VL-7B on AffectEdit's own human scores with LoRA (Hu et al., 2021) (rank 16, $\alpha{=}32$, dropout 0.05, adapters on the attention projections; AdamW, learning rate 10^{-4} held constant, 3 epochs), under 5-fold cross-validation stratified by target emotion (`StratifiedKFold`, 5 folds, seed 42), scoring each held-out fold with the model trained on the other four; folds are not grouped by source image, which if anything favours the tuned judge. The training-set size was swept over 0, 100, 200 and all $\sim$1,125 rated pairs per fold, a sweep fixed before training. The prompt is the standard rubric with the reply reduced to one digit, the training target is the rounded human mean, and the score is the expectation $\sum_i i\,p_i$ over the softmax of the five digit tokens' logits from a single forward pass, so scoring is deterministic. The five-fold mean τ_b rises from $+0.211$ with no training to $+0.270$, $+0.291$ and $+0.332$; PAEF's mean on the same folds is $+0.218$. Against PAEF on the same folds the tuned judge's advantage is $+0.052$, $+0.073$ and $+0.114$, with a 95% interval excluding zero only for the full training set. Pooled over the five held-out folds, the fully tuned judge's top fifth holds 2.5% failures and 39.0% successes, against PAEF's 2.1% and 31.6% (`experiments/iclr_04_extras.py`, section P). Training took 1.9 GPU-hours for the whole sweep and scoring 0.23 s per pair on one A100.

General-purpose learned evaluators. We also scored all 1,406 pairs with four learned evaluators, each run with its official code and pinned weights on V100 GPUs: the editing reward models EditReward (Wu et al., 2025b) (the mean of its two heads) and EditScore (Luo et al., 2026) (its official overall score, sampled at temperature 0.7), both given the instruction "Make this image feel more *emotion*"; the preference model HPSv3 (Ma et al., 2025), which scores the edited image against "a photo that feels *emotion*", with the difference between its scores for the edit and the original as a secondary readout; and VIEScore (Ku et al., 2024) with a Qwen2.5-VL-7B backbone. The protocols, the readouts and the rule for choosing a headline variant were fixed before any correlation was computed and ship with the release. VIEScore's official parser rejects 197 of its answers under the headline setting and 856 under the protocol as first written, reported as a second row; those pairs are left out rather than filled in, so its τ_b covers only the pairs it parsed, on which PAEF reaches $+0.201$ against its $+0.163$ (PAEF higher in 95% of resamples).

Table 6 lists the results. EditReward agrees with the raters better than PAEF overall and on both waves, HPSv3 is ahead overall, EditScore is level and VIEScore behind, and every evaluator's top fifth holds more human-rated failures than PAEF's. Table 7 shows where the agreement comes from,

in an analysis we chose after seeing these results. All four evaluators order the five editors exactly as the raters do, from Qwen-Image-Edit to AIF-Word, so their per-editor means alone reach τ_b=0.249, against 0.138 for PAEF, whose classifier term favours EmoEdit's edits. Within editors, EditReward stays ahead of PAEF (0.151 against 0.088), HPSv3 and EditScore are close (0.10), and all four are negative on AIF-Word edits, where PAEF reaches +0.248.

Table 6: **General-purpose learned evaluators** (Kendall τ_b against the human mean). Seconds per pair come from the same single-A100 timing run as Tab. 3; the repetition-penalty variant of VIEScore was not timed. P: share of resamples in which PAEF scores higher; for the VIEScore rows it compares scores on different pair sets (see text). Failures: share of human-rated failures in the top fifth, as in Tab. 16.

Evaluator	Backbone	s/pair	Parsed	Dist.	Wave 1	Wave 2	All	P	Fail. %
PAEF	CLIP, DINOv2	0.059	1406	1406	+0.213	+0.229	+0.219	—	2.1
EditReward	MiMo-VL-7B	0.37	1406	1406	+0.301	+0.285	+0.283	<0.01	4.3
HPSv3	Qwen2-VL-7B	0.31	1406	1406	+0.275	+0.255	+0.253	0.05	7.1
HPSv3, edit − original	Qwen2-VL-7B	0.31	1406	1406	+0.156	+0.189	+0.171	0.99	7.4
EditScore	Qwen2.5-VL-7B	4.7	1406	165	+0.231	+0.234	+0.227	0.35	8.4
VIEScore	Qwen2.5-VL-7B	5.8	1209	23	+0.172	+0.162	+0.163	0.99	3.8
VIEScore (pen. 1.05)	Qwen2.5-VL-7B	—	550	20	+0.044	+0.197	+0.139	0.99	6.9

Table 7: **Where the learned evaluators' agreement comes from** (exploratory). Editor means: τ_b of each metric's per-editor mean score alone. Within editor: τ_b after subtracting that mean. The last five columns give τ_b within each editor.

Metric	All	Editor means	Within editor	AIF-Word	EmoEdit	EmoEditor	FLUX.1	Qwen
PAEF	+0.219	+0.138	+0.088	+0.248	+0.300	+0.100	+0.082	+0.090
EditReward	+0.283	+0.249	+0.151	−0.034	+0.143	+0.196	+0.172	+0.185
HPSv3	+0.253	+0.249	+0.103	−0.064	+0.045	+0.228	+0.141	+0.159
EditScore	+0.227	+0.249	+0.102	−0.101	+0.060	+0.182	+0.095	+0.139
VIEScore	+0.163	+0.249	+0.062	−0.138	+0.072	+0.177	+0.091	+0.012

B PER-EMOTION AND PER-MODEL BREAKDOWN

PAEF is positive on every one of the 8 target emotions, 5 editors and 3 image sources (Tabs. 8 to 10; heatmap in Fig. 7), with a 95% interval above zero on 15 of the 16 slices. Table 11 summarises every metric and judge except SSIM, PSNR and LPIPS, which were added last without a slice analysis (overall they carry no signal or a negative one; Tab. 13). Emo-S, Emo-A, FindingEmo-DA, PAEF's two single-signal variants, the additive (ES, C) ablation, Qwen3-VL-32B and the matched-runner Qwen2.5-VL-7B control of Sec. A.5 are also positive on all 16 slices; ESMI is negative on 6, and the scene-only metrics on 5 to 14. Intervals come from the cluster bootstrap over source images of Sec. E. PAEF's worst slice (τ_b=+0.08, FLUX.1 Kontext) is the best worst slice of any metric, but it is within bootstrap noise of the others that stay positive, so we claim the absence of a negative slice, not an advantage in it. Eight of the nine judges are negative on AIF-Word edits, where PAEF and the emotion-only metrics stay positive (Tab. 8); we have no mechanism for this, and edit size does not explain it (Sec. J).

Editor ranking. Humans rank Qwen-Image-Edit first, then FLUX.1 Kontext, EmoEdit, EmoEditor and AIF-Word; $\mathrm{DA}{\cdot}C$ reproduces this order exactly, as do the four learned evaluators of Sec. A.5. PAEF ranks EmoEdit first, as do Emo-S, Emo-A and $\mathrm{ES}{\cdot}C$, which rest on the same classifier as ES, and FindingEmo-DA, ESMI and CLIP-I, which do not; gating the direction signal alone, $\mathrm{DA}{\cdot}C$, reproduces the human order, and seven of the nine judges rank FLUX.1 Kontext first (Tab. 12). EmoEdit's training pairs were selected with that classifier (Yang et al., 2025), so a metric built on it partly scores that editor on its own selection signal, and the fire shortcut of Sec. 4.6 concentrates in EmoEdit edits as well.



Figure 7: **Cross-cut Kendall τ_b on AffectEdit** (N=1,406). Per-cell τ_b vs. human mean across 16 slicings: 5 edit models (left), 3 image sources (middle), 8 Mikels emotions (right); n above each column. PAEF is positive on all 16 cuts ($[+0.08, +0.30]$), as are Emo-S, Emo-A and FindingEmo-DA (Tab. 11); InternVL3.5-14B spans a wider range ($[-0.09, +0.50]$), and each of the three judges shown is negative on AIF-Word only.

Table 8: **Per-edit-model Kendall's τ_b** on AffectEdit (point estimates on 100 to 446 pairs each) for PAEF, three MLLM judges and two metrics in current use. PAEF is positive on $5/5$ editors; InternVL3.5-14B is negative on AIF-Word. Red marks anti-correlation. IVL = InternVL3.5, Qwen-7B = Qwen2.5-VL-7B.

Editing model	n	PAEF	IVL-14B	IVL-8B	Qwen-7B	Emo-S	ESMI
AIF-Word	100	0.250	-0.09	-0.10	-0.06	0.220	0.040
EmoEdit	260	0.300	0.320	0.190	0.060	0.270	0.090
EmoEditor	155	0.100	0.120	0.130	0.100	0.040	-0.07
FLUX.1 Kontext	446	0.080	0.260	0.220	0.130	0.080	-0.03
Qwen-Image-Edit	445	0.090	0.190	0.130	0.120	0.070	-0.03

Per-emotion observations. The lowest PAEF per-emotion τ_b is on *disgust* (+0.10); we hypothesize that disgust edits often involve visceral content (food spoilage, decay) that humans rate viscerally but the V/A predictor cannot localize as a clean direction shift. *Anger* and *awe* (+0.28 each) are the easiest for PAEF. InternVL3.5-14B's advantage over PAEF is largest on *contentment* (+0.21) and *awe* (+0.22). Qwen2.5-VL-7B is weakest on *fear* and *sadness*.

C AffectEdit Datasheet

We follow the datasheet for datasets format (Gebru et al., 2021).

Motivation. AffectEdit was created to enable reliability-validated evaluation of affective image editing metrics, an emerging task with no prior reliability-validated benchmark.

Composition. N=1,406 edit pairs, each consisting of (original image, edited image, target Mikels emotion). Sources: 5 editing models (AIF-Word, EmoEdit, EmoEditor, Flux Kontext, Qwen-Image-Edit), 3 image sources (OASIS-80, EmoPair-504, EmoEdit-416), 8 target Mikels emotions (amusement, awe, contentment, excitement, anger, disgust, fear, sadness). Every edited image is machine-generated; the originals are photographs from the three sources. By a validated category probe (Sec. J), 38.9% of pairs show scenes, 28.3% people, 18.8% objects and 14.0% animals. The benchmark was assembled in two collection rounds. Each pair has $k \geq 4$ human ratings on a 0–4

Table 9: **Per-target-emotion Kendall's τ_b** on AffectEdit (IVL: InternVL3.5; FE-DA: FindingEmo-DA). Point estimates; Tab. 11 gives each metric's worst slice with its interval. PAEF is positive on $8/8$ emotions; the lowest cell is *disgust*, and the highest are *anger*, *awe*.

Emotion	n	PAEF	IVL-14B	IVL-8B	Qwen-7B	Emo-S	FE-DA
amusement	173	+0.18	+0.29	+0.28	+0.21	+0.13	+0.06
anger	244	+0.27	+0.33	+0.35	+0.21	+0.23	+0.30
awe	124	+0.28	+0.50	+0.23	+0.31	+0.20	+0.26
contentment	124	+0.19	+0.40	+0.37	+0.29	+0.12	+0.24
disgust	236	+0.10	+0.26	+0.20	+0.15	+0.05	+0.11
excitement	140	+0.24	+0.41	+0.31	+0.32	+0.20	+0.25
fear	210	+0.25	+0.31	+0.21	+0.07	+0.19	+0.19
sadness	155	+0.22	+0.27	+0.12	+0.09	+0.15	+0.17

Table 10: **Per-source Kendall's τ_b** on AffectEdit. OASIS-80 (n=96) is small, so its values are the least precise.

Source	n	PAEF	InternVL-14B	InternVL-8B	Qwen2.5-VL-7B	Emo-S
EmoEdit-416	594	$+.20$	$+.\mathbf{33}$	$+.30$	$+.23$	$+.16$
EmoPair-504	716	$+.22$	$+.\mathbf{33}$	$+.24$	$+.18$	$+.16$
OASIS-80	96	$+.19$	$+.\mathbf{23}$	$+.07$	$+.13$	$+.15$

BARS-style scale, k=5 for 1,271 of the 1,406 pairs (median 5, range 4–10); 7,144 total individual ratings.

Collection. Edits were generated using each editing model's standard inference pipeline with default hyperparameters. Human ratings were collected via Amazon Mechanical Turk in two rounds (Sec. 4). The first export of the second round's results mis-mapped part of the scores; 593 ratings on 261 pairs were restored to the scores the workers had submitted, and every number in this paper uses the restored ratings. The release marks each restored rating. Workers were qualified ($\geq$ 99% approval, $\geq$ 5,000 prior HITs, US/CA/UK/AU/NZ) and passed a comprehension quiz with visual examples (6% fail rate). Workers were paid \$0.65 per HIT. At least 196 unique workers rated the second round's pairs; the first round's worker identities were not retained, only each rating's position within its pair.

Preprocessing. A pair's human score is the mean of all its ratings. For the ICC only, pairs with k>5 are trimmed to five ratings by dropping the rating furthest from the median, and the 30 pairs with k=4 are padded with their mode. No personally identifiable information is collected from workers (only worker IDs from MTurk's anonymization scheme).

The AIF-Word template. AIF-Word is our shorthand, not a released variant: Affective Image Filter (Weng et al., 2023) takes free-text emotional descriptions rather than discrete emotion labels, so we drive it with the single-word template *"a photo that feels [emotion]"* to give it a Mikels target. Its edits are not visually subtle: their mean pixel change is above that of EmoEditor edits and close to that of EmoEdit edits, so edit size does not explain why the judges fail on them.

Editor coverage and sampling rationale. Per-editor pair counts in AffectEdit are unbalanced by design: AIF-Word (n=100), EmoEditor (n=155), EmoEdit (n=260), Qwen-Image-Edit (n=445), and Flux-Kontext (n=446). We deliberately favoured zero-shot editors (FLUX.1 Kontext, Qwen-Image-Edit) for full coverage because they are not trained against any of PAEF's components, so they yield the cleanest evaluation of the metric on out-of-distribution edits. We reduced coverage of EmoEdit, whose training pairs were selected with the same classifier that supplies ES, and of EmoEditor, whose training pairs are labelled and whose outputs are selected at inference by its own emotion predictor trained on EmoSet, the data behind that classifier; both are close to one of our components and could inflate the metric. This is the in-distribution effect we report in Tab. 12. We further reduced coverage of AIF-Word: because the AIF-Word editor does not accept explicit

Table 11: **Every metric on the** 16 **fixed slices** (5 editors, 3 image sources, 8 target emotions). Negative: slices with $\tau_b<0$. CI>0: slices whose 95% interval lies above zero. Worst: the lowest slice τ_b with the interval of the minimum over slices. P: share of resamples in which PAEF's worst slice is higher than the metric's. PAEF's worst slice is the highest point estimate, but it is within noise of the other metrics that stay positive, so we claim no advantage from it.

Metric	Negative	CI>0	Worst [95% CI]	Best	P
PAEF	0	15	+0.08 [−0.02, +0.10]	+0.30	—
DA·C	0	15	+0.06 [−0.01, +0.10]	+0.31	0.52
ES·C	0	14	+0.05 [−0.06, +0.08]	+0.28	0.91
FindingEmo-DA	0	14	+0.06 [−0.06, +0.08]	+0.30	0.74
Emo-S	0	12	+0.04 [−0.08, +0.07]	+0.27	0.94
Emo-A	0	8	+0.03 [−0.13, +0.07]	+0.29	0.90
OASIS-DA	1	9	−0.01 [−0.15, +0.03]	+0.31	0.98
CLIPScore	2	7	−0.03 [−0.17, +0.01]	+0.24	0.99
S_{emo}	2	8	−0.06 [−0.18, −0.01]	+0.25	>0.99
ESMI	6	2	−0.15 [−0.28, −0.07]	+0.12	>0.99
DINOv2 cosine	5	0	−0.15 [−0.25, −0.09]	+0.09	>0.99
S_{str}	14	0	−0.19 [−0.33, −0.13]	+0.03	>0.99
CLIP-I	14	0	−0.22 [−0.31, −0.15]	+0.04	>0.99
InternVL3.5-1B	3	5	−0.21 [−0.35, −0.08]	+0.18	>0.99
InternVL3.5-2B	3	1	−0.06 [−0.21, −0.03]	+0.13	>0.99
InternVL3.5-4B	1	14	−0.01 [−0.18, +0.08]	+0.42	0.90
InternVL3.5-8B	1	14	−0.10 [−0.25, +0.03]	+0.37	0.97
InternVL3.5-14B	1	14	−0.09 [−0.24, +0.06]	+0.50	0.95
InternVL3.5-38B	1	14	−0.15 [−0.30, −0.00]	+0.42	0.99
Qwen2.5-VL-7B	1	10	−0.06 [−0.23, +0.05]	+0.32	0.95
Qwen2.5-VL-32B	1	15	−0.08 [−0.23, +0.07]	+0.45	0.94
Qwen3-VL-32B	0	15	+0.05 [−0.09, +0.14]	+0.44	0.53
EmoVIT	3	6	−0.10 [−0.28, −0.03]	+0.18	>0.99
EmoVIT-recipe Qwen2.5-VL-7B	1	11	−0.07 [−0.23, +0.05]	+0.35	0.95
Qwen2.5-VL-7B (matched runner)	0	12	+0.01 [−0.16, +0.05]	+0.40	0.94

emotion labels in its instruction interface, we queried it with the templated prompt *"a photo that feels [emotion]"* for each Mikels target. In pilot inspection its outputs under this prompt were frequently unresponsive to the emotion target, changing the image without moving it toward the requested emotion; the 100 pairs we kept span the cases where it did produce a discernible edit. Per-emotion stratification was preserved within each editor budget.

Uses. AffectEdit is intended for evaluating affective image editing metrics. The source images come from public research datasets, and some show people, including children. Edits toward negative emotions can transform them in unflattering or unsettling ways, so the data must not be used to identify or characterise the people shown; the figures that show such edits (Figs. 1 and 5) do so only to illustrate how metrics score them.

Distribution and licensing. Our ratings, per-pair scores and annotations are released under CC BY 4.0; the images keep the terms of their sources and editors, which differ, so there is no blanket licence. OASIS images (Kurdi et al., 2017) were selected by their authors as free to edit and redistribute. The EmoEdit-416 originals are Unsplash and Pexels photographs, whose licences permit modification and redistribution. The EmoPair-504 originals are EmoSet photographs (Yang et al., 2023) obtained through EmoPair (**?**), which declares no licence; we do not redistribute them and release a script that restores them from their source by checksum. Edits keep the terms of the editor that produced them: FLUX.1 Kontext outputs fall under its non-commercial licence, which requires disclosing that they are AI-generated and forbids using them to train a competing model; Qwen-Image-Edit is Apache 2.0; EmoEdit and EmoEditor declare no licence. Every edited image is AI-generated, and the release says so.

Table 12: **Editor ranking.** Spearman ρ between each metric's per-editor mean score and the human per-editor mean over the 5 editors, and the editor each ranks first. With five editors, ρ moves in steps of 0.1 and a single swap separates most rows, so we read this table for the top-ranked editor, not for small differences in ρ. Metrics built on the EmoEdit classifier rank EmoEdit first; DA·C reproduces the human order.

Metric	ρ	Ranked first
Humans	—	Qwen-Image-Edit
PAEF	+0.60	EmoEdit
ES·C	+0.60	EmoEdit
DA·C	+1.00	Qwen-Image-Edit
Emo-S	+0.60	EmoEdit
Emo-A	+0.60	EmoEdit
FindingEmo-DA	+0.60	EmoEdit
ESMI	+0.10	EmoEdit
CLIP-I	−0.10	EmoEdit
InternVL3.5-1B	+0.60	FLUX.1 Kontext
InternVL3.5-2B	+1.00	Qwen-Image-Edit
InternVL3.5-4B	+0.90	FLUX.1 Kontext
InternVL3.5-8B	+0.90	FLUX.1 Kontext
InternVL3.5-14B	+0.80	FLUX.1 Kontext
InternVL3.5-38B	+0.90	FLUX.1 Kontext
Qwen2.5-VL-7B	+0.90	FLUX.1 Kontext
Qwen2.5-VL-32B	+0.90	FLUX.1 Kontext
Qwen3-VL-32B	+0.90	Qwen-Image-Edit

Maintenance. We will host AffectEdit at a public address given in the final version and maintain it for at least three years post-publication.

D Human study details

D.1 HIT interface

Each HIT shows the original and edited images side by side, the target emotion, the question (*"How well did this edit make the image feel more [emotion] while keeping the same scene?"*), and a 0–4 BARS-style rating scale with anchors:

- 0: Failed (no change or different photo)
- 1: Barely changed
- 2: Emotion shifted but messy (senseless changes)
- 3: Good edit with purposeful changes
- 4: Excellent / seamless (atmospheric only)

A screenshot of the rendered HIT is shown in Fig. 8; the full HIT HTML and consent text are included with the released code.

D.2 Worker qualification and quality filters

Workers were required to have $\geq 99\%$ HIT approval rate, $\geq 5{,}000$ prior approved HITs, and to be located in the US, Canada, UK, Australia, or New Zealand. A pre-task comprehension quiz with visual examples filtered out workers who could not distinguish identity edits, content swaps, and successful emotional shifts; 6% of attempted workers failed and were excluded. Within-task quality filters excluded workers who failed > 1 attention check, had a median response time below 1.5s or above 120s, or showed $> 80\%$ straight-lining on any rating dimension.

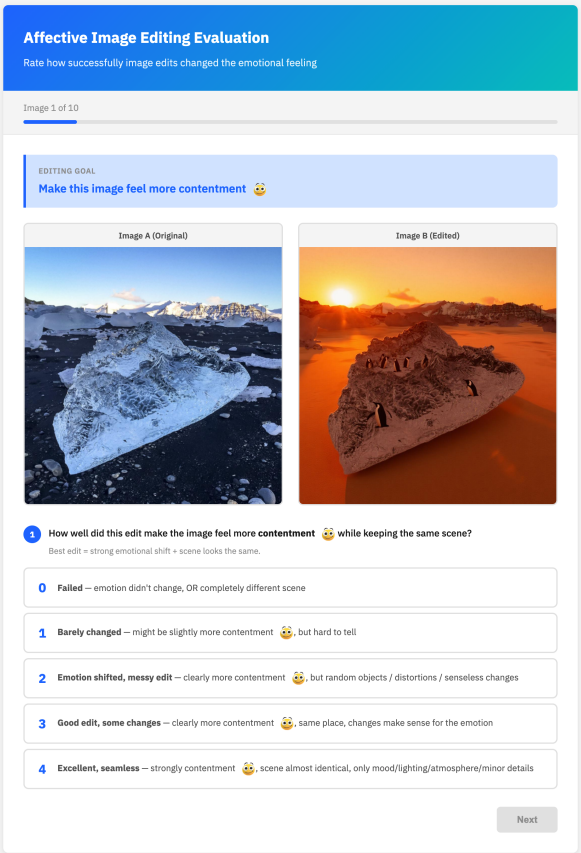

Figure 8: **AffectEdit Mechanical Turk HIT interface.** Each HIT shows the original and edited images side by side, the target Mikels emotion, the question *"How well did this edit make the image feel more [emotion] while keeping the same scene?"*, and a 0–4 BARS-style rating scale with the anchors listed in Sec. D. Workers were qualified ($\geq$ 99% approval, $\geq$ 5,000 prior HITs, US/CA/UK/AU/NZ) and passed a comprehension quiz with visual examples before rating.

D.3 Compensation and ethics

Workers received \$0.65 per HIT, equivalent to $\sim$\$13/hour at the median observed completion time of $\sim$3 minutes per HIT, above the U.S. federal minimum wage. Raters judged ordinary photographs and their edits toward eight emotions, four of them negative; no personal information was collected beyond the platform-assigned worker identifiers used to pay and de-duplicate raters. It was conducted under a standing institutional ethics approval covering crowdsourced collection of human responses to visual stimuli; the issuing committee and protocol number are withheld here for anonymity and will be named in the camera-ready.

D.4 Why one joint rating

Raters answered one question, with both images side by side: how well did the edit make the image feel more like the target *while keeping the same scene*? The 0–4 anchors encode both requirements. We did not collect separate emotion and preservation ratings. The claim under test is that the two requirements must not be scored apart and combined afterwards, which is how additive metrics fail; separate human ratings would have rebuilt that decomposition on the human side and left no joint judgment to validate a metric against. The ratings nevertheless respond to both requirements. Holding the gate C fixed, the human score rises with PAEF's emotion signal E (rank partial correlation 0.30, 95% CI $[0.25, 0.35]$), and holding E fixed it still rises with C (0.07, $[0.02, 0.13]$), so raters applied the preservation criterion as well as the emotional one.

D.5 INTER-RATER RELIABILITY

We compute ICC$(2,1)$ (single-rater absolute agreement) and ICC$(2,k)$ (average-of-k absolute agreement) following Shrout & Fleiss (1979), on the preprocessing of Sec. C. On all $1{,}406$ pairs, ICC$(2,1)$=0.483 ("fair" by Cicchetti's criteria (Cicchetti, 1994)) and ICC$(2,5)$=0.824 ("excellent"). By collection wave, ICC$(2,5)$ is 0.880 on the first $(462$ pairs) and 0.785 on the second (944); by content category it ranges from 0.809 (scenes) to 0.837 (objects); all are in the "excellent" band. The gap between single-rater and five-rater agreement reflects the subjectivity of affective judgment: individual raters disagree on edge cases, but the mean of five is reliable, so we never treat a single rater as a substitute for the aggregate. Even with five raters per pair, this residual subjectivity bounds the correlation any metric can reach with the human aggregate. Krippendorff's α (interval), which needs neither equal numbers of ratings nor the same raters across pairs, is 0.459 on all ratings of all pairs (0.595 and 0.387 by wave), in line with the single-rater ICC. All values are computed by `experiments/iclr_03_reliability.py` from the released ratings.

E SIGNIFICANCE TESTING PROTOCOL

Single-metric significance (τ_b=0 null). Two-sided p-values under the null hypothesis τ_b=0 are computed from the asymptotic Kendall–Mann distribution as implemented by `scipy.stats.kendalltau(variant='b', method='auto')`. For N=1,406 this approximation is exact to within numerical tolerance of an exact permutation null on the same data; we spot-checked agreement with a B=10,000 paired-sign-flip permutation null on a random 50-pair subset and p-values agreed to within 10^{-4}. We do not use significance stars. Table 1 greys every value whose 95% interval includes zero and reports the paired-bootstrap probability against PAEF directly; Fig. 3(a) shows the intervals.

Paired-metric significance (τ_A vs. τ_B). For paired comparisons ("does metric A correlate with humans better than metric B?"), we use a paired bootstrap. Each iteration b resamples the source images *with replacement*, takes every pair of each drawn image, and computes *both* metrics' τ_b on the same resampled pairs, recording $\Delta\tau^{(b)} = \tau_A^{(b)} - \tau_B^{(b)}$. This preserves the dependence between the two metrics on the same pairs and yields valid CIs for $\Delta\tau$. The reported $P(\text{A}{>}\text{B})$ is the fraction of bootstrap resamples in which $\tau_A^{(b)} > \tau_B^{(b)}$.

Bootstrap budget. Because one source image appears in several pairs, we resample source images rather than pairs (a *cluster* bootstrap), which is more conservative; intervals are the 2.5–97.5 percentiles. Every interval and paired probability computed by `experiments/iclr_01_tables.py`, `iclr_04_extras.py`, `iclr_05_pixel_metrics.py` or `iclr_06_config_sequence.py`, which covers Tabs. 1, 11, 13, 15 and 16 and Fig. 3, comes from one global draw (731 images, B=2,000, seed 20260725), so these comparisons are all paired. The operator sweep (Tab. 14), the probes of Sec. I (per row, seed 42) and the analyses of Sec. J use cluster draws of their own; the fine-tuned judges are compared with PAEF fold by fold (Sec. A.5). Where a family of tests is reported, we state its size and correct within it: Holm's procedure (Holm, 1979) for the operator sweep (nine comparisons per scaling) and Bonferroni for the cue audit (11 cells).

F QUALITATIVE FAILURE CASES

Figure 9 extends the bottom rows of Fig. 5 with the pairs from each tail of the agreement distribution, decomposing each failure into its DA, ES and coherence components. **False positives** ($\bar{h}{\leq}1$ but PAEF in its top fifth) have large ES ($+0.53$ to $+0.89$) and high C (0.73–0.92). Their DA ($+0.03$ to $+0.13$) does not set them apart: over all six false positives in the top fifth its median is 0.10, against 0.11 for the top fifth's 89 true successes. What lifts them into the top fifth is ES. Three of the four shown add festive objects (bunting, balloons, hanging decorations). **False negatives** ($\bar{h}{\geq}3$ but PAEF in its bottom 3%) have *negative* ES (-0.10 to -0.32): the classifier already gave these sources 0.22 to 0.72 probability of the target, against a median of 0.002 over all pairs, so it reports a drop. The pattern is general: 22 of the 30 human-rated successes in PAEF's bottom fifth have negative ES, and

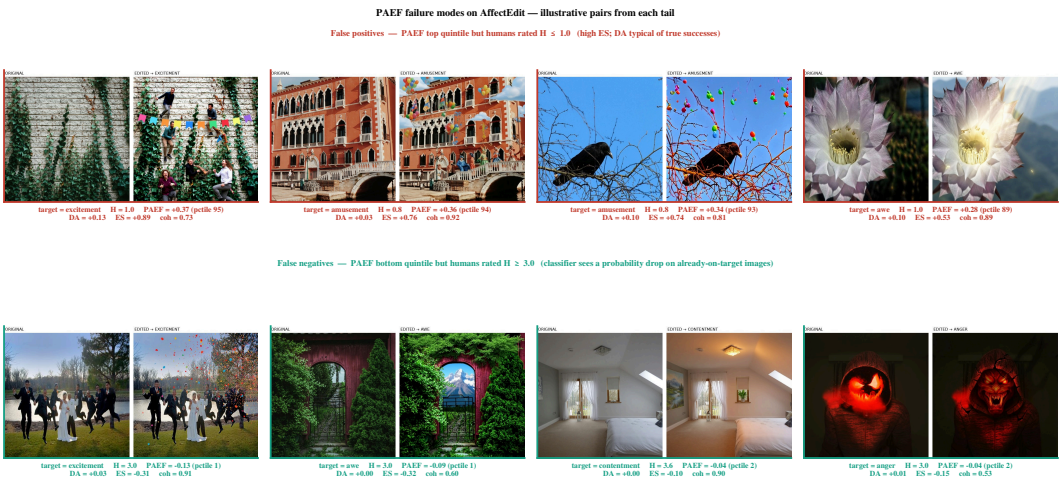

Figure 9: **PAEF failure modes on AffectEdit.** Illustrative pairs from each tail of the agreement distribution. **False positives (top row):** edits in PAEF's top quintile (pctile ≥ 89) that humans rate as failures ($\bar{h} \leq 1.0$). Pattern: ES is high ($+0.53$ to $+0.89$) and coherence is high (0.73–0.92); DA ($+0.03$ to $+0.13$) is in the range of the top fifth's true successes (median $+0.11$), so neither signal flags the failure. Three of the four edits add festive objects (bunting, balloons, hanging decorations). **False negatives (bottom row):** edits in PAEF's bottom 3% that humans rate as successes ($\bar{h} \geq 3.0$). Pattern: ES is *negative* (-0.10 to -0.32) because the classifier already gave the source 0.22 to 0.72 probability of the target (median over all pairs 0.002; e.g., a bedroom it read as contentment at 0.72), so it registers a *drop* rather than the expected rise. Both failure modes trace to the classifier signal ES rather than to PAEF's composition.

their sources' median target probability is 0.43. Both modes trace to the classifier signal ES rather than to PAEF's composition.

G Broader Impact, Risks, and Discussion

PAEF and **AffectEdit** are evaluation infrastructure for the emerging field of affective image editing. This appendix discusses the intended scope of the contribution, the beneficial uses we expect, the harms we foresee, the construct boundaries that constrain interpretation, the inherited and structural biases the system propagates, the worker-welfare considerations of the human study, the safeguards in place at release, and the scientific questions our findings raise but do not resolve.

Scope of the contribution. We release a benchmark and an evaluation metric. We do not release a generative model, a fine-tuned editor, or any new image-generation weights. Any downstream system that uses PAEF as a training signal must combine it with a separate generative model and is the responsibility of that system's deployer. PAEF is one input to a hypothetical pipeline; it is not the pipeline.

Intended beneficial uses. A reliability-validated affective-edit metric supports several legitimate applications. **(i) Better generative editors.** Researchers can use PAEF as a reproducible, fast, decomposable signal for model selection and for screening edits before human rating, a deterministic alternative to costly VLM-as-judge protocols. We have not tested it as a training reward (e.g., for DPO (Rafailov et al., 2023)); an optimiser could exploit the shortcuts documented in Sec. J, so such use needs its own validation. **(ii) Accessibility and creative tools.** Tools that adapt visual content to user-specified moods (mental-health-aware media, accessible storytelling, museum education) need a reproducible quality signal that does not require an LLM in the loop; PAEF provides one. **(iii) Affective-computing research.** AffectEdit's $1{,}406$ k=5-rated edit pairs enable reliability-validated downstream studies (cultural-variation analysis, intensity calibration, perceptual modeling) that prior datasets, none of which releases absolute ratings of edits with reported inter-rater agreement, cannot support. **(iv) Detection and moderation evaluation.** A directional affective metric can be

used as part of a screening pipeline to evaluate *whether* an automated edit is shifting affect at all and *toward what target*, useful for moderation systems that need to flag emotionally manipulated content before it propagates.

Foreseeable harms and dual-use risks. The same capability enables harm. **Misuse for emotional manipulation at scale.** A reliable affective-edit reward signal lowers the cost of training generative systems that produce emotionally manipulative imagery: political propaganda optimized to elicit fear or anger, deceptive advertising optimized to elicit attachment or aspiration, or fabricated images of real people calibrated to produce specific affective responses. PAEF's parameter-free, deterministic, decomposable nature is itself the dual-use property: it makes the metric easy to deploy as a training reward and is reproducible across actors and budgets, including adversarial ones. **Erosion of media trust.** At population scale, the proliferation of edited imagery whose perceived affective content is systematically optimized away from the source can degrade the assumption that images carry information about the world, with downstream consequences for journalism, evidentiary practice, and public discourse. PAEF does not cause this risk, but lowering the cost of affective-edit optimization is on the causal pathway to it. **Targeted distress.** A metric that reliably predicts which edits produce specific negative affect (anger, fear, disgust) makes it easier to optimize content for harm to specific viewers. The benchmark we release contains edits toward fear, disgust, anger and sadness, some of them unsettling, and the metric can be evaluated on any new content. **Reinforcement of categorical emotion assumptions.** By operationalizing the Mikels-8 taxonomy, PAEF implicitly endorses a discrete-categorical view of emotion that is contested in affective science (basic-emotion theories vs. constructionist accounts). Wide deployment of metrics built on this assumption can entrench it as a deployment standard even where it is psychologically inappropriate (Barrett, 2017).

Construct boundaries: what PAEF does *not* measure. Construct mis-application is the most likely source of inadvertent harm in a metric like ours. **Perceived $\neq$ felt.** PAEF measures the perceived affective content of an edited image (what a typical viewer reads from the image), aggregated over five qualified MTurk raters per pair (four to ten on 135 pairs). It does not measure the *felt* affect of any individual viewer, which depends on personal history, current state, and context. A high PAEF score indicates that the image will likely be *read* as expressing the target emotion, not that any viewer will *experience* it. **Perceived $\neq$ ground truth.** Edited images do not have a ground-truth emotion: the depicted scene is synthetic, and the expressed emotion is whatever viewers attribute to it. Treating PAEF as evidence about the mental state of any depicted human subject is a misuse. **Aggregate $\neq$ individual.** $\mathrm{ICC}(2,1)$=0.483 on AffectEdit indicates substantial disagreement among individual raters; the headline $\mathrm{ICC}(2,5)$=0.824 reliability is for the population mean. PAEF should be interpreted as a population-level estimator, not a per-individual prediction.

Inherited and structural biases. Several biases enter PAEF from upstream choices we did not make and could not easily correct. **Cultural and linguistic narrowness.** The Mikels-8 taxonomy was developed with American undergraduate participants (Mikels et al., 2005); the NRC-VAD lexicon (Mohammad, 2025) is English-language; CLIP and DINOv2 are trained on web-scraped corpora with documented over-representation of Western imagery. Our crowd workers were drawn from US/CA/UK/AU/NZ ($\geq$ 99% approval, $\geq$ 5,000 prior HITs); we did not stratify by age, gender, race, or first language. PAEF scores should not be assumed to generalize across cultural or linguistic contexts without revalidation. **Backbone representational biases.** CLIP and DINOv2 encode known biases related to gender, skin tone, age, and geographic representation (Oquab et al., 2023; Radford et al., 2021); the EmoEdit classifier inherits these and adds emotion-specific category boundaries from its training data. Because PAEF is a function of these backbones, every bias propagates through the metric. **Editor-distribution bias.** PAEF's component classifier is the same as one used by an editor in our benchmark (EmoEdit), producing the in-distribution effect we report in Tab. 12 (PAEF ranks EmoEdit first; humans rank it third). This is a generic property of classifier-based metrics evaluated on edits trained against that classifier; it is not unique to PAEF, but it should be acknowledged whenever PAEF is used to compare an editor that was trained against any of PAEF's components against one that was not.

Worker welfare in the human study. The AffectEdit human study used Amazon Mechanical Turk, which has well-documented labor-equity concerns (Hara et al., 2018). We mitigated as follows: workers were paid \$0.65 per HIT, equivalent to $\sim$\$13/hour at the median observed completion time

of $\sim$3 minutes per HIT (above the U.S. federal minimum wage of \$7.25); only qualified workers were eligible ($\geq$ 99% approval rate, $\geq$ 5,000 prior HITs); the rating task showed ordinary photographs and their edits toward eight emotions, four of them negative; attention checks and a comprehension quiz filtered low-quality work without retroactive non-payment of completed work; and worker IDs are anonymized in the released benchmark. We acknowledge that even fair pay on MTurk is not equivalent to formal employment with benefits and that the platform's structural asymmetries are not addressable at the level of any individual study.

Safeguards at release. **(i) Evaluation-only release.** We release PAEF as an evaluation metric and AffectEdit as a benchmark; the release contains no novel generative weights. Any harmful application requires combining PAEF with a separate generator that is not part of this release. **(ii) Licensing.** Our ratings and scores are released under CC BY 4.0; images keep their sources' and editors' terms, and originals whose redistribution is unclear are not redistributed (Sec. C). **(iii) Anonymization.** Worker IDs are anonymized; no personally identifiable information about workers is collected or released. **(iv) Datasheet.** The release includes a datasheet (Sec. C) documenting collection, intended uses, and known limitations. **(v) Documentation of failure modes.** The qualitative figures (Figs. 5 and 9) and the limitations section name PAEF's failure modes, so practitioners can identify known-bad inputs before deployment.

Open scientific and ethical questions. Our findings raise four questions we do not resolve in this paper: (i) why MLLM judges and classifier-based metrics err in opposite directions when ranking editors, (ii) cross-taxonomy and cross-cultural transfer of the PAEF framework, (iii) held-out generalization protocols when components are themselves trained on related distributions, and (iv) operationalizing the perceived/felt distinction with separate instruments calibrated against PAEF. Section 4.6 lists the limits of the present evidence that bear on them.

H COLLECTION WAVES, OPERATORS AND GATE STRENGTH

Every number in this section comes from the stored per-pair outputs through `experiments/iclr_01_tables.py`, which records its analysis protocol, fixed before its first run, in its header; uncertainty is the cluster bootstrap of Sec. E.

By collection wave. Table 13 extends Tab. 1 to every metric and judge. PAEF is the highest-scoring parameter-free metric on each wave, and most metrics score similarly on both. The exceptions are ESMI and its emotion term S_{emo}, near zero on the first wave and significant on the second, which is why we state ESMI's null result for the benchmark as a whole, and OASIS-DA, which gains about 0.06 on the second. On PAEF's components, the product ES$\cdot C$ beats their average $\frac{1}{2}(\text{ES}+C)$ by $+0.057$ on the first wave (paired P=0.99), $+0.021$ on the second (P=0.90) and $+0.031$ on all pairs (P=0.99); on ESMI's terms the product's gain is -0.013, $+0.085$ and $+0.046$ (P=0.16, 1.00 and 1.00).

Operator sweep. With the emotion signal $E=\frac{1}{2}(\text{DA}+\text{ES})$ and the gate C fixed, we compared ten parameter-free ways of combining them under five ways of putting the two signals on a common range (Tab. 14): the native scale (S0), per-axis min–max to $[0, 1]$ (S1), per-axis rank-uniform transformation (S2), min–max after winsorising at the 1st and 99th percentiles (S3), and clamping to $[0, 1]$ without a shift (S4, added after the other four). The Łukasiewicz, Einstein and Hamacher t-norms need inputs in $[0, 1]$ and are undefined on the native scale. Each operator is compared with the product on identical resamples, with Holm's correction (Holm, 1979) over the nine comparisons within each scaling. The sweep supports one conclusion under every scaling: addition is among the three lowest-ranked operators. It does not single out the product. The product ties the geometric mean on the native scale, which ranks every pair with $E\geq 0$ identically; after min–max rescaling (S1, S3) the minimum, harmonic mean and Hamacher product score significantly higher than the product, and after rank-uniform rescaling (S2) no operator differs significantly from it. The rescalings that favour them are exactly those that remove the property we choose the product for: on the native scale $E=0$ means no emotional change and C near 0 means a different scene, so a product of the two is zero whenever either requirement fails, without assuming the two signals share a scale.

Table 13: **Every metric and judge by collection wave** (Kendall τ_b against the human mean). P is the share of resamples in which PAEF scores higher on all pairs. Grey: 95% interval includes zero. The matched-runner Qwen2.5-VL-7B is the control for the EmoVIT-recipe judge (Sec. A.5).

Metric	Wave 1 (462)	Wave 2 (944)	All (1,406)	P
PAEF	+0.213	+0.229	+0.219	—
DA·C	+0.193	+0.213	+0.204	0.82
ES·C	+0.198	+0.185	+0.184	>0.99
$S_{\text{emo}}\cdot S_{\text{str}}$	−0.051	+0.166	+0.066	>0.99
$\frac{1}{2}$(ES+C)	+0.142	+0.164	+0.154	>0.99
FindingEmo-DA	+0.166	+0.194	+0.183	0.99
Emo-S	+0.179	+0.175	+0.172	>0.99
Emo-A	+0.177	+0.163	+0.167	0.99
OASIS-DA	+0.115	+0.172	+0.150	>0.99
CLIPScore	+0.114	+0.097	+0.101	>0.99
S_{emo}	−0.029	+0.189	+0.079	>0.99
ESMI	−0.038	+0.081	+0.020	>0.99
DINOv2 cosine	+0.008	+0.020	+0.017	>0.99
S_{str}	−0.072	−0.030	−0.045	>0.99
CLIP-I	−0.061	−0.063	−0.063	>0.99
SSIM	−0.104	−0.040	−0.064	>0.99
PSNR	−0.161	−0.182	−0.173	>0.99
−LPIPS	−0.053	+0.012	−0.015	>0.99
InternVL3.5-1B	+0.141	+0.064	+0.089	>0.99
InternVL3.5-2B	+0.090	+0.049	+0.058	>0.99
InternVL3.5-4B	+0.319	+0.299	+0.297	<0.01
InternVL3.5-8B	+0.290	+0.253	+0.259	0.03
InternVL3.5-14B	+0.364	+0.325	+0.329	<0.01
InternVL3.5-38B	+0.344	+0.317	+0.318	<0.01
Qwen2.5-VL-7B	+0.227	+0.195	+0.200	0.76
Qwen2.5-VL-32B	+0.333	+0.295	+0.299	<0.01
Qwen3-VL-32B	+0.371	+0.338	+0.339	<0.01
EmoVIT	+0.097	+0.081	+0.091	>0.99
EmoVIT-recipe Qwen2.5-VL-7B	+0.321	+0.210	+0.241	0.17
Qwen2.5-VL-7B (matched runner)	+0.246	+0.205	+0.214	0.57

Table 14: **Operator sweep** (Kendall τ_b against the human mean, all 1,406 pairs). Bold: differs from the product at Holm-adjusted $p<0.05$ within that scaling. S4 was added after S0–S3.

Operator on (E, C)	S0 native	S1 min–max	S2 rank	S3 robust	S4 clamp
product $E \cdot C$ (PAEF)	+0.219	+0.166	+0.175	+0.197	+0.219
geometric mean	+0.219	+0.166	+0.175	+0.197	+0.219
minimum	+0.211	**+0.210**	+0.175	**+0.220**	+0.212
harmonic mean	**+0.207**	**+0.184**	+0.174	**+0.216**	+0.214
Hamacher product	—	**+0.184**	+0.174	**+0.216**	+0.214
Einstein product	—	**+0.155**	+0.173	**+0.185**	+0.220
Łukasiewicz t-norm	—	**+0.142**	**+0.145**	**+0.155**	**+0.163**
additive $\frac{1}{2}(E+C)$	**+0.148**	**+0.142**	+0.161	**+0.157**	**+0.144**
E alone (no gate)	**+0.203**	+0.203	+0.203	+0.203	+0.204
C alone	**+0.017**	**+0.017**	**+0.017**	**+0.017**	**+0.017**

Gate strength. To test whether the gate over-penalises the scene change an affective edit needs, we weakened and strengthened it as $E \cdot \text{sign}(C)\,|C|^{\gamma}$ on a grid fixed in advance (Tab. 15; the sign term keeps the 2 pairs with $C<0$ defined and leaves $\gamma=1$ identical to PAEF). τ_b rises steadily up to $\gamma=1$ and flattens beyond it; no stronger gate beats $\gamma=1$ significantly, and the ungated signal E alone scores $+0.203$ (Tab. 14). A metric that over-penalised necessary scene change would have peaked below $\gamma=1$.

Table 15: **Gate strength.** τ_b of $E \cdot \mathrm{sign}(C)|C|^\gamma$ with its 95% interval, and the share of resamples in which it beats γ=1.

γ	τ_b	95% CI	$P(>\gamma{=}1)$
0.00	+0.207	[+0.170, +0.240]	0.01
0.25	+0.211	[+0.175, +0.244]	0.03
0.50	+0.215	[+0.180, +0.248]	0.07
0.75	+0.217	[+0.182, +0.250]	0.13
1.00 (PAEF)	+0.219	[+0.184, +0.252]	—
1.25	+0.220	[+0.185, +0.254]	0.90
1.50	+0.221	[+0.186, +0.255]	0.87
1.75	+0.221	[+0.186, +0.256]	0.79
2.00	+0.221	[+0.186, +0.256]	0.72

Top fifth. Table 16 lists, for every metric and judge, the share of human-rated successes and failures among its 282 top-scored pairs, one fifth of the benchmark, with ties at the cut broken by exact expectation. Adding ES to DA leaves τ_b within noise but cuts the failures in the top fifth from 5.3% to 2.1%; PAEF's top fifth contains fewer failures than that of every baseline and zero-shot judge in at least 95% of resamples (than the learned evaluators' in at least 92%), and fewer successes than the best judges'.

Table 16: **Top fifth of each metric** (282 of 1,406 pairs). Successes: $\bar{h}\geq 3$ (base rate 20.3%). Failures: $\bar{h}\leq 1$. P: share of resamples in which PAEF's failure rate is lower.

Metric	Successes % [95% CI]	Failures % [95% CI]	P
PAEF	31.6 [26.4, 37.4]	2.1 [0.4, 3.9]	—
DA·C	29.4 [23.9, 35.0]	5.3 [2.8, 8.1]	>0.99
ES·C	31.9 [26.3, 37.1]	2.1 [0.4, 4.0]	0.14
Emo-S	29.8 [24.5, 35.5]	5.0 [2.5, 7.7]	>0.99
Emo-A	29.6 [25.4, 33.8]	8.5 [6.0, 11.0]	>0.99
FindingEmo-DA	29.4 [23.7, 34.5]	7.4 [4.9, 11.2]	>0.99
OASIS-DA	28.0 [22.6, 33.0]	8.9 [5.9, 12.6]	>0.99
CLIPScore	27.7 [22.5, 33.9]	8.2 [4.9, 11.3]	>0.99
S_{emo}	24.1 [18.8, 28.9]	15.6 [12.0, 20.3]	>0.99
ESMI	24.1 [19.1, 29.1]	16.7 [12.2, 20.9]	>0.99
S_{str}	17.4 [13.2, 22.3]	18.4 [13.8, 23.2]	>0.99
DINOv2 cosine	16.7 [12.1, 21.3]	21.6 [17.1, 26.2]	>0.99
CLIP-I	16.0 [12.2, 20.4]	25.5 [20.4, 30.2]	>0.99
SSIM	19.5 [15.0, 24.3]	16.3 [12.4, 20.6]	>0.99
PSNR	13.8 [9.8, 18.1]	26.2 [20.9, 30.5]	>0.99
−LPIPS	15.6 [11.6, 20.6]	18.8 [14.5, 23.4]	>0.99
InternVL3.5-1B	23.5 [19.9, 27.1]	14.7 [11.7, 17.7]	>0.99
InternVL3.5-2B	21.4 [18.7, 24.3]	15.4 [13.4, 17.4]	>0.99
InternVL3.5-4B	28.9 [25.4, 32.5]	8.1 [6.2, 10.1]	>0.99
InternVL3.5-8B	29.0 [25.5, 32.6]	8.0 [6.0, 10.2]	>0.99
InternVL3.5-14B	35.9 [31.6, 40.3]	5.2 [3.1, 7.3]	>0.99
InternVL3.5-38B	32.2 [28.4, 36.2]	6.5 [4.5, 8.5]	>0.99
Qwen2.5-VL-7B	26.8 [23.6, 30.2]	10.4 [8.4, 12.5]	>0.99
Qwen2.5-VL-32B	38.7 [33.2, 44.3]	3.8 [1.9, 6.3]	0.95
Qwen3-VL-32B	30.0 [26.7, 33.3]	5.2 [3.6, 6.7]	>0.99
EmoVIT	25.1 [20.9, 29.6]	12.2 [9.1, 15.4]	>0.99
EmoVIT-recipe Qwen2.5-VL-7B	28.6 [25.0, 32.2]	8.8 [6.7, 11.1]	>0.99
EditReward	36.2 [31.0, 42.6]	4.3 [2.1, 6.8]	0.96
HPSv3	33.3 [28.2, 39.4]	7.1 [4.0, 9.9]	>0.99
HPSv3 (edit minus original)	23.0 [18.2, 28.6]	7.4 [4.7, 11.0]	>0.99
EditScore	30.1 [24.9, 35.9]	8.4 [5.2, 11.7]	>0.99
VIEScore (Qwen2.5-VL-7B)	25.5 [20.4, 30.8]	3.8 [1.7, 6.4]	0.92

What each metric's top fifth holds. Table 17 describes the human-rated failures that each metric ranks in its top fifth. A failure counts as *scene replaced* when the DINOv2 cosine between original and edit is below 0.5 (13.7% of all pairs), and as *emotion unchanged* when $|E|<0.02$ (24.2%); both thresholds were set once, at round values, and never varied. Both labels use PAEF's own signals, so they describe the baselines' failures and are no evidence for PAEF. Emotion-only metrics fail disproportionately on scene replacements (30 to 64% of their failures), scene-only metrics on edits that leave the emotion unchanged (50 to 58%), and ESMI on both.

Table 17: **Human-rated failures in each metric's top fifth** (descriptive). Scene replaced: $C<0.5$. Emotion unchanged: $|E|<0.02$. Both labels use PAEF's signals. The last two rows give the same shares over all failures, the base rate each metric row should be read against, and over all pairs.

Metric	Failures	Scene replaced %	Emotion unchanged %
Emo-S	14	64.3	0.0
FindingEmo-DA	21	57.1	0.0
Emo-A	20	35.0	15.0
CLIPScore	23	30.4	34.8
DINOv2 cosine	61	0.0	54.1
CLIP-I	72	2.8	58.3
S_{str}	52	32.7	50.0
SSIM	46	8.7	52.2
PSNR	74	6.8	54.1
−LPIPS	53	3.8	50.9
ESMI	47	44.7	23.4
PAEF	6	16.7	0.0
all 222 failures	222	21.6	42.3
all 1,406 pairs	—	13.7	24.2

Combining DA and ES. After the design was frozen, we compared the average with other ways of combining the two signals inside E, each multiplied by C (Tab. 18; ES is clipped at zero for the product and the geometric mean). On the first wave the choice makes no difference. On the held-out wave, the AND-like minimum and product lose agreement, while the maximum does slightly better than the average; we report the full sweep and keep the average, which was fixed before the held-out wave was examined.

Table 18: **Combining DA and ES inside** E (post hoc; each combination multiplied by C). P: share of resamples in which the average scores higher, on the held-out wave and on all pairs. Failures: share in the top fifth.

Combination	Wave 1	Wave 2	All	P (wave 2)	P (all)	Fail. %
average (= PAEF)	+0.213	+0.229	+0.219	—	—	2.1
max	+0.215	+0.239	+0.225	0.01	0.05	2.1
min	+0.214	+0.183	+0.191	>0.99	0.99	2.8
product	+0.214	+0.203	+0.204	0.99	0.93	4.3
geometric mean	+0.219	+0.205	+0.208	0.99	0.88	2.8

I CONTENT SWAPS AND IDENTITY EDITS

Two probes, fixed before they were scored, test the two failures of Sec. 4.1 on stimuli outside the benchmark. *Content swaps* are 62 pairs that show a real original next to a genuine edit, toward the same target, of a *different* source image; five crowd workers rated each on the benchmark's scale (mean 0.95, none at or above 3), and they are not among the 1,406 pairs. *Identity edits* are 200 originals shown as their own edit. For every metric we set the bar at which it accepts as many of the 1,406 genuine edits as the raters did (20.3%, up to ties; column "Genuine"), and count the share of each probe it accepts at that bar. A metric passes a probe if the upper end of the 95% cluster-bootstrap interval of that share is below 20.3% and it scores the probe items significantly below the genuine

edits (one-sided Mann–Whitney $p<0.05$). The learned evaluators take the two images separately, as on the benchmark. Table 1 applies the same test to the configurations between ESMI and PAEF (`experiments/iclr_06_config_sequence.py`). Two further probes were registered with these: degenerate inputs, where the edit is replaced by a black, white, grey or noise image (800 items), and direction reversals, where a genuine pair is scored against the target most opposite on the circumplex (400). Every classical metric we ran on them passes the degenerate inputs, and only the scene-only metrics, which never see the target, fail the reversals; PAEF passes both, and so does the additive $\frac{1}{2}(E+C)$, so neither probe separates addition from multiplication. The identity probe was added for that reason, and registered before it was run; all four probes and their results ship with the release.

Emotion-only metrics accept content swaps and reject identity edits; scene-only metrics do the reverse, and ESMI fails both (Tab. 19). PAEF accepts none of either: the swaps have $C=0.03$ on average (at most 0.25) against 0.75 for the genuine edits, so the gate rejects them however strongly they carry the target emotion, and identity edits have no emotion shift to gate. EditScore passes both; EditReward rejects the swaps and accepts no identity edit at its bar, but scores identity edits almost as high as genuine ones (AUC 0.54), so it fails the second test; HPSv3, which scores the edit without the original, accepts 19% of the swaps.

Table 19: **Content swaps and identity edits accepted** (%) at each metric's matched bar, with the share of genuine edits it accepts at the same bar. ✓: passes the probe; ×: fails it. [a]VIEScore parses 16 of the 62 swaps and 134 of the 200 identity edits.

Metric	Genuine	Content swaps	Identity edits
PAEF	20	0 ✓	0 ✓
Emotion only			
Emo-S (= ES)	20	32 ×	0 ✓
Emo-A	32	39 ×	8 ✓
CLIPScore	20	29 ×	4 ✓
FindingEmo-DA (= DA)	20	26 ×	0 ✓
OASIS-DA	20	29 ×	0 ✓
S_{emo} (ESMI)	20	26 ×	0 ✓
Both, added			
ESMI	20	21 ×	100 ×
Scene only			
DINOv2 cosine (= C)	20	0 ✓	100 ×
CLIP-I	20	0 ✓	100 ×
SSIM	20	0 ✓	100 ×
PSNR	20	0 ✓	100 ×
−LPIPS	20	0 ✓	100 ×
S_{str} (ESMI)	20	16 ×	100 ×
Learned evaluators			
EditReward	20	3 ✓	0 ×
HPSv3	20	19 ×	14 ✓
EditScore	21	0 ✓	0 ✓
VIEScore[a]	21	6 ✓	0 ✓

J Robustness and bias audit

PAEF is built from frozen backbones, so it can inherit their shortcuts, for instance a learned association between warm colours and joy. We measured such inheritance against the human ratings rather than against constructed stimuli alone.

Method. For each pair we take the rank residual, PAEF's percentile minus the human percentile, and regress it on a standardised cue: colour warmth, luminance, saturation and contrast measured on the pixels, and the presence of added fire, balloons or celebration objects measured by CLIP ViT-L/14 detectors, a different backbone from ES's ViT-B/32. A cue that humans weight as heavily as PAEF does cancels to a zero slope, so only *excess* reliance registers; slopes are in percentile points per

standard deviation of the cue, with cluster-bootstrap intervals. We tested 11 cue–target cells chosen in advance (Tab. 20), so a cell survives Bonferroni correction at $p<0.05/11$.

Findings. The warm-colour shortcut is absent: for positive targets the warmth slope is -0.74 with an interval spanning zero. Darkening and desaturation move the classifier toward fear and sadness, but PAEF does not weight them above the raters. One shortcut is real. PAEF over-rewards added fire when the target is anger or fear, the only cell that survives the correction. The excess comes from ES and not from DA (Tab. 21): the EmoSet-trained classifier learned that fire means anger, while the V/A probe did not, and averaging the two dilutes the prior. It concentrates in EmoEdit edits (slope $+12.99$, 95% CI $[+3.54, +20.97]$), the editor whose training pairs were selected with the same classifier. It does not carry the anger result: removing the fire pairs, or regressing the fire cue out of both scores, lowers PAEF's τ_b on anger from $+0.275$ to between $+0.140$ and $+0.249$ depending on the rule, still clearly positive. Pooled over all $1{,}406$ pairs with each target emotion as its own baseline, no cue shows excess reliance above one percentile point.

Table 20: **Excess reliance on visual cues**: slope of PAEF's rank residual on each cue, in percentile points per standard deviation. Positive means PAEF rewards the cue more than raters do. The pooled rows centre residual and cue within each target emotion.

Cue → target	n	Slope	95% CI	p
warm → positive	561	-0.74	[-3.78, +2.26]	0.6260
warm → amusement	173	+3.67	[-2.10, +9.22]	0.2213
fire → anger	244	+5.04	[+1.27, +8.83]	0.0100
fire → anger+fear	454	+4.22	[+1.46, +7.05]	0.0013
dark → sadness	155	+0.61	[-5.57, +6.61]	0.8367
dark → fear	210	-1.90	[-5.80, +2.04]	0.3460
dark → sad+fear	365	-1.08	[-4.33, +2.27]	0.5440
balloon → amusement	173	+5.76	[+0.23, +10.88]	0.0367
balloon → excitement	140	-5.64	[-12.16, +1.04]	0.1007
balloon → amus+exc	313	+1.61	[-2.59, +5.95]	0.4600
celebr → amus+exc	313	+0.63	[-3.28, +4.45]	0.7407
pooled over all targets: warmth (R-B)	1406	-0.18	[-1.88, +1.57]	0.8447
pooled over all targets: luminance	1406	-0.46	[-2.20, +1.29]	0.6187
pooled over all targets: saturation	1406	-0.67	[-2.28, +1.07]	0.4440
pooled over all targets: contrast	1406	-0.87	[-2.64, +0.82]	0.3173
pooled over all targets: fire	1406	+0.31	[-1.37, +1.98]	0.7073
pooled over all targets: balloon	1406	+0.84	[-1.11, +2.81]	0.4093
pooled over all targets: celebration	1406	-0.88	[-2.57, +0.87]	0.2907

Table 21: **Which signal carries the excess.** The same residual slope computed with ES alone, DA alone, and PAEF.

Cue → target	n	ES only [95% CI]	DA only [95% CI]	PAEF
warm → positive	561	-1.92 [-4.88, +1.42]	+0.97 [-2.08, +4.15]	-0.74
fire → anger	244	+6.37 [+2.49, +10.26]	-0.62 [-4.05, +2.86]	+5.04
dark → sad+fear	365	+0.51 [-2.84, +3.81]	-0.57 [-3.81, +2.63]	-1.08
balloon → amus+exc	313	+1.10 [-3.30, +5.36]	+0.69 [-3.38, +5.01]	+1.61

Subtle edits, faces and conflicting cues. Table 22 scores PAEF on subsets of real benchmark pairs, with the human ratings as the arbiter; the subsets were fixed before any number was computed, and those added afterwards are marked exploratory. **Edit size.** Splitting pairs into thirds by mean pixel change, agreement is lowest on the largest edits, not the smallest. **Faces.** A CLIP ViT-L/14 probe trained on 160 hand-labelled source images (92.5% cross-validated accuracy; an OpenCV Haar detector reached 71% and was discarded) flags 330 pairs whose source contains a legible face; PAEF's agreement there does not differ detectably from the rest (difference -0.015, 95% CI $[-0.099, +0.070]$). The probe detects that a face is present, not that the edit changed its expression, so expression-level sensitivity remains untested. **Conflicting cues.** A metric that tracked only "warmer and brighter" would fail where the colour change points against the instruction, such as a darker,

greyer image for *amusement*. On the 428 pairs where it does, PAEF's agreement does not drop. Its weakest subset is the opposite one, a strong colour change that agrees with the target, where it partly follows the cue, as Emo-S and FindingEmo-DA do while InternVL3.5-14B does not. Among the 12 contrasts of this family, only the strong-conflict versus strong-agreement contrast survives Bonferroni correction, and the conflict advantage is carried by the EmoEdit source images, so we claim only that conflicting cues do not break PAEF. **Balloons.** On 39 pairs that add balloons, mostly for amusement, PAEF carries no signal while InternVL3.5-14B does; Emo-S and FindingEmo-DA fail there too. **Styling alone.** On 2,000 constructed stimuli that change only colour styling (250 source images, a "happy" or "sad" look, scored toward each target), PAEF averages 17% of the score of a genuine edit, 29% in its worst cell (a sad look on positive content), and 1.3% of the stimuli reach the top fifth of real edits.

Table 22: **PAEF on subsets of AffectEdit** (Kendall τ_b against the human mean, cluster-bootstrap 95% intervals over source images).

Subset	Family	n	τ_b	95% CI
ALL (full benchmark)	reference	1406	+0.219	[+0.186, +0.253]
Person	protocol	350	+0.188	[+0.121, +0.258]
not Person	protocol	1056	+0.224	[+0.185, +0.263]
Fire added (>0.5)	protocol	67	+0.281	[+0.102, +0.442]
no fire added	protocol	1339	+0.205	[+0.171, +0.242]
Balloon added (>0.5)	protocol	39	−0.003	[−0.249, +0.242]
no balloon added	protocol	1367	+0.223	[+0.190, +0.258]
Strong colour change (top tercile)	protocol	469	+0.219	[+0.158, +0.278]
weak/medium colour change	protocol	937	+0.207	[+0.164, +0.250]
Edit size: small tercile	exploratory	469	+0.195	[+0.135, +0.253]
Edit size: medium tercile	exploratory	468	+0.259	[+0.200, +0.319]
Edit size: large tercile	exploratory	469	+0.126	[+0.062, +0.187]
Stroop conflict (cue vs target valence)	exploratory	428	+0.286	[+0.223, +0.345]
congruent cue	exploratory	978	+0.187	[+0.145, +0.226]
Strong AND conflicting cue	exploratory	115	+0.361	[+0.246, +0.468]
Strong AND congruent cue	exploratory	354	+0.085	[+0.014, +0.152]
Face [probe_face]	exploratory	330	+0.203	[+0.131, +0.279]
no face [probe_face]	exploratory	1076	+0.218	[+0.178, +0.257]
Face [face_best]	exploratory	319	+0.218	[+0.145, +0.295]
no face [face_best]	exploratory	1087	+0.214	[+0.176, +0.254]
Face [cat_person]	exploratory	350	+0.188	[+0.121, +0.258]
no face [cat_person]	exploratory	1056	+0.224	[+0.185, +0.263]
Face [clip_face]	exploratory	214	+0.180	[+0.101, +0.259]
no face [clip_face]	exploratory	1192	+0.226	[+0.189, +0.266]
Face [haar_frontal_any]	exploratory	374	+0.233	[+0.170, +0.302]
no face [haar_frontal_any]	exploratory	1032	+0.211	[+0.172, +0.251]

Content categories. A ViT-L/14 probe trained on the 843 OASIS images outside the benchmark, using OASIS's human category labels, labels the source images; on the 57 OASIS images inside the benchmark it is 93.0% accurate (95% CI $[83.3, 97.2]$). Scenes account for 38.9% of pairs, people 28.3%, objects 18.8% and animals 14.0%. Edit quality depends only slightly on category (Kruskal–Wallis on the human score, H=10.59, p=0.014, ε^2=0.005). PAEF is positive in all four categories, with every interval above zero (Tab. 23), and rater agreement is flat across them (ICC$(2, 5)$ from 0.809 to 0.837).

Table 23: **By content category.** Kendall τ_b against the human mean with 95% intervals; ICC$(2, 5)$ of the ratings within each category.

Category	n	Share	PAEF	Emo-S	IVL-14B	ICC$(2, 5)$
Scene	547	38.9%	+0.20 [+0.14, +0.25]	+0.16 [+0.10, +0.22]	+0.33 [+0.27, +0.39]	0.809
Person	398	28.3%	+0.20 [+0.14, +0.27]	+0.13 [+0.06, +0.19]	+0.31 [+0.24, +0.37]	0.824
Object	264	18.8%	+0.28 [+0.20, +0.36]	+0.24 [+0.16, +0.33]	+0.36 [+0.27, +0.45]	0.837
Animal	197	14.0%	+0.19 [+0.09, +0.29]	+0.15 [+0.06, +0.25]	+0.30 [+0.20, +0.40]	0.835