EMOWEAVE: UNIFIED AFFECTIVE IMAGE MANIPULATION VIA EMOTION-AWARE UNDERSTANDING AND GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Affective Image Manipulation (AIM) aims to modify images according to emotional intents, enabling them to evoke desired affective responses. Existing methods typically focus on either appearance or semantic manipulation, require predefined target emotions, and lack explicit mechanisms for coordinating content and appearance edits toward a shared emotional goal. We introduce Unified Affective Image Manipulation (U-AIM), a task formulation that connects emotion intent inference, editing planning, and image editing. Given a source image with an optional target emotion, U-AIM supports both emotion enhancement and emotion migration by generating and executing coordinated content and appearance editing instructions. To provide supervision for this process, we propose Emotion-Grounded Retrieval Augmentation (EGRA), which retrieves emotion-relevant evidence conditioned on the source-image state and target emotion to construct grounded editing instructions. Combining affective image data with retouching data, we build UAIM-14K, linking source images, emotional intents, coordinated instructions, and edited results. We further introduce UAIMBench to evaluate both planning and end-to-end editing under user-specified and model-inferred emotional intents. Based on this supervision, we develop EmoWeave, a unified multimodal model that integrates emotional reasoning, instruction generation, and instruction-conditioned image editing within a single framework. Experiments on UAIMBench demonstrate that EmoWeave outperforms evaluated unified multimodal baselines and applicable AIM methods in both planning quality and end-to-end editing performance.

1 INTRODUCTION

Photographs convey not only visual content but also emotional intent, which shapes how viewers perceive captured scenes. Affective Image Manipulation (AIM) aims to modify images according to emotional intents, enabling them to evoke specific affective responses (Yang et al., 2025). However, emotional intents are abstract and do not directly specify which visual elements should change. The same emotion may require different visual changes across images, so editing decisions must account for the source content and context. AIM therefore requires translating abstract emotional intents into image-specific editing decisions guided by source content and context. Existing AIM methods express target emotions through appearance adjustments (Fu et al., 2022; Liu et al., 2018; Wang et al., 2013; Weng et al., 2023), semantic modification (Yang et al., 2026; 2025), or agent-based planning with tool execution (Mao et al., 2026). However, these approaches typically assume a user-specified target emotion and do not explicitly model how content and appearance edits should be coordinated toward a shared emotional goal. Therefore, AIM requires not only emotion understanding but also the ability to infer editing intent and translate it into coherent image modifications.

To this end, we introduce Unified Affective Image Manipulation (U-AIM), a unified task that connects target emotion recommendation, emotional-driven planning, and image editing (Fig. 1). Given a source image with an optional target emotion, U-AIM infers or follows the target emotion, generates coordinated content and appearance editing instructions, and executes them to produce the edited image. U-AIM covers both emotion enhancement, which strengthens existing emotional expression, and emotion migration, which shifts the image toward a different emotion.

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.

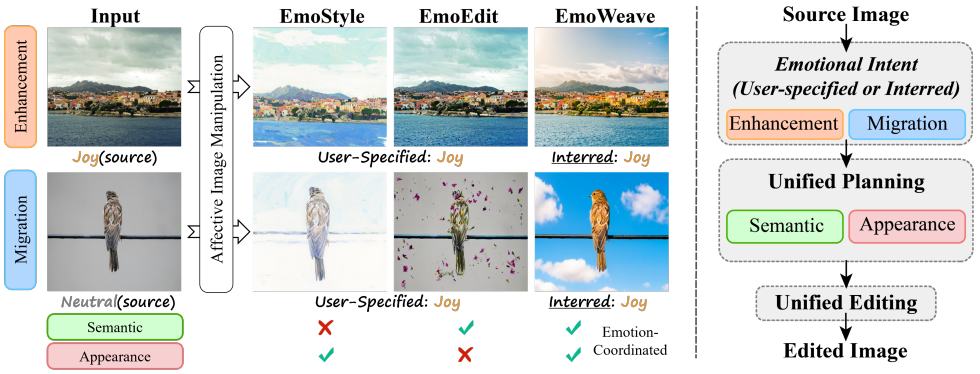

Figure 1: Overview of existing affective image editing paradigms and Unified Affective Image Manipulation (U-AIM). **(a)** Existing methods mainly support either appearance editing or content editing, whereas EmoWeave explicitly coordinates both levels toward a shared emotional intent. **(b)** U-AIM takes a source image and an optional emotional intent as input, connecting emotion inference, coordinated planning, and image editing to support both emotion enhancement and migration.

Realizing U-AIM requires a model to understand emotional cues, formulate coordinated editing plans, and execute the planned edits. Unified multimodal models (UMMs) integrate understanding and generation within a single model, providing a natural foundation for this process (Deng et al., 2025; Tian et al., 2026; Diao et al., 2026). However, general understanding and generation capabilities alone are insufficient for effective affective editing. We observe that general-purpose vision-language models (VLMs) tend to over-rely on existing cues in the source image when planning edits, particularly in emotion enhancement (§4.4). This motivates explicit supervision that connects emotional intents, editing plans, and editing outcomes, enabling the model to learn how to plan and execute edits that evoke the target emotion.

Existing affective image datasets provide emotion annotations but lack editing instructions and corresponding edited results (Yang et al., 2023; Guo et al., 2026b). To obtain before-and-after image pairs with matched instructions, we turn to a professional retouching dataset (Guo et al., 2026a), which provides detailed editing descriptions and emotionally expressive results. To bridge these complementary sources, we propose Emotion-Grounded Retrieval Augmentation (EGRA), which retrieves emotion-relevant evidence to construct grounded editing supervision. EGRA represents the source-image state in textual form and retrieves evidence conditioned on the source state and emotional intent to construct coordinated content and appearance editing instructions. Based on this strategy, we construct UAIM-14K, linking source images, emotional intents, coordinated instructions, and edited results to supervise the full U-AIM process. We further introduce UAIMBench to evaluate both planning and editing under model-inferred and user-specified emotional intents.

Building on this supervision, we develop EmoWeave based on SenseNova-U1.5 (Diao et al., 2026), integrating affective reasoning, editing planning, and image editing within a single UMM. Given a source image, EmoWeave infers an emotional intent or follows a user-specified one, generates coordinated semantic and appearance editing instructions, and executes them to produce the edited image. Experiments on UAIMBench show that EmoWeave outperforms evaluated UMM baselines and applicable AIM methods in planning quality and end-to-end performance. Comparisons with direct VLM planning and ablation studies further validate the effectiveness of EGRA.

Our main contributions are as follows:

- We introduce U-AIM, a unified task formulation that connects model-inferred or user-specified emotional intents with coordinated semantic and appearance editing, supporting both emotion enhancement and migration.
- We propose EGRA to construct coordinated planning supervision from retrieved affective evidence, and build UAIM-14K and UAIMBench for training and evaluation.
- We develop EmoWeave, which unifies emotional intent inference, editing-instruction generation, and instruction-conditioned image editing within a single UMM. Extensive experiments demonstrate its effectiveness and consistent advantages over existing UMM baselines and AIM methods.

2 Related Work

Affective Image Manipulation. Conventional computational photography and retouching methods primarily optimize perceptual quality or aesthetic preference rather than emotional intent (Gharbi et al., 2017; Chen et al., 2018; Zeng et al., 2022; Wang et al., 2021; Yang et al., 2022; Ke et al., 2023; Guo et al., 2026a; Zeng et al., 2026). Existing affective image manipulation methods express emotions through appearance adjustment, such as color and style editing (Lin et al., 2025; Yang et al., 2026), or content-level modification under user-specified emotional targets (Yang et al., 2025; Luo et al., 2026). Agent-based approaches further introduce planning and tool execution (Mao et al., 2026), but often rely on prompted orchestration rather than learned editing decisions. Overall, existing methods lack the ability to translate emotional intent into image-specific, coordinated content and appearance modifications across different editing scenarios.

Retrieval-Augmented Supervision Construction. Existing retouching datasets provide editing supervision without explicit emotional objectives (Bychkovsky et al., 2011; Liang et al., 2021), while emotion datasets provide affective annotations without corresponding executable edits (Yang et al., 2023). Retrieval-augmented methods offer a complementary way to construct supervision by distilling planning knowledge or retrieving prior experience (Yeh et al., 2025; Rajan et al., 2026; Kagaya et al., 2024; Wang et al., 2025b). Building on this direction, we use retrieval to construct emotion-grounded supervision that connects emotional intents with editing plans and image realizations.

Unified Multimodal Models. Recent unified multimodal models integrate visual understanding, reasoning, and generation within a shared architecture, enabling joint modeling of perception and generation (Diao et al., 2026; Deng et al., 2025; Xie et al., 2025; Fu et al., 2026; Tian et al., 2026). Compared with pipelines that separately connect vision-language models and image generators, unified models reduce the need to transfer intermediate decisions across independently optimized components. This makes native UMMs a suitable backbone for U-AIM, which requires emotional understanding, editing-plan generation, and image synthesis within a unified framework.

3 Methodology

3.1 Unified Affective Image Manipulation

Problem Definition. We formalize *Unified Affective Image Manipulation* (U-AIM) as a task that connects emotional intent determination, coordinated planning, and image editing. Given a source image I_0 and an optional user-specified target emotion e_{user}, U-AIM produces an edited image I^* that conveys the target emotion e^*:

$$I^* = \mathcal{F}(I_0, e^*), \quad e^* = \begin{cases} e_{\text{user}}, & \text{if provided,} \\ g(I_0), & \text{otherwise,} \end{cases} \tag{1}$$

where $g(\cdot)$ denotes the inference of a suitable target emotion from the source image. Let e_0 denote the source emotion conveyed by I_0, which may be `Neutral`. U-AIM encompasses two types of affective manipulation: (a) *enhancement*, which strengthens the existing emotion when $e_0 = e^*$, and (b) *migration*, which shifts the image toward a different target emotion when $e_0 \neq e^*$. Both are unified as image editing conditioned on the target emotion e^*.

Overview. We realize U-AIM through emotion-grounded planning and unified visual execution. A planner φ produces an editing plan $P = (u_s, u_a)$ from the source image I_0 and target emotion e^* (inferred or provided as defined in Eq. 1), where u_s and u_a specify semantic- and appearance-level edits, respectively. A unified operator D executes both instructions to produce the edited image:

$$P = \varphi(I_0, e^*), \qquad I^* = D(I_0, u_s, u_a). \tag{2}$$

This decomposition separates emotion-aware planning from visual realization while explicitly coordinating semantic- and appearance-level edits. An overview of our approach is shown in Fig. 2.

3.2 EGRA: Emotion-Grounded Retrieval Augmentation

Learning U-AIM requires supervision that links emotional intent to coordinated semantic- and appearance-level edits. As shown in Fig. 2, EGRA constructs such supervision in three stages:

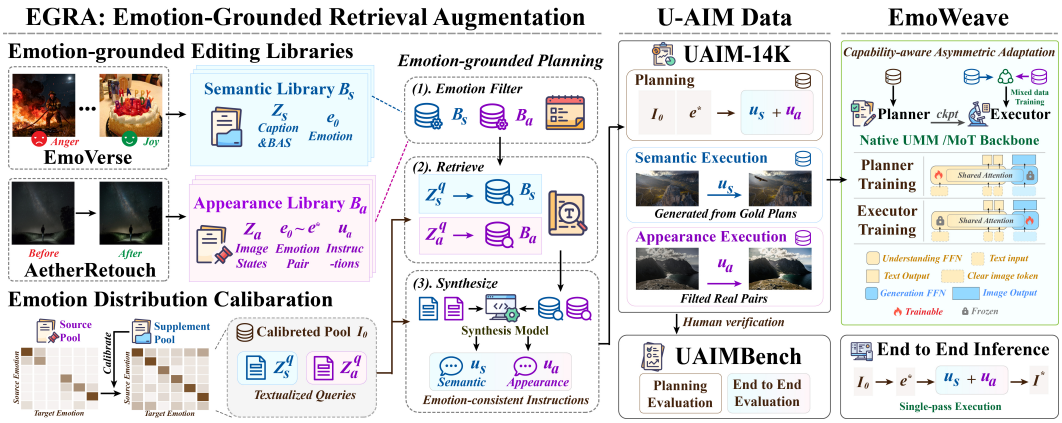

Figure 2: Overview of U-AIM. EGRA constructs emotion-grounded supervision and evaluation, while EmoWeave performs unified emotional planning and editing using UAIM-14K.

building emotion-grounded editing libraries (§3.2.1), calibrating their emotion distributions for enhancement and migration (§3.2.2), and generating planning annotations that, together with semantic- and appearance-editing data, form UAIM-14K (§3.2.3).

3.2.1 Emotion-Grounded Editing Libraries

Since no single data source jointly provides emotional intent, visual state, and editing operations, we organize AetherRetouch (Guo et al., 2026a) and EmoVerse Guo et al. (2026b) into two complementary retrieval memories for semantic and appearance editing. AetherRetouch provides real image transformation pairs with appearance-editing instructions but lacks explicit emotion grounding, whereas EmoVerse provides emotion-related semantic knowledge without executable edits:

$$B_s = \{(Z_s, e_0)\}, \qquad B_a = \{(I, Z_a, e_0, e^*, u_a, I^*)\}, \tag{3}$$

where B_s and B_a denote the semantic and appearance retrieval memories, respectively.

- **Semantic library** B_s, built from EmoVerse, provides emotion-grounded semantic evidence through Z_s, constructed from semantic descriptions and B-A-S annotations paired with emotion label e_0. To obtain a compact and representative retrieval memory, we embed Z_s and apply Mini-Batch K-Means clustering to reduce redundancy while preserving emotion-semantic coverage.
- **Appearance library** B_a, built from AetherRetouch, retains editing pairs (I, I^*), appearance instructions u_a, textualized visual states Z_a, and emotion labels e_0, e^* estimated from I and I^*, respectively. Each Z_a captures scene content, mood tonal and color attributes, and region-level visual states, linking the image state, target emotion, and executable appearance edits.

We unify all emotion annotations into the Ekman Six Emotion Space (Ekman, 1992). The representations Z_s and Z_a are embedded and indexed with FAISS Johnson et al. (2019) to construct the corresponding retrieval indices. Together, the two libraries provide complementary evidence for emotion-grounded planning. More detailed construction information refer to Appendix A.2.1.

3.2.2 Emotion Distribution Calibration

We construct the source training pool primarily from AetherRetouch, whose professionally edited image pairs reflect real-world photographic practice. However, its emotion distribution reflects common editing preferences, limiting coverage of U-AIM scenarios. We therefore calibrate this distribution to broaden visual and emotional diversity while preserving photographic realism.

Based on the emotion annotations in the appearance-editing resources (§3.2.1), we observe that Appearance library B_a is dominated by enhancement cases, *i.e.*$e_0 = e^*$. We retain these editing pairs, supplement underrepresented emotions with diverse samples from EmoVerse, and incorporate neutral-source cases to construct migration supervision. The calibrated source pool $\mathcal{S}$ is defined as

$$\mathcal{S} = \mathcal{S}_a^e \cup \mathcal{S}_a^0 \cup \mathcal{S}_e, \tag{4}$$

where $\mathcal{S}_a^e$ and $\mathcal{S}_a^0$ denote non-neutral and neutral source cases from AetherRetouch, respectively. $\mathcal{S}_e$ contains supplementary images from EmoVerse covering underrepresented emotion categories.

Supplementation follows the two U-AIM scenarios. For enhancement ($e_0 = e^*$), the source emotion already matches the target; for migration ($e_0 \neq e^*$), the source emotion is preserved while a different target is assigned, with $\mathcal{S}_a^0$ providing natural neutral starting points. Detailed sampling and quota allocation are provided in the Appendix A.1.

Fig. 3 shows the resulting source-target emotion coverage after calibration, where diagonal and off-diagonal entries correspond to enhancement and migration, respectively. The calibrated pool balances realistic photographic priors with broader U-AIM scenario coverage, providing the source basis for subsequent supervision construction.



Figure 3: Calibrated source-target emotion transition matrix.

3.2.3 UAIM-14K: Unified Affective Image Manipulation Supervision

Given the calibrated source pool, EGRA retrieves emotion-grounded evidence from B_s and B_a to construct coordinated semantic- and appearance-level supervision.

For each source image I_0, we derive textualized query representations Z_s^q and Z_a^q, aligned with the representations indexed in B_s and B_a. These queries retrieve emotion-consistent evidence from the two libraries, which is synthesized with the target emotion to produce the edits,

$$(u_s, u_a) = \Psi\Big(e^*, Z_s^q, Z_a^q, \text{retrieve}(Z_s^q, B_s), \text{retrieve}(Z_a^q, B_a)\Big), \tag{5}$$

yielding the planning tuple $(I_0, e_0, e^*, u_s, u_a)$ by a reasoning model Ψ (*e.g.*, Deepseek-V4-pro (DeepSeek-AI, 2026)). Before similarity search, e^* filters candidates associated with the target emotion, while retrieval is performed using Z_s^q and Z_a^q. We additionally collect the reasoning behind target emotion as auxiliary supervision to support autonomous target emotion recommendation.

We compare reasoning models and input modalities in §4.4. Experiments show that directly conditioning on raw images reduces planning quality, motivating EGRA to use textualized visual states and retrieved evidence for plan generation. The resulting planning annotations, selected appearance-editing pairs, and semantic-editing pairs generated by a strong editing model from gold plans together constitute **UAIM-14K**, providing supervision for both planning and editing execution. We further construct **UAIMBench** through sample filtering and manual review to evaluate planning and end-to-end editing under both model-inferred and user-specified emotional intents.

3.3 EmoWeave: Unified AIM with Capability-Aware Adaptation

We build **EmoWeave** on a native unified multimodal backbone and train it in two stages to integrate emotional intent determination, coordinated planning, and image editing, as shown in Fig. 2.

Planner Adaptation. Using the planning supervision in UAIM-14K, we impose the mapping $(I_0, e^*) \rightarrow (u_s, u_a)$ described in §3.2.3 into the planner ϕ. We train the planner ϕ to acquire emotion-related knowledge for interpreting emotional intent and generating coordinated editing instructions. During this stage, we freeze the executor to preserve its visual generation capabilities while adapting the planner for emotion-grounded understanding and planning.

Executor Adaptation. Starting from the planner-adapted checkpoint, we freeze the planner φ and the shared layers of the Mixture-of-Transformers architecture, updating only the generation-specific components of the executor D to preserve emotion-grounded planning capabilities. We mix semantic- and appearance-editing pairs from UAIM-14K to learn the mappings $(I_0, u_s) \rightarrow I_s$ and $(I_0, u_a) \rightarrow I_a$ within the same executor. At inference, we jointly condition the executor on the coordinated instructions (u_s, u_a) to generate a single edited image I^* incorporating both semantic and appearance changes in a single editing pass, $I^* = D(I_0, u_s, u_a)$. This joint execution is enabled by the coordinated semantic and appearance instruction pairs in UAIM-14K and the base model's ability to follow composite editing instructions, with its effectiveness validated under end-to-end in §4.2 and fixed joint instructions in Appendix D.3.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Implementation. EmoWeave is built on SenseNova-U1.5-8B-MoT-Preview (Diao et al., 2026) and trained in two stages on 8×H20 GPUs. DeepSeek-V4-Pro (DeepSeek-AI, 2026) generates both instructions from textual states and retrieved evidence. The planner is trained on UAIM-14K (14,907 samples) for 2 epochs with a learning rate of 5×10^{-5} while freezing generation modules. The executor is then initialized from the planner checkpoint and trained for 1 epoch at 1×10^{-5} on 23,333 editing pairs, including 13,537 appearance and 9,796 semantic editing pairs, with the understanding modules frozen. At inference, one planner call produces (u_s, u_a), followed by a single unified editing pass. Full details are provided in the Appendix B.

Benchmark. UAIMBench contains 600 images: 250 from AetherRetouch (Guo et al., 2026a) and 350 from EmoVerse (Guo et al., 2026b). The former includes 144 enhancement and 106 migration cases. Reference source and target emotions are fixed during benchmark construction using EEmo-Logic (Gao et al., 2026) and dataset annotations. Scenario A provides the target emotion (325 samples), whereas Scenario B requires the model to infer it (275 samples). Planner evaluation uses all 600 samples, while end-to-end evaluation uses a stratified subset of 100 to manage the cost of API calls across compared methods. All benchmark samples are disjoint from the training data.

Baselines. We compare with emotion-oriented methods EmoEdit (Yang et al., 2025), EmoStyle (Yang et al., 2026), and VeraRetouch (Guo et al., 2026a); general image editors FireRed-Edit (Super Intelligence Team et al., 2026) and Qwen-Image-Edit-2511 (Wu et al., 2025a); and unified multimodal models BAGEL (Deng et al., 2025) and InternVL-U (Tian et al., 2026). Target-conditioned methods are evaluated only when applicable. In Scenario B, no baseline receives the reference target emotion or our planner outputs. Planner baselines further include InternVL3.5-8B/38B (Wang et al., 2025a), Qwen3-VL-8B/32B (Bai et al., 2025), InternVL-U (Tian et al., 2026), and BAGEL-7B (Deng et al., 2025) , all evaluated zero-shot under the same task specification.

Metrics. End-to-end metrics include EmoACC (target-emotion accuracy), Emo-S (target-emotion probability gain), Semantic and Appearance (blind 1–10 ratings by GPT-5.5 and Gemini-3.1-Pro), and UniPercept Aesthetic and Quality (Cao et al., 2025). Planner metrics cover target recommendation accuracy, state grounding, executability, instruction coherence, and overall quality, macro-averaged over six emotions. Significance uses paired bootstrapping for planner/end-to-end and two-sided Wilcoxon tests for EGRA, both Holm-corrected within families ($\alpha = 0.05$; Appendix C.1).

4.2 END-TO-END EVALUATION ON U-AIM

Quantitative Results. As shown in Tab. 1 and Tab. 2, EmoWeave consistently achieves leading performance in emotional intent alignment and editing quality across both target-specified and autonomous settings. In Scenario B, EmoWeave matches the strongest baseline in EmoACC while achieving stronger emotion realization and editing quality, demonstrating effective autonomous affective manipulation. In Scenario A, EmoWeave achieves the strongest emotional alignment and semantic–appearance editing performance on both tasks, with larger gains observed in migration where the two editing levels must be jointly coordinated.

Table 1: Autonomous affective image manipulation without target emotion (Scenario B).

Method	EmoACC	Emo-S	Semantic	Appearance
VeraRetouch-Auto	**0.6363**	<u>-0.0121</u>	4.6932	4.6818
FireRed-Edit	0.4318	-0.1211	5.3182	5.5909
Qwen-Image-Edit	0.5455	-0.0354	5.5568	<u>5.9091</u>
BAGEL	<u>0.5682</u>	-0.0356	<u>5.8864</u>	4.6818
InternVL-U	0.5455	-0.0308	5.4091	4.8522
EmoWeave	**0.6363**	**0.0190**	**6.2046**	**6.4887**

Qualitative Results. Fig. 4 presents representative end-to-end comparisons under both target-specified (Scenario A) and autonomous (Scenario B) affective image manipulation. Existing methods often struggle to coordinate semantic and appearance modifications toward the intended emotional intent, whereas EmoWeave produces more coherent and image-grounded edits aligned with the target affective objective while preserving scene consistency.

Table 2: Comparison on target-specified emotional retouching (Scenario A). Semantic evaluates semantic editing quality, while Appearance evaluates appearance retouching quality.

Method	EmoACC ↑	Emo-S ↑	Semantic ↑	Appearance ↑	Aesthetic ↑	Quality ↑
(a) Emotion Enhancement						
EmoEdit	0.6000	0.0131	4.9333	–	58.5167	59.5500
EmoStyle	0.4333	-0.1960	4.0333	–	63.6083	57.1208
VeraRetouch	0.6667	0.0193	–	6.4833	62.8292	63.1667
FireRed-Edit	0.7333	0.0897	6.0667	6.3500	64.4917	**69.9500**
Qwen-Image-Edit-2511	0.7333	0.0637	6.2833	6.4500	63.4625	65.8250
BAGEL	0.6333	0.0169	5.9833	4.9500	48.2417	49.0792
InternVL-U	0.6667	-0.0046	5.7667	5.0667	57.0333	59.6708
EmoWeave	**0.8333**	**0.1491**	**6.9667**	**7.0500**	**66.0792**	66.2750
(b) Emotion Migration						
EmoEdit	0.3077	0.1777	4.4231	–	55.7981	57.9135
EmoStyle	0.2692	0.1834	3.8269	–	66.0048	56.4423
VeraRetouch(Style-mode)	0.1923	0.1015	–	5.3654	61.8029	59.3558
FireRed-Edit	0.3846	0.2393	5.8846	6.1731	**66.9231**	**69.9231**
Qwen-Image-Edit-2511	0.3462	0.1863	6.0000	6.3462	61.1298	63.6058
BAGEL	0.3846	0.2680	5.6346	4.7885	49.3654	49.4856
InternVL-U	0.3076	0.1738	5.5200	5.2600	56.8450	59.0000
EmoWeave	**0.5000**	**0.3572**	**6.3654**	**6.7731**	60.6154	64.1923

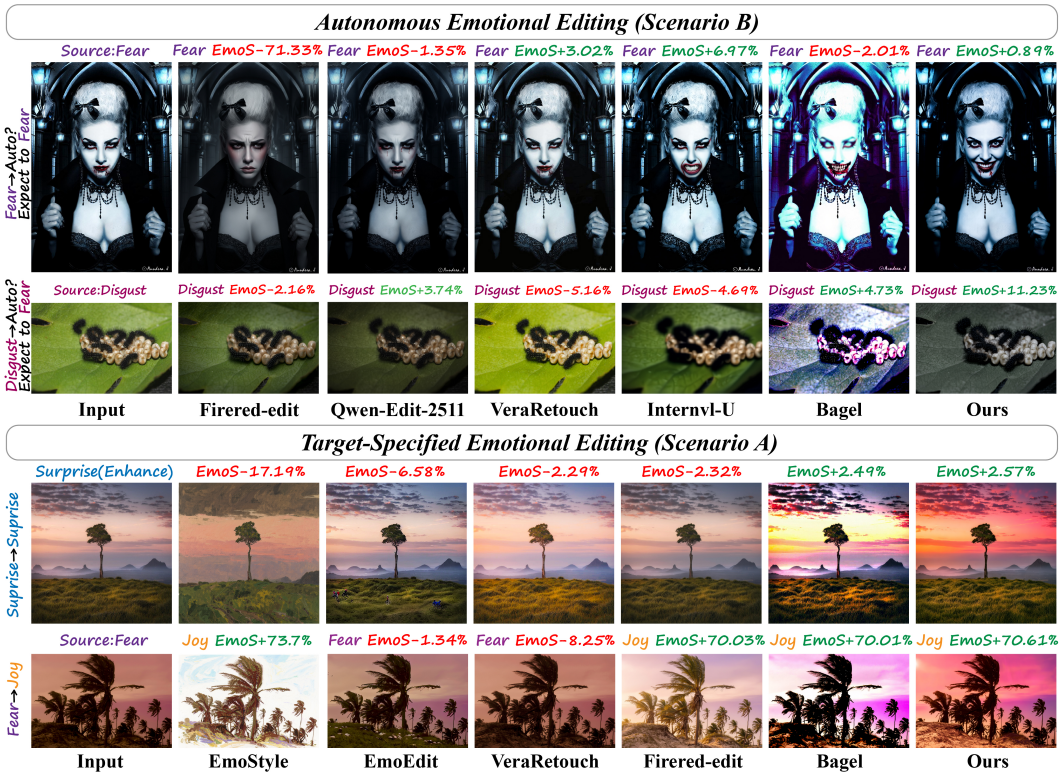

Figure 4: **Qualitative end-to-end comparison.** Examples of autonomous affective image manipulation (Scenario B, top) and target-specified affective image manipulation (Scenario A, bottom). Annotations indicate the target/predicted emotion and Emo-S change.

4.3 Planner Evaluation

Quantitative Results. EmoWeave's planner achieves the best overall planning quality across both scenarios (Tab. 3). Its clearest advantage lies in Executability, substantially outperforming VLM baselines in both target-specified and autonomous settings. Although some baselines achieve higher Instruction Coherence in Scenario B, coherent instructions alone are insufficient when the inferred emotional intent is incorrect, as the resulting plan may optimize toward the wrong objec-

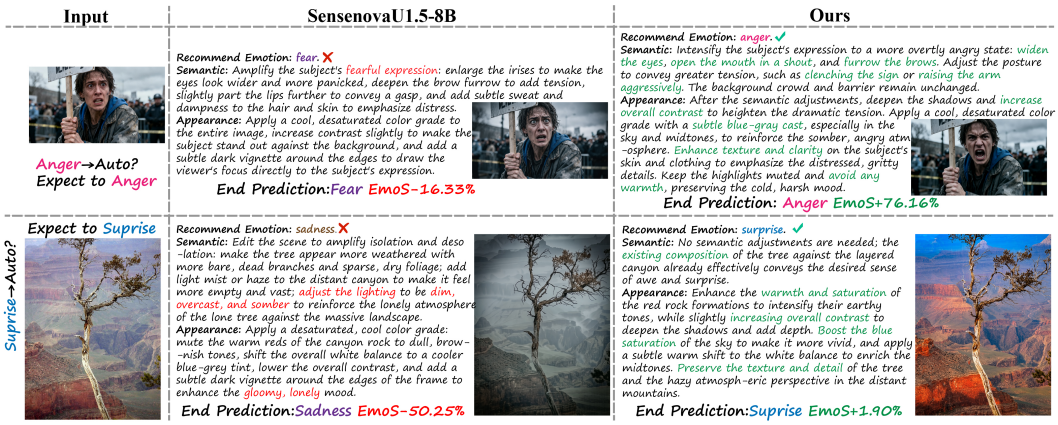

Figure 5: **Qualitative comparison of autonomous emotion planning and editing.** SenseNova-U1.5-8B and EmoWeave predictions with emotional intent inference and image-grounded instructions. ✓/× denote correct/incorrect emotion inference.

Table 3: **Comparison of emotion-grounded planning quality.** Scenario B requires autonomous emotional intent inference, while Scenario A evaluates planning with a specified target emotion. We evaluate state grounding, executability, instruction coherence, and overall planning quality.

Scenario B: Target-Recommendation Planning

Method	Recommend Acc. ↑	State Grounding ↑	Executability ↑	Instruction Coherence ↑	Overall Quality ↑
InternVL3.5-8B	0.4861	0.6764	0.3972	0.6524	0.6173
InternVL3.5-38B	<u>0.5571</u>	0.6784	0.3560	0.6447	0.6075
Qwen3-VL-8B	0.4974	<u>0.7712</u>	0.4440	<u>0.7284</u>	0.6944
Qwen3-VL-32B	0.4571	0.7398	0.3733	0.6966	0.6688
SenseNova-U1.5-8B	0.4685	0.7662	<u>0.5508</u>	**0.7872**	<u>0.7425</u>
InternVL-U	0.4764	0.6968	0.4509	0.6721	0.6457
BAGEL-7B	0.5311	0.5897	0.4321	0.4950	0.5413
EmoWeave Planner	**0.7236**	**0.8262**	**0.8579**	0.7126	**0.8121**

Scenario A: Target-Specified Planning

Method	State Grounding ↑	Executability ↑	Instruction Coherence ↑	Overall Quality ↑
InternVL3.5-8B	0.6108	0.3765	0.6215	0.5955
InternVL3.5-38B	0.6381	<u>0.5521</u>	<u>0.7106</u>	<u>0.6701</u>
Qwen3-VL-8B	0.6645	0.3842	0.6951	0.6500
Qwen3-VL-32B	<u>0.6978</u>	0.2957	0.6388	0.6251
SenseNova-U1.5-8B	0.6844	0.4237	0.6033	0.6357
InternVL-U	0.6455	0.4074	0.6364	0.6078
BAGEL-7B	0.5447	0.4801	0.5215	0.5503
EmoWeave Planner	**0.7624**	**0.8263**	**0.7696**	**0.7952**

tive. EmoWeave instead achieves the highest emotional intent inference accuracy while maintaining grounding and executability, yielding the best overall planning quality under autonomous planning.

Qualitative Results. The planning outputs shown in the lower panel of Fig. 5 provide qualitative evidence for the quantitative improvements. Compared with the unified multimodal baseline, EmoWeave generates more image-grounded and executable instructions, while the baseline often produces generic editing descriptions. In Scenario B, EmoWeave further infers more appropriate emotional intents together with coherent editing plans, consistent with its higher emotion intent inference accuracy and planning scores.

4.4 Ablation Study

Emotion-Grounded Planning Supervision. We compare four strategies for constructing emotion-grounded planning supervision under two independent judges (Fig. 6). DeepSeek+EGRA achieves the highest or joint-highest top-ranked rate across judges and task subsets, validating EGRA as an effective supervision strategy. The results also suggest that state-only textual grounding reduces reliance on existing pixel-level cues during affective reasoning.

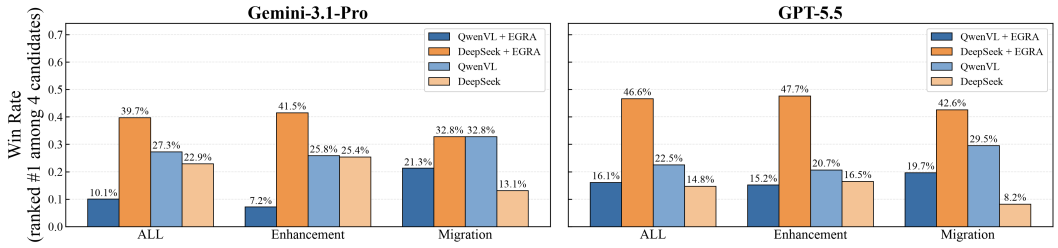

Figure 6: **Ablation of EGRA for emotion-grounded planning supervision.** We compare four supervision construction strategies using pairwise win rates from two independent judges, Gemini-3.1-Pro and GPT-5.5. DeepSeek+EGRA achieves the highest win rate, demonstrating the effectiveness of retrieval-grounded supervision construction.

Table 4: Ablation of planning and editing adaptation for unified affective image manipulation.

Training Variant	EmoACC ↑	Emo-S ↑	Semantic ↑	Appearance ↑
Full Model	**0.6600**	**0.1459**	**6.4750**	**6.7050**
w/o Executor Training	0.6300	0.1006	6.3250	6.5850
w/o Planner & Executor Training	0.6100	0.1128	6.3244	6.5123

Training Stage Gains. Tab. 4 evaluates the contribution of planner and executor adaptation. The full model consistently performs best across all reported metrics; executor removal mainly degrades Semantic and Appearance, while removing both stages further weakens overall editing performance, despite a non-monotonic Emo-S trend in the intermediate variants.

Qualitative Ablation. Fig. 7 further visualizes how the learned capabilities accumulate. The planner-only variant produces limited visual changes, semantic adaptation introduces target-related content modifications, and the full model further achieves more coherent appearance editing, consistent with the quantitative gains above.

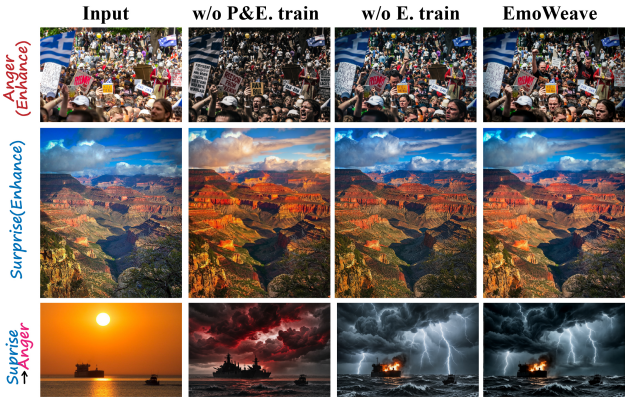

Figure 7: Qualitative ablation of planner and executor adaptation. Removing either stage weakens emotional realization, while EmoWeave produces more coherent semantic and appearance changes.

5 Conclusion

We introduced U-AIM, which connects emotional intent determination, coordinated semantic and appearance planning, and image editing to support both emotion enhancement and migration. To address the supervision gap, we developed EGRA and constructed UAIM-14K, alongside UAIM-Bench for evaluating planning and editing. Building on this supervision, EmoWeave learns emotion-grounded planning and instruction-conditioned editing within a single unified multimodal model through capability-aware asymmetric training. Experiments demonstrate improved overall planning quality, emotional alignment, and editing appropriateness under both user-specified and model-inferred targets. Ablations further support the benefits of retrieval-grounded supervision and adapting both planning and execution. Together, these findings highlight the value of explicitly connecting emotional intent with image-specific editing decisions, providing a foundation for more nuanced and emotion-guided image manipulation.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Vladimir Bychkovsky, Sylvain Paris, Eric Chan, and Frédo Durand. Learning photographic global tonal adjustment with a database of input/output image pairs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 97–104, 2011.

Shuo Cao, Jiayang Li, Xiaohui Li, Yuandong Pu, Kaiwen Zhu, Yuanting Gao, Siqi Luo, Yi Xin, Qi Qin, Yu Zhou, Xiangyu Chen, Wenlong Zhang, Bin Fu, Yu Qiao, and Yihao Liu. Unipercept: Towards unified perceptual-level image understanding across aesthetics, quality, structure, and texture. *arXiv preprint arXiv:2512.21675*, 2025.

Yu-Sheng Chen, Yu-Ching Wang, Man-Hsin Kao, and Yung-Yu Chuang. Deep photo enhancer: Unpaired learning for image enhancement from photographs with gans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6306–6314, 2018.

DeepSeek-AI. Deepseek-v4: Towards highly efficient million-token context intelligence. *arXiv preprint arXiv:2606.19348*, 2026.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, Guang Shi, and Haoqi Fan. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025.

Haiwen Diao, Penghao Wu, Hanming Deng, Jiahao Wang, Shihao Bai, Silei Wu, Weichen Fan, Wenjie Ye, Wenwen Tong, Xiangyu Fan, Li Yan, Yubo Wang, Zhijie Cao, Zhiqian Lin, Zhitao Yang, Zhongang Cai, Yuwei Niu, Yue Zhu, Bo Liu, et al. Sensenova-u1: Unifying multimodal understanding and generation with neo-unify architecture. *arXiv preprint arXiv:2605.12500*, 2026.

Paul Ekman. An argument for basic emotions. In Nancy L. Stein, Barry Leventhal, and Tom Trabasso (eds.), *Cognition and Emotion*, pp. 169–200. Lawrence Erlbaum Associates, 1992.

Fengyi Fu, Mengqi Huang, Shaojin Wu, Yunsheng Jiang, Yufei Huo, Hao Li, Yinghang Song, Fei Ding, Jianzhu Guo, Qian He, Zheren Fu, Zhendong Mao, and Yongdong Zhang. Lance: Unified multimodal modeling by multi-task synergy. *arXiv preprint arXiv:2605.18678*, 2026.

Tsu-Jui Fu, Xin Eric Wang, and William Yang Wang. Language-driven artistic style transfer. In *European conference on computer vision*, pp. 717–734. Springer, 2022.

Lancheng Gao, Ziheng Jia, Zixuan Xing, Wei Sun, Huiyu Duan, Guangtao Zhai, and Xiongkuo Min. Eemo-logic: A unified dataset and multi-stage framework for comprehensive image-evoked emotion assessment. *arXiv preprint arXiv:2602.01173*, 2026.

Michaël Gharbi, Jiawen Chen, Jonathan T. Barron, Samuel W. Hasinoff, and Frédo Durand. Deep bilateral learning for real-time image enhancement. *ACM Transactions on Graphics*, 36(4):118, 2017.

Yihong Guo, Youwei Lyu, Jiajun Tang, Yizhuo Zhou, Hongliang Wang, Jinwei Chen, Changqing Zou, and Qingnan Fan. Veraretouch: A lightweight fully differentiable framework for multi-task reasoning photo retouching. *arXiv preprint arXiv:2604.27375*, 2026a.

Yijie Guo, Dexiang Hong, Weidong Chen, Zihan She, Cheng Ye, Xiaojun Chang, Yongdong Zhang, and Zhendong Mao. Emoverse: A mllms-driven emotion representation dataset for interpretable visual emotion analysis. *IEEE Transactions on Affective Computing*, 2026b.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547, 2019.

Tomoyuki Kagaya, Thong Jing Yuan, Yuxuan Lou, Jayashree Karlekar, Sugiri Pranata, Akira Kinose, Koki Oguri, Felix Wick, and Yang You. Rap: Retrieval-augmented planning with contextual memory for multimodal llm agents. *arXiv preprint arXiv:2402.03610*, 2024.

Zhanghan Ke, Yuhao Liu, Lei Zhu, Nanxuan Zhao, and Rynson W. H. Lau. Neural preset for color style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14173–14182, 2023.

Jie Liang, Hui Zeng, Miaomiao Cui, Xuansong Xie, and Lei Zhang. Ppr10k: A large-scale portrait photo retouching dataset with human-region mask and group-level consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 653–661, 2021.

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

Da Liu, Yaxi Jiang, Min Pei, and Shiguang Liu. Emotional image color transfer via deep learning. *Pattern Recognition Letters*, 110:16–22, 2018.

Jiamin Luo, Xuqian Gu, Jingjing Wang, and Jiahong Lu. Towards LLM-centric affective visual customization via efficient and precise emotion manipulating. In *Proceedings of the ACM Web Conference 2026*. ACM, 2026. doi: 10.1145/3774904.3792585.

Qi Mao, Haobo Hu, Yujie She, Difei Gao, Haokun Chen, and Libiao Jin. Emoagent: A multi-agent framework for diverse affective image manipulation. *IEEE Transactions on Affective Computing*, 17(2):1618–1635, 2026. doi: 10.1109/TAFFC.2026.3672501.

Anirudh Sundara Rajan, Krishna Kumar Singh, and Yong Jae Lee. From plans to pixels: Learning to plan and orchestrate for open-ended image editing. *arXiv preprint arXiv:2605.15181*, 2026.

Super Intelligence Team, Changhao Qiao, Chao Hui, Chen Li, Cunzheng Wang, Dejia Song, Jiale Zhang, Jing Li, Qiang Xiang, Runqi Wang, Shuang Sun, Wei Zhu, Xu Tang, Yao Hu, Yibo Chen, Yuhao Huang, Yuxuan Duan, Zhiyi Chen, and Ziyuan Guo. Firered-image-edit-1.0 technical report. *arXiv preprint arXiv:2602.13344*, 2026.

Changyao Tian, Danni Yang, Guanzhou Chen, Erfei Cui, Zhaokai Wang, Yuchen Duan, Penghao Yin, Sitao Chen, Ganlin Yang, Mingxin Liu, Zirun Zhu, Ziqian Fan, Leyao Gu, Haomin Wang, Qi Wei, Jinhui Yin, Xue Yang, Zhihang Zhong, Qi Qin, Yi Xin, Bin Fu, Yihao Liu, Jiaye Ge, Qipeng Guo, Gen Luo, Hongsheng Li, Yu Qiao, Kai Chen, and Hongjie Zhang. Internvl-u: Democratizing unified multimodal models for understanding, reasoning, generation and editing. *arXiv preprint arXiv:2603.09877*, 2026.

Tao Wang, Yong Li, Jingyang Peng, Yipeng Ma, Xian Wang, Fenglong Song, and Youliang Yan. Real-time image enhancer via learnable spatial-aware 3d lookup tables. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2471–2480, 2021.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, Guanzhou Chen, Zichen Ding, Changyao Tian, Zhenyu Wu, Jingjing Xie, Zehao Li, Bowen Yang, Yuchen Duan, Xuehui Wang, Songze Li, Xiangyu Zhao, Haodong Duan, Nianchen Deng, Bin Fu, Yinan He, Yi Wang, Conghui He, Botian Shi, Junjun He, Yingtong Xiong, Han Lv, Lijun Wu, Wenqi Shao, Kaipeng Zhang, Huipeng Deng, Biqing Qi, Jiaye Ge, Qipeng Guo, Wenwei Zhang, Wanli Ouyang, Limin Wang, Min Dou, Xizhou Zhu, Tong Lu, Dahua Lin, Jifeng Dai, Bowen Zhou, Weijie Su, Kai Chen, Yu Qiao, Wenhai Wang, and Gen Luo. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025a.

Xiaohui Wang, Jia Jia, and Lianhong Cai. Affective image adjustment with a single word. *The Visual Computer*, 29(11):1121–1133, 2013.

Zihan Wang, Yaohui Zhu, Gim Hee Lee, and Yachun Fan. Navrag: Generating user demand instructions for embodied navigation through retrieval-augmented llm. *arXiv preprint arXiv:2502.11142*, 2025b.

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 Zhu, Yuxuan Zhang, Hao Liu, Zhe Xu, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025a.

Daiqing Wu, Dongbao Yang, Can Ma, and Yu Zhou. Emocaliber: Advancing reliable visual emotion comprehension via confidence verbalization and calibration. *arXiv preprint arXiv:2512.15528*, 2025b.

Jinheng Xie, Zhenheng Yang, and Mike Zheng Shou. Show-o2: Improved native unified multimodal models. *arXiv preprint arXiv:2506.15564*, 2025.

Canqian Yang, Meiguang Jin, Xu Jia, Yi Xu, and Ying Chen. Adaint: Learning adaptive intervals for 3d lookup tables on real-time image enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17522–17531, 2022.

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

Jingyuan Yang, Zihuan Bai, and Hui Huang. Emostyle: Emotion-driven image stylization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 594–603, 2026.

Chun-Hsiao Yeh, Yilin Wang, Nanxuan Zhao, Richard Zhang, Yuheng Li, Yi Ma, and Krishna Kumar Singh. Beyond simple edits: X-planner for complex instruction-based image editing. *arXiv preprint arXiv:2507.05259*, 2025.

Hui Zeng, Jianrui Cai, Lida Li, Zisheng Cao, and Lei Zhang. Learning image-adaptive 3d lookup tables for high performance photo enhancement in real-time. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(4):2058–2073, 2022. doi: 10.1109/TPAMI.2020.3026740.

Ying Zeng, Miaosen Luo, Guangyuan Li, Yang Yang, Ruiyang Fan, Linxiao Shi, Qirui Yang, Jian Zhang, Chengcheng Liu, Siming Zheng, Jinwei Chen, Bo Li, and Peng-Tao Jiang. Smartphotocrafter: Unified reasoning, generation and optimization for automatic photographic image editing. *arXiv preprint arXiv:2604.19587*, 2026.

APPENDIX

This appendix provides supporting details for reproducibility and further analysis. We describe the data sources and statistics (§A.1), the EGRA construction pipeline (§A.2), and the construction and splits of UAIMBench (§A.3). We then detail the training configuration and CAST (§B.1), followed by the inference procedure and baseline configurations (§B.2). The evaluation protocol includes complete metric definitions, aggregation rules, and statistical testing (§C.1), together with judge rubrics and prompts (§C.2). We then present additional quantitative and qualitative results, executor-level ablations, and operator-specific capability analysis (§D). Finally, we discuss limitations (§E) and provide a reproducibility statement (§F).

A DATA CONSTRUCTION AND BENCHMARK DETAILS

A.1 DATA SOURCES AND STATISTICS

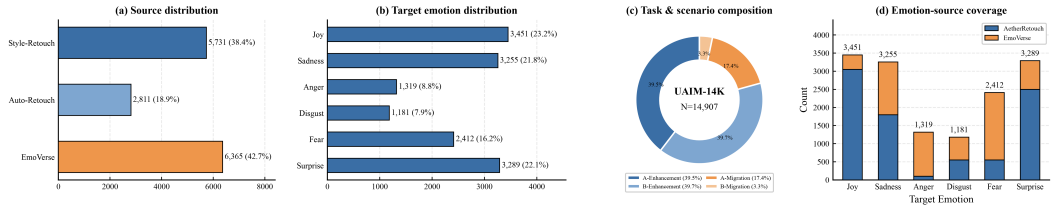

Figure 8: Statistics of UAIM-14K. (a) Source composition. (b) Coverage of the six target emotions. (c) Composition across task types and scenarios. (d) Per-source contribution to emotional coverage.

UAIM-14K combines appearance-editing supervision from AetherRetouch with emotion-grounded B-A-S annotations from EmoVerse (Fig. 8). AetherRetouch provides preset-based Style-Retouch pairs verified against editing parameters and professionally graded Auto-Retouch pairs with instructions VLM-translated from difference-derived pseudo-parameters.

EmoVerse supplements emotions sparse in retouching data, particularly anger and disgust. Its labels are aligned to Ekman's six emotions by merging amusement, contentment, and excitement into joy and mapping awe to surprise, with the remaining categories unchanged. Class balancing combines clustering of majority classes with broad-match retrieval expansion of minority classes, without synthetic oversampling (§A.2).

Enhancement and neutral-source migration are split approximately equally between Scenarios A and B via deterministic hashing, whereas non-neutral-source migration is assigned exclusively to Scenario A, accounting for its higher prevalence within migration. UAIM-14K and the executor training set contain 12,393 and 13,381 unique source images, respectively, with 8,217 shared images. The difference between records and unique images results from multiple edits associated with the same source image. This overlap primarily reflects semantic-editing pairs constructed from UAIM-14K, while appearance-editing pairs are drawn from the broader AetherRetouch corpus (§B.1).

A.2 EGRA CONSTRUCTION PIPELINE

Overview. EGRA constructs offline planning supervision by generating *directional* instructions from textual image states and retrieved evidence, while leaving edit *magnitude* to the image-conditioned operator. This supports one-shot planning while allowing execution to adapt to intermediate image changes.

The pipeline comprises four stages: (1) **state extraction**, completing semantic and appearance representations for each image; (2) **unified retrieval**, querying B_s and B_a by target emotion and state similarity; (3) **instruction generation**, jointly producing $\{e^\star, u_s, u_a\}$ from textual states and retrieved evidence; and (4) **verification**, retaining samples that pass all four image-conditioned audit checks. Database construction is detailed in §A.2.1, and retrieval, generation, and verification in §A.2.2.

A.2.1 RETRIEVAL DATABASE CONSTRUCTION

Appearance database B_a. B_a combines two confidence tiers from AetherRetouch-1M+. Tier L1 contains preset-based Style-Retouch pairs verified against the associated parameters θ, retaining their original instructions and all valid presets per input image. Tier L2 contains professionally graded Auto-Retouch pairs without native instructions: difference-derived pseudo-parameters $\hat{\theta}$ are VLM-translated into instructions u_a, which are checked against $\hat{\theta}$ before inclusion. Each entry is stored as $r_a = \{I, I^\star, Z, e, e^\star, u_a\}$, with $e = \mathrm{VER}(I)$ and $e^\star = \mathrm{VER}(I^\star)$, where VER denotes EEmo-Logic (Gao et al., 2026).

The raw distribution is dominated by joy (66%) and neutral (15%), with sparse anger and disgust coverage. Neutral-target entries are excluded from the six-emotion retrieval pool. We bucket entries by $e^\star$ and balance each class separately. Joy is compressed using MiniBatch K-Means ($K = 2{,}968$) on `scene_brief` embeddings to preserve scene diversity within the emotion class. Each cluster retains its highest-confidence L1 representative, or its medoid if only L2 entries are available. Sadness, surprise, and fear are retained unchanged. Broadened dominant-$\cup$-ranking matching expands anger from 38 to 55 entries (38 strict + 17 broad) and disgust from 244 to 538 entries (244 strict + 294 broad), with broad-matched entries retained directly in the bucket rather than as a separate larger candidate pool. These matches are tagged `match_type=broad` as per-entry metadata, distinguishing them from `strict` matches for downstream re-ranking discount. Together, these steps reduce the largest-to-smallest class ratio from approximately $435\times$ (joy vs. anger) to $54\times$ (joy vs. anger) without oversampling. Figure 9 visualizes operation-type diversity within B_a after balancing.

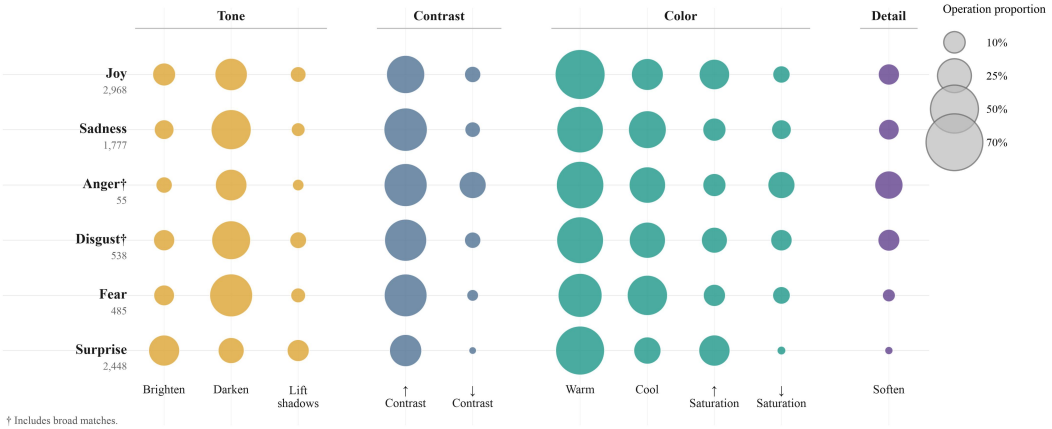

Figure 9: Distribution of appearance-editing operation types across the six target emotions in B_a after category balancing. Bubble size indicates the proportion of retained entries containing each operation (Tone: Brighten/Darken/Lift Shadows; Contrast: increase/decrease; Color: Warm/Cool/increase/decrease Saturation; Detail: Soften). Sample counts per emotion are shown below each label; † indicates broad-match-expanded classes.

Operation composition varies systematically by emotion rather than collapsing to a single dominant technique: *sadness*, *fear*, and *disgust* skew heavily toward Darken and increased Contrast, *surprise* skews toward Brighten and Warm color shifts, and *anger* shows a comparatively balanced distribution between increasing and decreasing Contrast rather than favoring one direction as the other five emotions do. This operation-level diversity, preserved even for the broad-match-expanded anger and disgust buckets, supports our category-balancing strategy (§A.2.1) as retaining varied editing evidence rather than converging on templated single-operation entries per class.

Semantic database B_s. B_s is built from EmoVerse's 234,191 Background-Attribute-Subject (B-A-S) triplets with emotion labels and semantic descriptions, using the Mikels-to-Ekman mapping described above. Amusement, contentment, and excitement are pooled into joy before clustering to avoid overrepresenting this class. Awe maps to surprise based on its expectation-violation component, while excitement remains grouped with joy. Across all six target emotions, B-A-S text is embedded and clustered within each class, retaining the highest-intensity original record per cluster

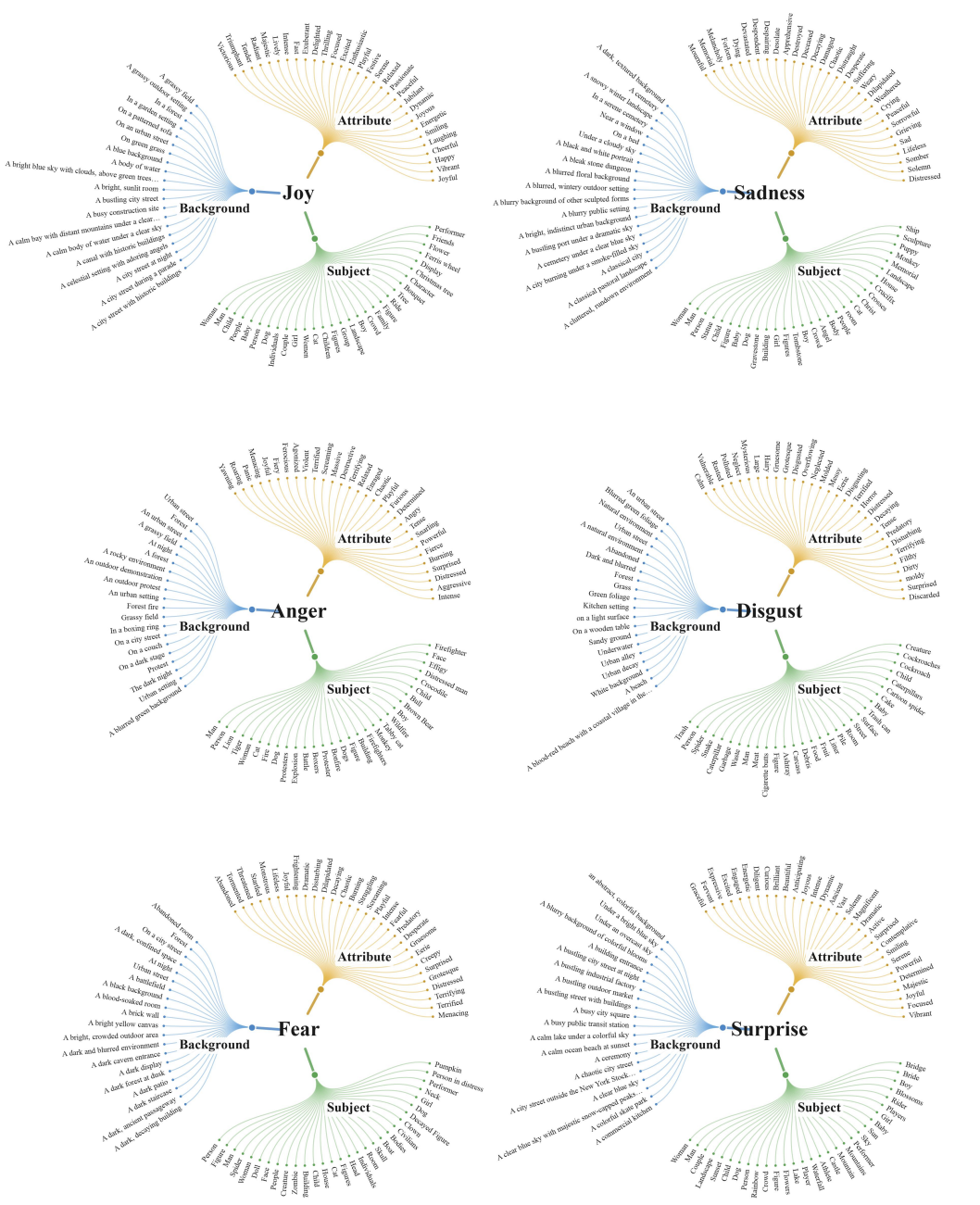

Figure 10: Representative Background/Attribute/Subject (B-A-S) vocabulary in B_s for each of the six target emotions, visualized as radial trees rooted at the emotion label.

to obtain approximately 300 representatives per emotion. Unlike B_a, B_s provides textual semantic evidence without paired edited images. Figure 10 visualizes representative B-A-S vocabulary per emotion after alignment and clustering. Each field captures a distinct axis of variation: Background spans concrete settings (e.g., *a grassy field*, *a bustling city street* for joy; *a dark, decaying building*, *a blood-soaked room* for fear), Attribute captures emotion-congruent descriptors at varying intensity (e.g., *peaceful* through *joyful* for joy; *despairing* through *desolate* for sadness), and Subject spans both human and non-human entities (e.g., *cockroaches*, *caterpillars* for disgust; *firefighter*, *bull* for anger). This three-field structure gives B_s coverage across scene, tone, and content diversity within each emotion class, complementing B_a's appearance-operation diversity (Fig. 9) with semantic-level variation the appearance database does not capture.

State extraction. We retain source-provided representations where available. For EmoVerse, EEmo-Logic extracts B-A-S annotations and semantic descriptions to form Z_s. For AetherRetouch, Qwen3-VL-32B extracts the appearance state Z_a, completing the representations required for dual-library retrieval.

Representation and indexing. Both libraries use 768-dimensional `bge-base-en-v1.5` embeddings with L2 normalization and FAISS `IndexFlatIP` for exact cosine-similarity search. For B_s, B-A-S annotations and semantic descriptions are embedded jointly as Z_s, with one index per target emotion. For B_a, Z_a is indexed through three channels: `scene_brief_mood` (concatenated scene and mood text), `tone_color`, and `region`, yielding 18 indexes across six emotions. Each region is embedded separately. Region matching compares equally named regions and averages their similarities over query regions, assigning zero to unmatched names.

A.2.2 Emotion-Grounded Instruction Generation

Unified retrieval. During offline construction, the reference target emotion $e^\star$ selects the corresponding pools in B_s and B_a, without filtering by source emotion. Each search retrieves 20 neighbors, and each library returns $K = 5$ references. For B_a, candidates from the three channels are merged by record ID and re-scored using the stored embeddings. Channel similarities are averaged with equal weights over channels available in the query. For B_s, candidates are scored by cosine similarity between the joint semantic embeddings. Both libraries use MMR with $\lambda = 0.7$ and a pool-size threshold of 600; larger pools use direct top-K selection. For appearance retrieval, MMR computes redundancy using the normalized mean of the participating channel embeddings.

Instruction generation. We use DeepSeek-V4-Pro to jointly generate $\{e^\star, u_s, u_a\}$ from textual states, reference emotions, and retrieved evidence. Text-only generation is the default; image-conditioned regeneration is triggered by low retrieval confidence, weak evidence, or a directional error detected during verification. The teacher receives both reference emotions regardless of scenario, while Scenario A/B determines the downstream planner prompt and output schema. Scenario A specifies the target and uses $\{e^\star, u_s, u_a\}$; Scenario B withholds it and uses $\{e^\star, u_s, u_a, \text{rationale}\}$, with the generated target constrained to the reference and the rationale retained for interpretability. The teacher's access to reference targets during construction does not extend to Scenario-B benchmark inference. Scenario B is assigned at a 50% rate among enhancement ($e = e^\star$) and neutral-source migration samples via deterministic per-record hashing; non-neutral-source migration is always Scenario A. Retrieved top-K appearance instructions anchor u_a: their editing directions and techniques must be preserved, while the wording is adapted to Z_I rather than copied verbatim.

Multi-dimension verification. Qwen3.7 plus audits generated plans using the source image, target emotion, and editing instructions. Images are resized to a maximum side length of 1024 pixels and encoded as JPEG with quality 90. The audit returns pass/fail judgments for image–instruction fit, emotion consistency, cross-instruction consistency, and domain validity. Strict filtering retains only samples passing all four checks. The pipeline retains 14,907 of 14,952 source records (99.7%), with exclusions covering failed audits and unresolved image paths.

Human verification. Human spot checks were conducted throughout data construction, with two to three reviewers involved across stages. These checks supplemented automated verification in retrieval database construction, instruction generation, and benchmark preparation.

A.3 UAIMBench Construction and Splits

Source pool. UAIMBench comprises the four source categories in Table 5. For AetherRetouch, source and target emotions are obtained using EEmo-Logic (Gao et al., 2026) on the input and reference edited images, respectively. For EmoVerse, source labels follow the Mikels-to-Ekman mapping (§A.2.1), with targets assigned by the sampling rule. Anger is absent from categories ②/③.

Scenario assignment. For AetherRetouch, enhancement and neutral-source migration are assigned to Scenarios A/B via deterministic hashing with equal probability; non-neutral-source migration is always Scenario A. EmoVerse retains its predefined subsets: Scenario-B enhancement,

Table 5: Source composition of UAIMBench.

Category	Source category	N
①	AetherRetouch: enhancement	144
②	AetherRetouch: migration from non-neutral sources	57
③	AetherRetouch: migration from neutral sources	49
④	EmoVerse	350
	Total	600

Scenario-A enhancement, and Scenario-A migration with assigned targets. The resulting benchmark contains 325 Scenario A and 275 Scenario B instances.

Evaluation protocols. UAIMBench evaluates three complementary aspects. **Planning quality** uses the full source pool under both scenarios, with target recommendation assessed only in Scenario B. **Joint execution** evaluates $(I_0, u_s, u_a) \rightarrow I^\star$ using the same high-quality instruction pairs generated by our planner for all editing models, isolating execution capability from planning quality. **End-to-end realization** evaluates complete methods on a stratified 100-instance subset under Scenarios A and B, measuring whether the outputs realize the intended emotion. Operator-specific ablations are reported in §D.4; metric definitions and aggregation rules are provided in §C.1.

Split integrity. Benchmark pools are screened before assembly. Retouch candidates use held-out image IDs after removing those present in executor training pools; EmoVerse candidates exclude planner-training record IDs. Checks over the 600-instance benchmark found no overlap with planner training data by record ID or filename stem, or with appearance-executor training/validation data and B_a by retouch image ID. The 100-instance end-to-end and 200-instance instruction-following sets are subsets of this benchmark.

B Implementation Details

B.1 Training Configuration and CAST

Capability-Aware Asymmetric Training (CAST). CAST adapts planning and generation sequentially using different parameter subsets and objectives. Planner adaptation uses textual supervision, while executor adaptation updates generation components for semantic and appearance editing while preserving the planner-adapted understanding components.

Parameter scope. Planner. We fully fine-tune the understanding tower, including the vision encoder path and language/output head. Generation-related MoT expert blocks and patch encoding/decoding layers are frozen with `requires_grad=False` before optimizer construction. Training uses cross-entropy only, with image-generation loss weight set to zero. **Executor.** Semantic and appearance editing share the same trainable scope: generation-side MoT expert blocks and patch encoding/decoding layers are fully fine-tuned, while the understanding tower and LLM remain frozen. Understanding loss is disabled and image-generation loss weight is 1.0; no separate semantic-reasoning adaptation is performed.

Training schedule. Table 6 summarizes the training runs. Stage 1 starts from the official backbone checkpoint. All Stage-2 variants start from the same planner-adapted checkpoint: semantic-only and appearance-only variants serve as ablations, while the final model jointly trains on both editing objectives in a single Mix SFT run. Appearance pairs are selected from a 23,980-pair training split by requiring a long edge of at least 768 pixels and a short edge of at least 512 pixels. The 9,796 semantic training pairs are constructed from UAIM-14K using FireRed-Edit (Super Intelligence Team et al., 2026): each source image I_0 and semantic instruction u_s are provided to FireRed-Edit to generate the target image $I^\star$, forming training tuples $(I_0, u_s, I^\star)$. No resolution filtering is applied to these pairs.

Table 6: Training schedules. All learning rates are constant.

Stage / variant	Training data	Schedule	Learning rate
Stage 1: Planner	14,907 planning instances	2 epochs / 112 steps	5×10^{-5}
Stage 2: Appearance-only	13,537 editing pairs	1 epoch / 160 steps	1×10^{-5}
Stage 2: Semantic-only	9,796 editing pairs	1 epoch / 116 steps	1×10^{-5}
Stage 2: Mix SFT	23,333 editing pairs	1 epoch / 276 steps	1×10^{-5}

Engineering configuration. We use the official SenseNova-U1.5-8B-MoT-Preview backbone on 8×H20 GPUs with weight parallelism WP=8, TP=1, PP=1, ZeRO-1 disabled, and gradient accumulation 1. The optimizer is AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.95$, $\epsilon = 10^{-8}$) with zero weight decay. Sequence lengths are 28,672 for Stage 1 and 23,000 for Stage 2; both stages use native dynamic image resolution (256–2048 pixels). We use seed 42 throughout and save model-only checkpoints each epoch, excluding optimizer and EMA states. Unlisted hyperparameters follow the official `sensenovau1_8b_mot_sft.py` configuration.

B.2 Inference and Baseline Configuration

EmoWeave inference. A single planner call produces (u_s, u_a), followed by one unified image-editing pass. The planner uses greedy decoding (`do_sample=False`, `max_new_tokens=2048`). The executor uses `num_steps=50`, `cfg_scale=4.0`, `img_cfg_scale=1.0`, `cfg_norm=none`, `timestep_shift=3.0`, and seed 42. CFG is applied throughout the denoising trajectory (`cfg_interval=(0.0, 1.0)`), and reasoning tokens are disabled during image generation (`think_mode=False`).

Planner baselines. We evaluate InternVL3.5-8B/38B, Qwen3-VL-8B/32B, SenseNova-U1.5-8B, InternVL-U, and BAGEL-7B zero-shot, without fine-tuning on UAIM-14K. All use official HuggingFace checkpoints run locally. SenseNova-U1.5-8B is the unadapted backbone and also serves as the CAST ablation reference.

Generation-side baselines. We likewise use official HuggingFace checkpoints run locally. General editing baselines include FireRed-Edit, Qwen-Image-Edit-2511, BAGEL, and InternVL-U; BAGEL and InternVL-U use their official default configurations. Specialized baselines include EmoEdit and EmoStyle (Scenario A only, following their original protocols), VeraRetouch, and Lance-Edit (appendix ablations only, §D.1). All baseline prompt and instruction templates will be released in our open-source repository.

Fixed-instruction execution. For joint execution, all editing models receive the identical (I_0, u_s, u_a) inputs, with both instructions generated by our planner. Operator-specific ablations instead provide (I_0, u_a) or (I_0, u_s) to evaluate appearance or semantic execution separately. Executors do not generate their own instructions in these comparisons, isolating execution capability from planning quality. These fixed-instruction settings are separate from end-to-end evaluation, where Scenario-B baselines receive neither the reference target emotion nor our planner outputs.

C Evaluation Protocol

C.1 Metric Definitions and Aggregation

Planning quality. We evaluate five aspects. *Target Emotion Accuracy* measures target recommendation in Scenario B, macro-averaged over the six emotions:

$$\mathrm{Acc} = \frac{1}{6} \sum_{k \in \mathcal{E}} \frac{1}{|\mathcal{D}_{B,k}|} \sum_{i \in \mathcal{D}_{B,k}} \mathbb{1}[\hat{e}_i^\star = e_i^\star],$$

where $\mathcal{D}_{B,k}$ contains Scenario-B samples with reference target emotion k. *State Grounding* assesses alignment with the source image and emotional objective; *Executability* assesses whether an editor can directly follow the instructions; *Instruction Coherence* assesses consistency between semantic

and appearance instructions. These dimensions are rated on a 1–5 scale and divided by 5. *Overall Quality* is a holistic planning score assigned directly by the VLM judges and reported separately from the individual rubric dimensions. Planning ratings are obtained from Gemini-3.1-Pro and GPT-5.5.

End-to-end realization. Results follow the scenario and task breakdowns in the main text. We use EEmo-Logic (Gao et al., 2026) as the visual emotion recognizer (VER) over seven labels: anger, disgust, fear, joy, neutral, sadness, and surprise. For a reported subset of N samples,

$$\text{EmoACC} = \frac{1}{N}\sum_i \mathbb{1}[\text{VER}(I_i^\star) = e_i^\star],$$

$$\text{Emo-S} = \frac{1}{N}\sum_i \left[P_{\text{VER}}(e_i^\star \mid I_i^\star) - P_{\text{VER}}(e_i^\star \mid I_{0,i})\right].$$

EmoACC measures target-emotion recognition accuracy. Emo-S measures the change in target-emotion probability, ranging from -1 to 1, with positive values indicating improvement. **Semantic** and **Appearance** are VLM ratings (1–10) of how well the respective edits serve the target emotion; inapplicable dimensions are marked "–". **Aesthetic** and **Quality** are UniPercept no-reference scores (0–100), reported separately.

Fixed-instruction execution. With editing instructions fixed (§D.4), a VLM judge evaluates instruction following (M1) across four dimensions: modification executed, requirement completeness, no contradiction, and instruction consistency. An independent holistic Overall score is reported alongside these dimensions. Content preservation is measured by CLIP-I and DINO-I embedding similarities between I_0 and $I^\star$.

Aggregation. Planning results are macro-averaged over the six reference target emotions within each scenario; pooled micro averages are supplementary. End-to-end metrics use arithmetic means within each reported subset. For metrics evaluated by both judges, their ratings are first averaged per sample before aggregation. Additional breakdowns by emotion, task type, and source are provided in §D.1.

Statistical testing. All comparisons use samples available for both methods. Planner and end-to-end comparisons use paired bootstrap resampling with 10,000 replicates (seed 42), yielding pointwise 95% percentile confidence intervals and approximate two-sided p-values from bootstrap tail probabilities. Planner statistics are macro-averaged over six benchmark target-emotion classes; end-to-end resampling groups records by source image. EGRA overall-quality comparisons use two-sided Wilcoxon signed-rank tests, excluding zero differences. Holm correction is applied separately within the planner, end-to-end, and EGRA comparison families, with significance defined by adjusted $p < 0.05$. Results appear in §D.1.2, with EGRA comparisons in §D.1.1.

C.2 Judge Rubrics and Prompts

For EGRA supervision validation, Gemini-3.1-Pro and GPT-5.5 independently rate semantic coherence, emotion consistency, instruction coherence, and domain validity on a 1–5 scale. Dimension-wise scores and comprehensive pass rates are reported separately for each judge. Overall quality is rated holistically on a separate 1–10 scale. Method identities are hidden, and candidate order is shuffled using deterministic sample-specific permutations.

Full prompts for EGRA validation, planning, end-to-end, and instruction-following evaluation will be released with the evaluation code (§F).

D SUPPLEMENTARY RESULTS

D.1 ADDITIONAL QUANTITATIVE RESULTS

D.1.1 EGRA SUPERVISION VALIDATION

Figure 11 compares QwenVL (Qwen3-VL-235B-FP8) and DeepSeek (DeepSeek-V4-Pro), with and without EGRA, evaluated independently by Gemini-3.1-Pro and GPT-5.5. DeepSeek receives textual states and target emotions, with retrieval evidence added by EGRA. QwenVL+EGRA receives the image, textual states, target emotion, and retrieved evidence; its retrieval-free baseline receives only the image and target emotion. Although absolute AND pass rates differ substantially between judges, DeepSeek+EGRA ranks first under both, consistent with the main-text win-rate results. Relative to DeepSeek without retrieval, EGRA improves all four rubric scores under both judges and ranks first in emotion consistency and instruction coherence. QwenVL without retrieval leads in semantic coherence, while QwenVL+EGRA leads in domain validity, indicating that strengths in individual dimensions do not necessarily yield the highest joint pass rate. These results support our choice of DeepSeek+EGRA for constructing emotion-grounded planning supervision in UAIM-14K.

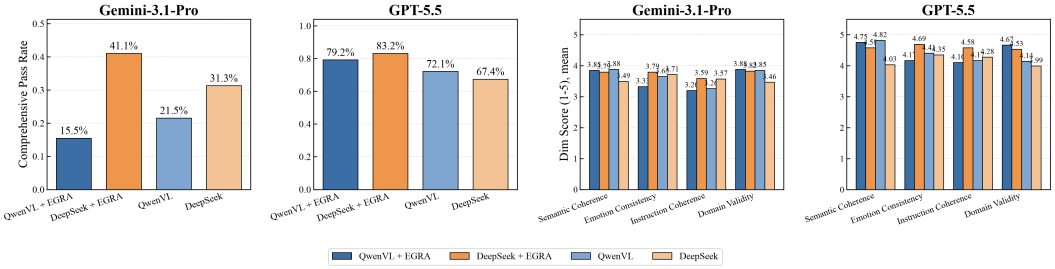

Figure 11: Quality of planning supervision generated by QwenVL and DeepSeek, with and without EGRA, assessed independently by Gemini-3.1-Pro and GPT-5.5. (a) Comprehensive pass rate under the AND criterion. (b) Scores for semantic coherence, emotion consistency, instruction coherence, and domain validity.

D.1.2 STATISTICAL SIGNIFICANCE

Table 7: Holm-adjusted significance of EmoWeave versus each baseline. +/− indicate significantly higher/lower scores (adjusted $p < 0.05$); n.s. indicates a nonsignificant difference. Ground.: State Grounding; Exec.: Executability; Coher.: Instruction Coherence; Acc.: target recommendation accuracy; Sem./App.: Semantic/Appearance.

(a) Planner

	Scenario A			Scenario B			
Baseline	Exec.	Coher.	Overall	Acc.	Exec.	Coher.	Overall
SenseNova-U1.5-8B	+	+	+	+	+	−	+
InternVL3.5-8B	+	+	+	+	+	+	+
Qwen3-VL-8B	+	+	+	+	+	n.s.	+
Qwen3-VL-32B	+	+	+	+	+	n.s.	+
InternVL-U	+	+	+	+	+	+	+
BAGEL-7B	+	+	+	+	+	+	+

(b) End-to-end

	A: Enhancement				A: Migration				Scenario B			
Baseline	EmoACC	Emo-S	Sem.	App.	EmoACC	Emo-S	Sem.	App.	EmoACC	Emo-S	Sem.	App.
Qwen-Image-Edit	n.s.	+	n.s.	+	n.s.	+	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.
FireRed-Edit	n.s.	n.s.	+	+	n.s.	n.s.	n.s.	n.s.	n.s.	+	n.s.	n.s.
BAGEL	n.s.	+	+	+	n.s.	n.s.	n.s.	+	n.s.	n.s.	n.s.	+
InternVL-U	n.s.	+	+	+	n.s.	n.s.	n.s.	+	n.s.	n.s.	+	+

Planner results. Table 7 shows significant gains in Overall and Executability under both scenarios, and in Scenario B recommendation accuracy, over all six tested baselines (adjusted $p = 0.015$). Against SenseNova-U1.5-8B, Scenario B gains are 0.0696 in Overall (95% CI: $[0.0553, 0.0849]$) and 0.2551 in recommendation accuracy ($[0.1931, 0.3161]$). Scenario B instruction coherence is lower than the backbone (adjusted $p = 0.015$).

End-to-end results. Significant gains depend on the metric and subset. EmoWeave improves Emo-S over Qwen-Image-Edit on both Scenario A tasks, and over FireRed-Edit in Scenario B. Against FireRed-Edit, enhancement gains also hold for Semantic and Appearance. Appearance gains over BAGEL and InternVL-U are significant across all three subsets, whereas EmoACC differences against these four baselines do not reach significance. These results support gains in emotion-probability improvement and editing appropriateness in specific settings, while EmoACC differences against the four general-purpose baselines remain statistically inconclusive.

D.1.3 Per-Emotion Stratified Results

Table 8: Per-emotion results on a shared diagnostic subset. All Scenario A samples are retained; Scenario B samples are retained only when the EmoWeave planner's target recommendation matches the reference. The same sample mask is applied to all methods, retaining 89 of 100 samples. Each cell reports EmoACC / Emo-S (N); [+X] denotes excluded samples.

Method	Task	Anger	Disgust	Fear	Sadness	Surprise	Joy
Qwen-Image-Edit	Enh.	0.33/0.07 (9)[+1]	0.73/-0.03 (11)[+2]	0.90/0.12 (10)[+2]	1.00/0.01 (14)[+2]	0.30/-0.02 (10)[+1]	1.00/-0.04 (7)[+2]
	Mig.	0.50/0.22 (4)	0.00/0.01 (3)	0.00/0.00 (2)[+1]	0.67/0.37 (6)	0.00/0.02 (4)	0.33/0.14 (9)
VeraRetouch	Enh.	0.11/-0.01 (9)[+1]	0.82/-0.02 (11)[+2]	0.70/0.02 (10)[+2]	1.00/0.01 (14)[+2]	0.40/-0.01 (10)[+1]	1.00/0.00 (7)[+2]
	Mig.	0.00/0.01 (4)	0.00/0.01 (3)	0.00/0.12 (2)[+1]	0.67/0.37 (6)	0.00/0.03 (4)	0.33/0.02 (9)
EmoWeave	Enh.	0.56/0.34 (9)[+1]	0.91/0.04 (11)[+2]	1.00/0.20 (10)[+2]	0.93/-0.04 (14)[+2]	0.40/0.03 (10)[+1]	1.00/0.04 (7)[+2]
	Mig.	0.75/0.47 (4)	0.00/0.03 (3)	0.00/0.09 (2)[+1]	0.83/0.68 (6)	0.00/0.01 (4)	0.78/0.43 (9)

Table 8 provides a conditional diagnostic breakdown, supplementing the full-sample end-to-end results in §4.2. On enhancement, EmoWeave achieves the best or tied-best EmoACC in five of six emotions. On migration, it leads on anger, sadness, and joy, while all three methods obtain zero EmoACC on disgust, fear, and surprise. These cells contain only 3, 2, and 4 samples, respectively: they reveal unsuccessful target realization in the evaluated cases but provide limited evidence for broader per-emotion comparisons.

D.1.4 Preservation and Treatment Effectiveness by Task Type

Table 9: Content preservation and emotion-directed treatment effectiveness by task type. CLIP-I/DINO-I measure similarity between I_0 and $I^\star$; Sem/Ret are VLM-rated semantic/appearance treatment scores (1–10). EmoEdit and EmoStyle are evaluated on Scenario A only (§4.1).

	Enhancement					Migration				
Method	CLIP-I	DINO-I	Sem	Ret	N	CLIP-I	DINO-I	Sem	Ret	N
VeraRetouch	**0.97**	**0.99**	5.11	5.44	71	**0.93**	**0.98**	4.12	5.31	29
Qwen-Image-Edit	0.93	0.96	5.94	6.23	71	0.91	0.96	5.76	6.07	29
EmoWeave	0.86	0.89	**6.59**	**6.79**	71	0.75	0.75	**6.21**	**6.50**	29
EmoEdit	0.80	0.77	4.93	5.47	30	0.80	0.72	4.42	5.19	26
EmoStyle	0.57	0.57	4.03	3.57	30	0.56	0.51	3.83	3.75	26

Among the three methods evaluated on the same samples in Table 9, EmoWeave achieves the highest Sem/Ret scores for both enhancement and migration, while VeraRetouch and Qwen-Image-Edit retain higher CLIP-I/DINO-I similarity. The preservation gap widens on migration, while EmoWeave maintains its lead in treatment effectiveness. These results illustrate a preservation–effectiveness tradeoff within this comparison.

Table 10: End-to-end results by source. AetherRetouch targets are estimated by VER on edited images; EmoVerse source labels follow the Mikels-to-Ekman mapping, with targets assigned during benchmark construction. EmoEdit and EmoStyle are evaluated on Scenario A only (§4.1).

	AetherRetouch						EmoVerse					
Method	CLIP-I	DINO-I	Sem	Ret	EmoACC	N	CLIP-I	DINO-I	Sem	Ret	EmoACC	N
VeraRetouch	**0.95**	**0.99**	4.89	5.77	0.73	33	**0.96**	**0.99**	4.78	5.22	0.43	67
Qwen-Image-Edit	0.93	0.96	5.53	6.00	0.67	33	0.92	0.96	6.07	6.28	0.49	67
EmoWeave	0.82	0.83	6.14	6.26	**0.76**	33	0.83	0.86	**6.64**	**6.93**	**0.61**	67
EmoEdit	0.78	0.68	4.93	5.38	0.62	21	0.81	0.78	4.56	5.31	0.37	35
EmoStyle	0.58	0.51	4.19	3.74	0.52	21	0.56	0.56	3.79	3.60	0.26	35

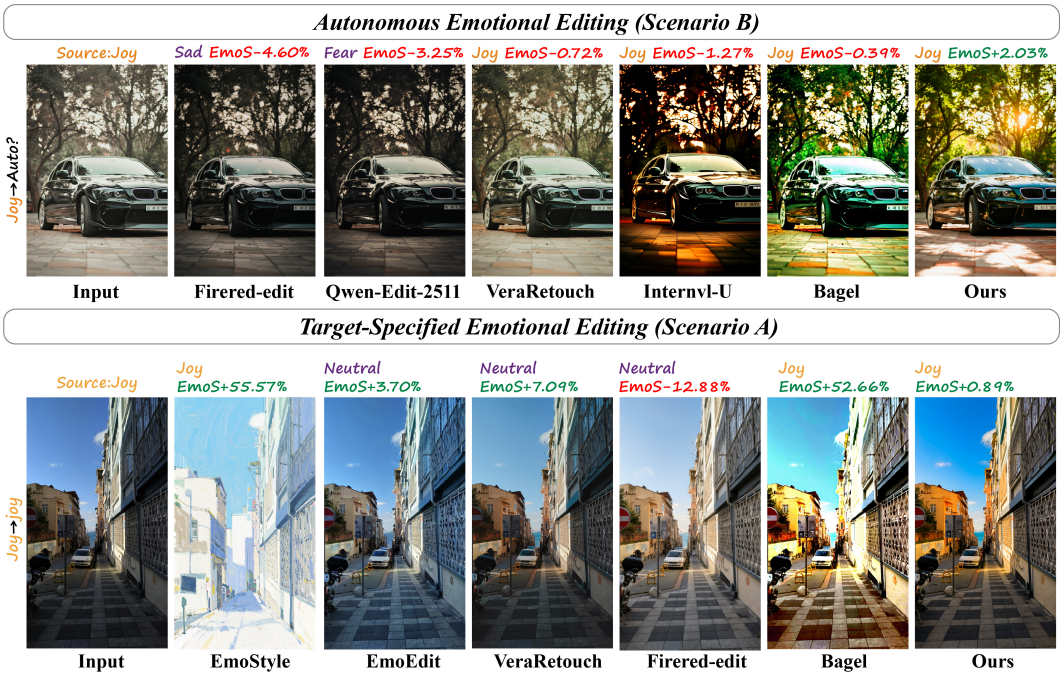

Figure 12: Additional end-to-end comparisons in Scenario B (top) and Scenario A (bottom), with predicted emotions and Emo-S annotated above each output.

D.1.5 AetherRetouch vs. EmoVerse Source Comparison

Table 10 shows lower EmoACC on EmoVerse than on AetherRetouch for every method. Among the three methods evaluated on the same samples, EmoWeave achieves the highest EmoACC and Sem/Ret scores on both sources, with the smallest decline in EmoACC across sources. Their preservation rankings remain unchanged, with VeraRetouch achieving the highest CLIP-I/DINO-I scores in both subsets. These results show consistent relative performance across the two source subsets.

D.1.6 Cross-VER Robustness Check

To verify that our end-to-end results are not an artifact of a single emotion recognizer, we re-score the enhancement-subset outputs from Table 2(a) using EmoCaliber (Wu et al., 2025b) in place of EEmo-Logic, querying it under the same closed-set, forced-choice protocol over our 7-label (Ekman6+neutral) taxonomy to align with EEmo-Logic's output space. Table 11 reports EmoACC and Emo-S under this protocol, with the change relative to the corresponding EEmo-Logic score in Table 2(a) shown in parentheses.

EmoWeave remains the top-ranked method under EmoCaliber on both metrics, and the relative ordering is preserved for all methods except a minor swap between InternVL-U and FireRed-Edit on EmoACC, whose EEmo-Logic scores were already close (0.6667 vs. 0.7333). All EmoACC deltas fall within ± 0.04 and all Emo-S deltas within ± 0.018, with no systematic direction favoring either

Table 11: EmoACC and Emo-S under EmoCaliber (enhancement subset), with the difference from the EEmo-Logic scores in Table 2(a) shown in parentheses.

Method	EmoACC (EmoCaliber)	Emo-S (EmoCaliber)
EmoEdit	0.57 (−0.03)	0.0089 (−0.0042)
EmoStyle	0.40 (−0.03)	−0.1785 (+0.0175)
VeraRetouch	0.63 (−0.04)	0.0225 (+0.0032)
FireRed-Edit	0.70 (−0.03)	0.0812 (−0.0085)
Qwen-Image-Edit-2511	0.77 (+0.04)	0.0751 (+0.0114)
BAGEL	0.60 (−0.03)	0.0203 (+0.0034)
InternVL-U	0.70 (+0.03)	−0.0089 (−0.0043)
EmoWeave	**0.80** (−0.03)	**0.1428** (−0.0063)

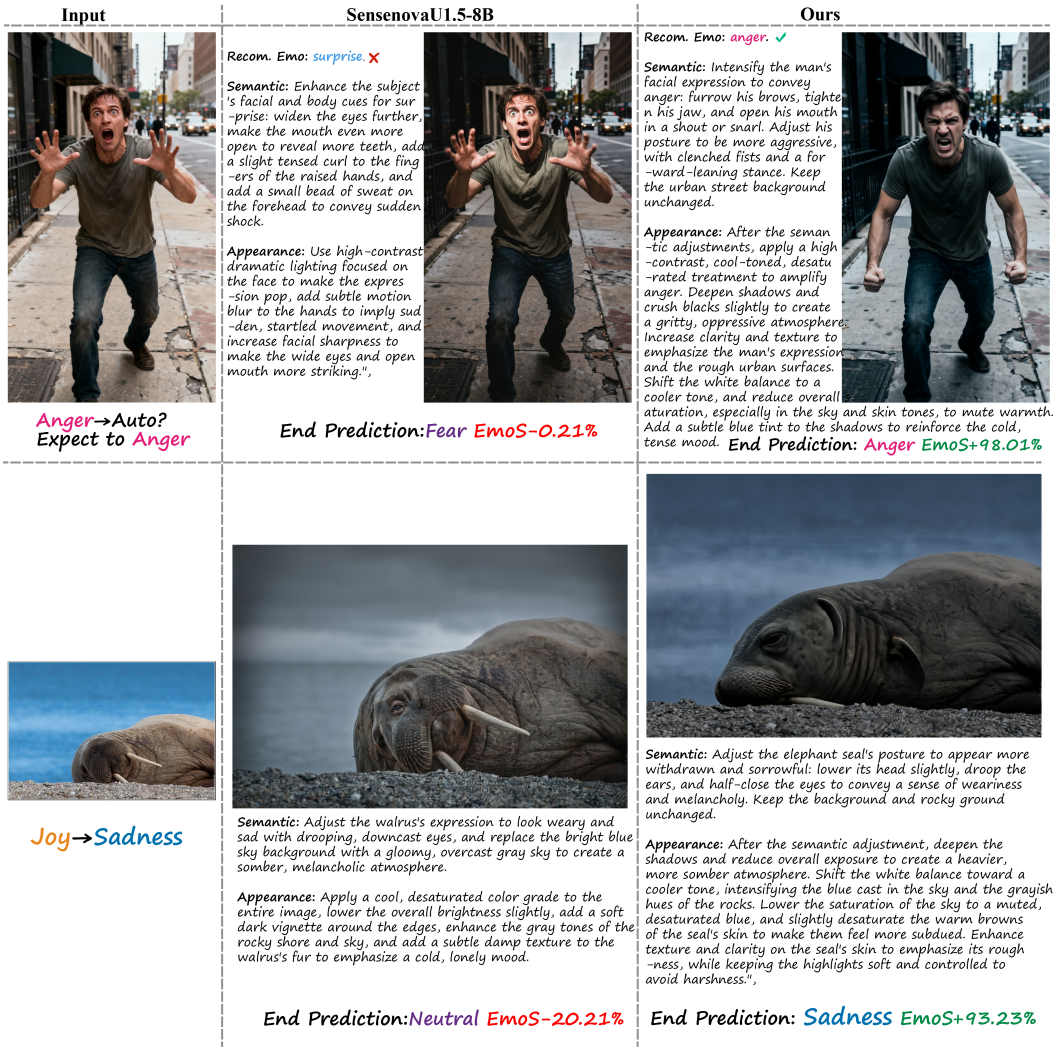

Figure 13: Visualization of EmoWeave planning outputs, including predicted target emotions and generated semantic/appearance editing instructions.

recognizer, indicating the end-to-end results are not dependent on a single VER model's labeling behavior.

D.2 Additional Qualitative Results

Planning visualization. Figure 13 visualizes the intermediate planning results generated by EmoWeave. Given different emotional objectives, the model produces coordinated semantic and appearance instructions grounded in the input content. The examples show that the generated plans identify both content-level changes and appearance-level adjustments required for achieving the target emotional expression.

Ablation visualization. Figure 14 visualizes the contribution of planner and executor adaptation. Removing either component leads to incomplete emotional transformations, while the full EmoWeave model better aligns semantic changes and appearance adjustments with the target emotion.

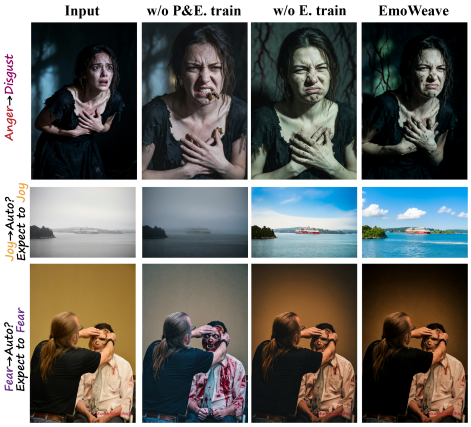

Figure 14: Qualitative ablation of planner and executor training in EmoWeave.

D.3 Executor Evaluation

We compare executors under identical (I_0, u_s, u_a) inputs, with both instructions generated by our planner and fixed across methods. This controlled setting isolates execution capability from planning quality. As shown in Table 12, EmoWeave achieves competitive instruction following across both semantic and appearance editing, while substantially improving over the unadapted SenseNova-U1.5-8B backbone on both dimensions. This confirms that the executor adaptation improves the model's ability to realize fixed editing instructions, rather than merely benefiting from stronger planning. Qwen-Image-Edit remains stronger in semantic following and content preservation, indicating a different trade-off between instruction realization and source-image preservation. Separate appearance-only and semantic-only analyses are provided in §D.4.1 and §D.4.2.

Table 12: Executor comparison using identical semantic and appearance instructions generated by our planner. Semantic and Appearance measure instruction following for semantic and appearance edits, respectively (higher is better). CLIP-I/DINO-I measure source-image similarity; higher similarity indicates greater resemblance to the source, but does not necessarily imply better instruction following. Bold and underlined values denote the highest and second-highest scores in each column, respectively.

Executor	Instruction Following		Source Similarity	
	Semantic ↑	Appearance ↑	CLIP-I	DINO-I
FireRed-Edit	0.6800	<u>0.7300</u>	0.8871	0.8499
Qwen-Image-Edit	**0.7163**	0.7175	**0.9034**	**0.8799**
BAGEL	0.6050	0.6175	0.8925	0.8629
InternVL-U	0.6148	0.5714	0.8798	0.8357
SenseNova-U1.5-8B	0.7088	0.7132	0.8914	<u>0.8710</u>
EmoWeave	<u>0.7150</u>	**0.7325**	<u>0.8978</u>	0.8415

D.4 Operator-specific Capability Analysis

We evaluate appearance-only (I_0, u_a) and semantic-only (I_0, u_s) execution using fixed planner-generated instructions, complementing joint execution in Table 12. All executors share the same sample subset: all Scenario A samples and Scenario B samples with correct EmoWeave target recommendations ($c_i = 1$).

We report requirement completeness and instruction consistency (1–5), omitting two near-ceiling dimensions: modification executed and no contradiction. Overall is an independent holistic VLM score (0–1); CLIP-I/DINO-I measure input–output similarity. Bold and underlined values denote column maxima and runners-up.

D.4.1 APPEARANCE OPERATOR (INSTRUCTION-ONLY EXECUTION)

Table 13 includes Lance-Edit and VeraRetouch as appearance-specialized baselines. EmoWeave leads on both instruction-following dimensions and Overall, while VeraRetouch preserves content most closely.

Table 13: Appearance-only execution with fixed (I_0, u_a) inputs. Higher is better for all metrics.

Executor	Requirement Completeness ↑	Instruction Consistency ↑	CLIP-I ↑	DINO-I ↑	Overall ↑
SenseNova-U1.5-8B	4.1361	<u>4.9409</u>	<u>0.9124</u>	<u>0.9173</u>	0.9422
InternVL-U	<u>4.7462</u>	4.8782	0.8755	0.8421	<u>0.9732</u>
BAGEL	3.5899	4.0492	0.9061	0.9070	0.8468
Lance-Edit	3.9578	4.6933	0.8665	0.8397	0.9154
VeraRetouch	4.3654	4.8563	**0.9193**	**0.9323**	0.9632
EmoWeave	**4.9181**	**4.9525**	0.9017	0.9025	**0.9887**

D.4.2 SEMANTIC OPERATOR (INSTRUCTION-ONLY EXECUTION)

In Table 14, EmoWeave leads on both instruction-following dimensions and Overall, followed by SenseNova-U1.5-8B. EmoEdit has the highest input–output similarity and EmoWeave the lowest; similarity alone cannot distinguish intended semantic changes from unintended content loss.

Table 14: Semantic-only execution with fixed (I_0, u_s) inputs. CLIP-I/DINO-I are reference similarities, as semantic edits may require content changes; higher is better for other metrics.

Executor	Requirement Completeness ↑	Instruction Consistency ↑	CLIP-I	DINO-I	Overall ↑
SenseNova-U1.5-8B	<u>4.6538</u>	<u>4.7331</u>	0.8432	0.8394	<u>0.9132</u>
InternVL-U	4.1035	4.6218	0.8677	0.8455	0.9046
BAGEL	3.4491	4.0863	<u>0.8791</u>	<u>0.8602</u>	0.8123
EmoEdit	4.0234	4.1094	**0.8902**	**0.8971**	0.9103
EmoWeave	**4.8932**	**4.8804**	0.8340	0.8247	**0.9402**

E LIMITATIONS

EmoWeave is evaluated only on SenseNova-U1.5-8B-MoT, and its transferability to other UMM architectures or scales remains untested. The current formulation is limited to post-capture editing and does not address capture-time decisions such as composition, framing, or lighting. It also models emotional expression at the single-image level, leaving temporal or narrative settings such as photo sequences and video unexplored. User-specific emotional preferences are not considered. Finally, edits may misrepresent emotional intent or be used misleadingly, especially for images of people, requiring attention to consent, privacy, and data licensing.

F REPRODUCIBILITY STATEMENT

We will release, upon publication, the full training and evaluation codebase, UAIM-14K and UAIM-Bench in their entirety, the EGRA construction pipeline (retrieval database construction, instruction generation, and verification scripts, §A.2), all training configurations for the planner and executor stages (§B.1), inference and decoding configurations for EmoWeave and every baseline (§B.2), and the complete evaluation suite including all judge prompts and rubrics (§C.2). Every quantitative result in this paper, including the appendix tables in §D.1, is reproducible from these released artifacts using the exact checkpoints, seeds, and decoding parameters documented in §B.1 and §B.2.

AI USE STATEMENT

AI tools assisted with manuscript editing and code development. Pretrained models were also used for data construction, editing-pair generation, and automated evaluation, as detailed in the corresponding methodology and appendix sections. The authors take responsibility for the final manuscript, experimental analyses, and reported results.