AFFECTRAG: CONTINUOUS AFFECTIVE IMAGE EDITING VIA RETRIEVAL-AUGMENTED SEMANTIC GROUNDING AND RESIDUAL CONTROL

Anonymous authors
Paper under double-blind review

ABSTRACT

Affective image editing (AIE) focuses on modulating the emotions evoked by images, yet existing methods typically rely on discrete emotion categories and fail to capture fine-grained affective variations. In this paper, we study **continuous affective image editing (CAIE)**, which aims to edit an image toward a target valence-arousal (V-A) state. Current methods face two key challenges in CAIE. First, V-A values lack explicit visual semantics, making the numeric-to-visual mapping difficult to learn. Second, existing editing mechanisms cannot precisely reach a specified V-A coordinate. To address these challenges, we propose **AffectRAG**, a retrieval-augmented planning framework that combines discrete semantic grounding with residual-conditioned directional control. Specifically, to compensate for the lack of visual semantics, we construct a large-scale Affective Knowledge Base to ground abstract V-A targets into editable semantic anchors. To further improve precision, we introduce **Retrieval-Augmented Residual Control Optimization** to learn the relationship between anchor-target residuals and ternary visual controls with plan-level rewards and retrieval-confidence gating. Experiments on multiple public datasets show that AffectRAG outperforms existing state-of-the-art AIE and general editing methods in V-A accuracy and semantic preservation.

1 INTRODUCTION

Emotion shapes how humans perceive, understand, and express the world (Pessoa, 2008). As a fundamental medium of visual communication, images can evoke subjective emotional responses from viewers. Therefore, the ability to modulate the affective impression of images is valuable for content creation, personalized generation and affective regulation (Yang et al., 2024). In recent years, affective image editing has emerged as an important direction in image editing, aiming to alter the emotion conveyed by an image while preserving its main content and visual naturalness.

Existing affective image editing methods mostly formulate the target emotion as a discrete category (Yang & Peng, 2008; Liu et al., 2018; Weng et al., 2023; Yang et al., 2025; Lin et al., 2025). Some methods divide emotions into positive and negative classes, while others use a small set of emotion adjectives. Such formulations deviate from the continuous nature of real human emotions. Human emotions cannot be adequately described by a few fixed categories. Instead, they vary continuously in positivity and intensity (Cowen & Keltner, 2017). For example, the transition from contentment to amusement is gradual rather than separated by clear categorical boundaries.

To address this limitation, we study **continuous affective image editing (CAIE)**. We adopt the valence-arousal (V-A) model (Russell, 1980), which represents emotion in a continuous Cartesian space. Valence measures the transition from negative to positive, while arousal measures the transition from low activation to high activation. Compared with finite discrete emotion categories, the continuous V-A space provides a more precise representation of emotional intensity and transition, offering a fine-grained control target for editing (Zhao et al., 2021; Paskaleva et al., 2024).

However, CAIE introduces new challenges. First, continuous V-A coordinates lack direct visual semantics, making it difficult for existing models to establish a stable mapping between affective numerical values and editable visual features. For a numerical target such as "$V = 7$, $A = 6$", they struggle to determine which visual cues should be introduced or adjusted to express the affective po-

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.

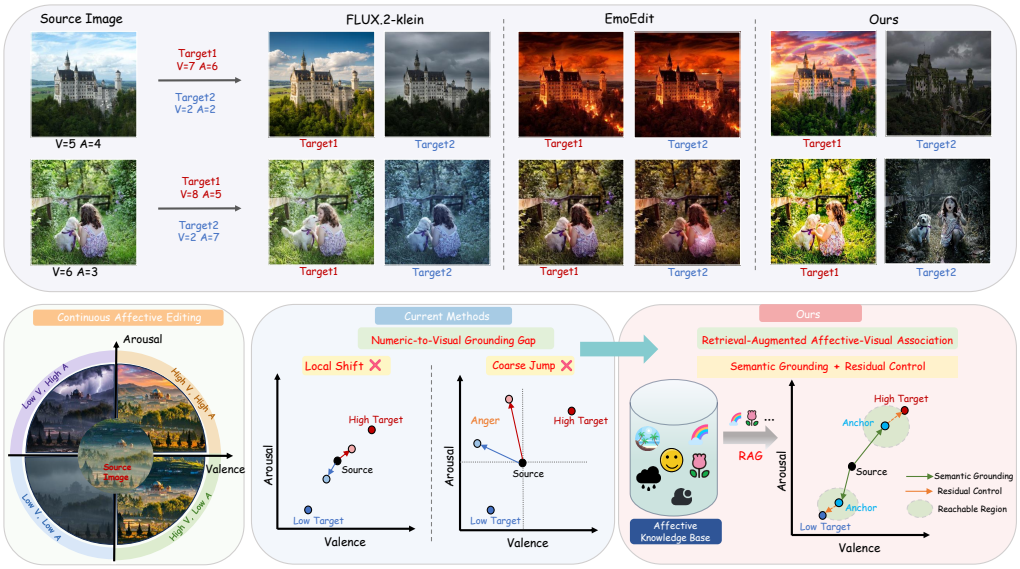

Figure 1: Overview of **AffectRAG** and existing methods for continuous affective image editing. Existing methods struggle to establish numeric-to-visual mapping, and their local or coarse editing cannot precisely reach the target. In contrast, AffectRAG retrieves affective-visual associations from Knowledge Base and combines semantic grounding with residual control, expanding each semantic anchor into a locally reachable region for more effective coverage of the continuous affective space.

sition. Second, existing editing mechanisms are not suitable for precise localization in a continuous affective space. They induce only local emotional shifts with lighting-and-color adjustments (Yang & Peng, 2008; Liu et al., 2018; Weng et al., 2023) or coarse affective jumps by element-level editing (Yang et al., 2025; Lin et al., 2025), and neither can reliably reach a specified V-A coordinate.

Our key insight is that CAIE should be achieved through the combination of **discrete semantic grounding** and **residual-conditioned directional control**. Discrete affective elements, such as flowers, beaches, and smiling faces, provide explicit emotional semantics and can serve as anchors for large-scale movement in the V-A space. Visual controls of lighting and color can then locally adjust the image around a given semantic anchor to further approach the target emotional position. Figure 1 illustrates the limitations of existing methods for CAIE and the key idea of our method.

Based on this idea, we propose **AffectRAG**, a retrieval-augmented planning framework for CAIE. AffectRAG employs Qwen3-VL-4B (Bai et al., 2025) as the Planner to generate editing instructions, and FLUX.2-klein (Black Forest Labs, 2025) as the Executor to perform visual editing. To address the lack of visual semantics in continuous affective values, we construct a large-scale Affective Knowledge Base, organizing affective-visual associations in a tree structure of **factors-attributes-coordinates**. Through retrieval, the Planner can ground an abstract coordinate into editable visual cues, obtaining more reliable semantic anchors. To further improve the precision of CAIE, we introduce **Retrieval-Augmented Residual Control Optimization (RACO)**. RACO treats the difference between the target V-A coordinate and the anchor V-A coordinate as an affective residual, and trains the Planner to map this residual into visual-control directions. Meanwhile, RACO introduces retrieval-confidence gating to adaptively regulate the coverage of retrieved semantics in instruction, balancing semantic grounding and residual control. The contributions of this paper are as follows:

1) We introduce the task of CAIE, extending affective image editing from discrete categories to continuous V–A space, and propose **AffectRAG**, a retrieval-augmented framework that decouples CAIE into coarse semantic anchoring and fine-grained visual refinement.

2) We construct an **Affective Knowledge Base** to compensate for the lack of visual semantics in affective numerical values. It organizes affective-visual associations into a tree structure, enabling the Planner to retrieve editable semantic cues from abstract V–A targets.

3) We propose **RACO** to improve emotional precision by learning the relationship between affective residuals and ternary visual controls via plan-level rewards and retrieval-confidence gating. Experiments show that AffectRAG outperforms SOTA methods in V–A accuracy and content preservation.

2 Related Work

Visual Emotion Analysis. Visual emotion analysis studies affective responses evoked by visual content. Early methods relied on low-level cues such as color, texture, and composition (Machajdik & Hanbury, 2010; Borth et al., 2013), while recent approaches emphasize semantic and contextual modeling (Rao et al., 2020; Yang et al., 2021). Based on the valence-arousal (V-A) model (Russell, 1980), datasets such as EMOTIC (Kosti et al., 2019) and FindingEmo (Mertens et al., 2024) support continuous affect recognition, and recent models further extend V-A representations to multimodal understanding and generation (Lian et al., 2025; Dang et al., 2025). However, these works mainly focus on affect perception or synthesis, rather than editing a given image toward a target V-A coordinate while preserving its content, which is the focus of this paper.

Affective Image Editing. Diffusion-based image editing methods (Meng et al., 2021; Hertz et al., 2022; Brooks et al., 2023; Zhang et al., 2023) and large-scale editors such as FLUX and Qwen-Image-Edit (Batifol et al., 2025; Black Forest Labs, 2025; Wu et al., 2025) have achieved strong instruction-following capabilities. Yet they rely on natural-language prompts and lack mechanisms for translating numerical affective targets into visual operations. Existing affective editing methods explore color/style transfer (Yang & Peng, 2008; Liu et al., 2018; Weng et al., 2023) or semantic element manipulation (Yang et al., 2025; Lin et al., 2025), but typically define targets as discrete emotion categories. In contrast, continuous affective image editing requires both semantic grounding of abstract V-A values and fine-grained visual calibration toward target coordinates.

Retrieval-Augmented Affective Planning. LLM planning methods, such as Chain-of-Thought (Wei et al., 2022), ReAct (Yao et al., 2022), and prompt optimization (Zhou et al., 2022), decompose complex goals into executable steps. Retrieval-augmented generation further enhances planning with external knowledge (Lewis et al., 2020; Borgeaud et al., 2022). For continuous affective editing, raw V-A coordinates do not directly specify editable visual elements or control directions. Therefore, we construct an affective knowledge base that grounds abstract V-A targets into semantic editing anchors, enabling retrieval-guided planning for precise affective image editing.

3 Method

3.1 Problem Formulation

We formulate CAIE in the V-A space, where the affective target is represented as a two-dimensional continuous coordinate. In this work, both valence and arousal are normalized to the range of $[0, 9]$. Given a source image $\boldsymbol{I}_s$ and a target affective coordinate $\boldsymbol{z}^* = (v^*, a^*)$, where v^* and a^* denote the target valence and arousal, respectively, the goal of CAIE is to generate an edited image $\boldsymbol{I}_e$ that conveys the target affective state while preserving the main semantics of the source image.

We decompose the editing process into two stages: planning and execution. First, a vision-language model Planner π_θ generates an editing instruction τ conditioned on the source image, the target affective coordinate, and the retrieved affective-visual knowledge. Then, a training-free image editing model $\mathcal{E}$ acts as the Executor and generates the edited image. We use an affective evaluator Φ to assess whether the edited image reaches the target affective state. Ideally, the edited image should exhibit minimal deviation from the target in V–A space while preserving the source semantics.

3.2 Affective Knowledge Base

The **Affective Knowledge Base** (AKB) organizes affective-visual associations in a tree structure of **factors-attributes-coordinates**. Each affective factor represents a semantic component. Under each factor, we collect multiple concrete visual attributes from real images, each associated with a corresponding V-A coordinate. This structure allows the Planner to retrieve relevant visual cues from a target coordinate, rather than relying solely on its implicit understanding of affective values.

We build the Affective Knowledge Base based on the training splits of EmoSet dataset (Yang et al., 2023). The construction process consists of three stages, as illustrated in Figure 2.

Stage 1: Visual-Affective Annotation We first use Qwen3-VL-8B (Bai et al., 2025) to perform structured annotation on selected images from EmoSet. For each image, the model extracts affective

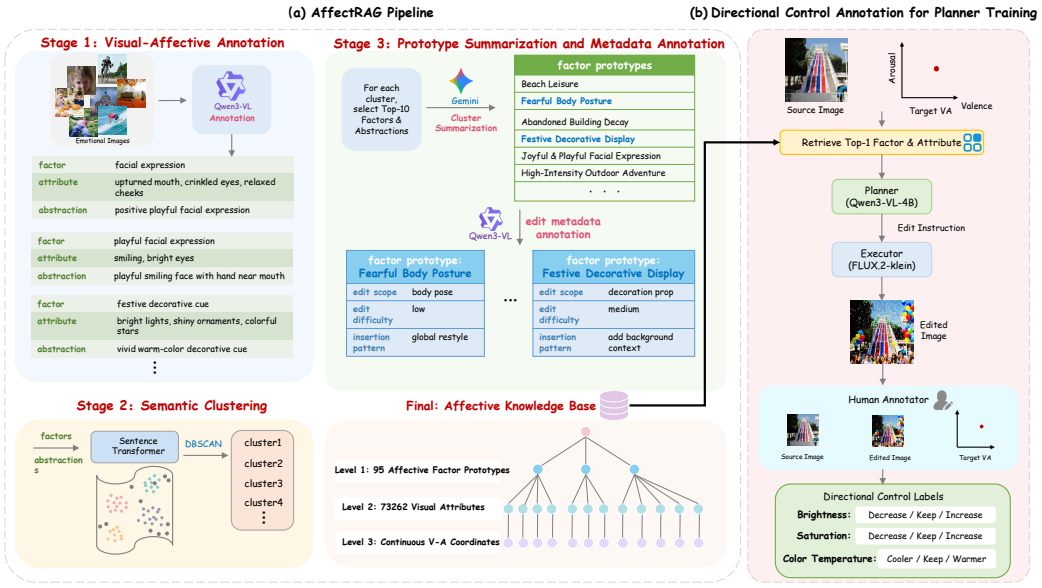

Figure 2: Overview of the AKB and dataset annotation. (a) Construction of the factor-attribute-coordinate tree structure in the AKB. (b) Generation of editing pairs and directional control labels.

factors, concrete visual attributes, and an image abstraction. Factors describe the main semantic elements that influence the image's emotional impression, while attributes describe how these factors appear in the image. Abstraction provides a holistic description of the overall affective context.

Stage 2: Semantic Clustering To derive stable affective factors from a large number of images, we use Sentence Transformer (Reimers & Gurevych, 2019) to encode each image's factors and abstractions into textual features. Then we apply DBSCAN (Ester et al., 1996) to group samples with similar semantics and affective roles into the same cluster, forming a set of candidate factors.

Stage 3: Prototype Summarization and Metadata Annotation For each cluster, we count the top-10 most frequent factors and abstractions, and employ Gemini3 Pro (Team et al., 2023) to summarize the cluster into a factor prototype name. We annotate each prototype with a set of editing metadata for compatibility filtering and retrieval ranking in later stages. The metadata includes:

Edit scope: where the factor mainly takes effect, such as body pose or local object.

Edit difficulty: the difficulty for an editing model to realize it, categorized as low, medium or high.

Insertion pattern: how to modify image content, such as global restyle or add background context.

To enable semantic coverage checking during training, we further generate synonym lists for each factor and attribute, allowing the model to use natural expressions instead of copying the original knowledge base entries verbatim. Both the editing metadata and synonym lists are first annotated by Qwen3-VL-8B and then manually corrected. The resulting AKB comprises 95 affective factors and 73,262 visual attributes. Each attribute is associated with a V–A coordinate derived from human judgments of its affective impression. The annotation protocol is detailed in Appendix C.

3.3 AffectRAG

Directly generating instructions from continuous V-A values is difficult. To address this, AffectRAG decomposes CAIE into two stages. The overall structure of AffectRAG is presented in Figure 3.

Discrete Semantic Grounding Given an image $\boldsymbol{I}_s$, the Planner first determines the set of compatible edit scopes $\mathcal{S}(\boldsymbol{I}_s)$ and insertion patterns $\mathcal{P}(\boldsymbol{I}_s)$. Each affective factor f_i in AKB is associated with an edit scope set $\mathcal{S}(f_i)$ and an insertion pattern set $\mathcal{P}(f_i)$. A factor is considered compatible with the source image if both its edit scope and insertion pattern are allowed by the current image:

$$\mathbb{I}(f_i, \boldsymbol{I}_s) = \mathbb{I}[\mathcal{S}(f_i) \subseteq \mathcal{S}(\boldsymbol{I}_s)] \cdot \mathbb{I}[\mathcal{P}(f_i) \subseteq \mathcal{P}(\boldsymbol{I}_s)]. \tag{1}$$

After compatibility filtering, we expand all visual attributes under compatible factors to form a candidate pool. Each candidate is ranked based on its V-A distance to the target and the editing

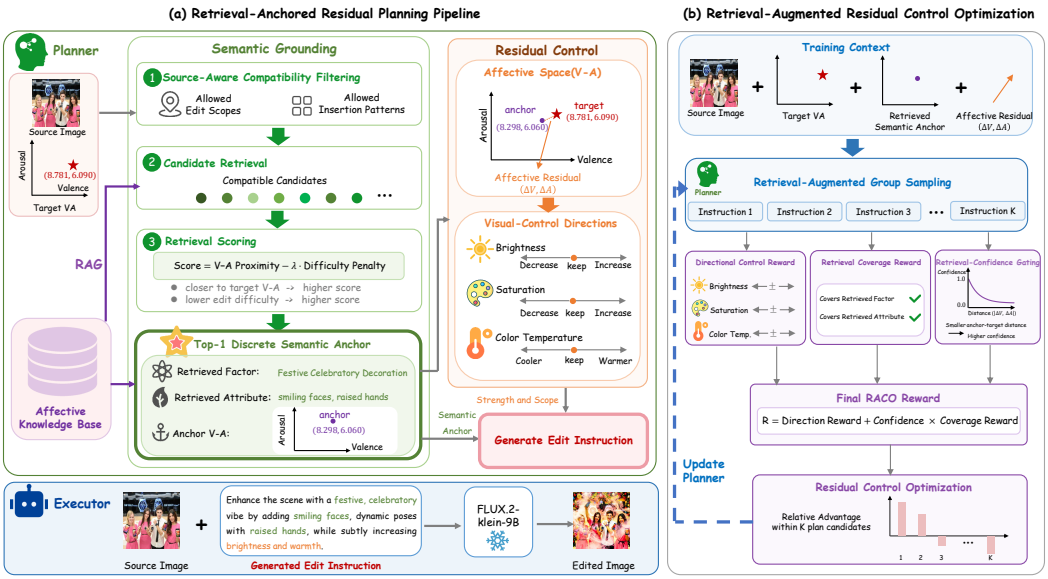

Figure 3: Overview of AffectRAG. (a) Pipeline: The planner retrieves a semantic anchor to ground V-A into editable cues, and predicts residual controls to further approach the target. (b) Optimization: RACO learns the relationship between residuals and ternary visual-control directions.

difficulty:

$$S_{ij} = -\left(|v_{ij} - v^*| + |a_{ij} - a^*|\right) - \lambda_d d_i, \tag{2}$$

where λ_d is the penalty coefficient, and d_i is the edit difficulty of factor f_i. Low, medium and high are mapped to 0, 0.5 and 1. The top-ranked candidate is selected as the retrieved semantic anchor.

Residual-Conditioned Directional Control Although the retrieved anchor guides the image emotion toward the target region, its V-A coordinate usually does not exactly match the target coordinate. We therefore define the affective residual as the coordinate-wise difference between the target affect $\boldsymbol{z}^*$ and the retrieved anchor $\boldsymbol{z}_r$. This residual describes the direction and magnitude by which the edited result still needs to move in the V-A space. Given the source image, target affect, retrieved anchor, and affective residual, the Planner predicts a ternary direction vector and generates the editing instruction. The residual magnitude is provided as contextual information to condition the strength and spatial scope of the generated instruction. Thus, targets sharing the same semantic anchor and direction vector can still yield different instructions. For example, an instruction may specify a slight or substantial increase in saturation, or restrict the adjustment to the sky rather than apply it across the image. Compared with editing directly from the target V-A, semantic retrieval narrows the planning problem from a mapping across the full V–A space to local refinement around an affectively grounded reference, reducing the difficulty of translating affective values to visual changes.

3.4 Retrieval-Augmented Residual Control Optimization

In the above pipeline, the Planner needs to learn two abilities: predicting correct visual-control directions from the residual, and properly utilizing the retrieved anchors to generate the editing instruction. To this end, we propose Retrieval-Augmented Residual Control Optimization, which optimizes the Planner with directional rewards and retrieval-confidence-gated coverage rewards.

Directional Control Annotation To train the Planner to learn the mapping from affective residuals to visual controls, we construct directional control labels for each training sample.

For image $\boldsymbol{I}_s$ and target $\boldsymbol{z}^*$, we first retrieve the top-1 semantic anchor from AKB using the strategy described in Sec. 3.3. We then generate an initial editing instruction and an edited image. Annotators are presented with the source image, the edited image, and the target coordinate. Through direct visual assessment, they indicate how the illumination, saturation, and color temperature of the edited image should be further adjusted to better approach the target emotion. Each dimension is labeled as one of three directions: -1 denotes decrease or cooler, 0 denotes keep, and $+1$ denotes increase or warmer. This yields a three-dimensional directional label $\boldsymbol{u}^{gt} = (u_l^{gt}, u_s^{gt}, u_t^{gt})$ used by RACO. For EmoSet, annotators additionally rate the source image's V-A during the same annotation session.

Table 1: Quantitative comparison on EmoSet, EMOTIC, and FindingEmo. Best scores are in blue, second-best in green. AffectRAG improves affective accuracy while preserving core semantics.

	EmoSet				EMOTIC				FindingEmo			
Method	V-Error ↓	A-Error ↓	CLIP-I ↑	LPIPS ↓	V-Error ↓	A-Error ↓	CLIP-I ↑	LPIPS ↓	V-Error ↓	A-Error ↓	CLIP-I ↑	LPIPS ↓
Seedream 5.0	1.491	1.826	0.879	0.457	2.129	1.706	0.872	0.376	1.697	1.759	0.843	0.437
Nano Banana Pro	1.632	1.950	0.888	0.413	2.169	1.755	0.876	0.405	2.321	1.893	0.850	0.410
Qwen-Image-Edit	1.830	1.888	0.858	0.649	2.433	1.863	0.894	0.594	2.705	2.054	0.804	0.825
FLUX.1-Kontext	1.868	1.948	0.870	0.669	2.446	1.867	0.873	0.672	2.774	2.058	0.809	0.673
FLUX.2-dev	1.667	1.927	0.872	0.663	2.400	1.865	0.898	0.517	2.735	2.032	0.800	0.666
EmoEdit	1.675	1.929	0.779	0.448	2.345	2.018	0.755	0.371	2.574	1.970	0.721	0.428
EmoEditor	1.568	1.918	0.831	0.586	2.252	1.938	0.813	0.503	2.570	1.990	0.693	0.649
AffectRAG	1.434	1.691	0.889	0.446	1.967	1.673	0.902	0.345	1.637	1.435	0.851	0.405

Directional Control Reward Given the Planner-generated control direction $\boldsymbol{u} = (u_l, u_s, u_t)$, we define the directional reward as:

$$R_{\text{dir}} = \frac{1}{3} \sum_{m \in \{l,s,t\}} \mathbb{I}\left(u_m = u_m^{gt}\right). \tag{3}$$

This reward serves as the primary learning signal in RACO, encouraging the Planner to predict ternary visual-control directions from the retrieval residual.

Retrieval Coverage Reward Predicting correct control directions alone does not guarantee that the instruction actually uses the retrieved factor and attributes. The Planner may ignore them and produce only lighting and color adjustments. To encourage grounding in the retrieved semantic anchor, we introduce a retrieval coverage reward. For the retrieved factor f_r and attribute set $\boldsymbol{p}_r = \{q_1, \ldots, q_N\}$, the AKB stores a factor synonym list $\mathcal{A}_{f_r}$ and attribute synonym lists $\mathcal{A}_{q_n}$. We define the match function as $\text{match}(\tau, \mathcal{A}) = \mathbb{I}\left[\exists a \in \mathcal{A},\ a \subseteq \tau\right]$, which means that if the editing instruction τ contains any expression in the synonym list, the corresponding semantic item is considered covered. The retrieval reward is a combination of factor coverage and attribute coverage:

$$R_{\text{ret}} = \text{match}(\tau, \mathcal{A}_{f_r}) + \frac{1}{N} \sum_{n=1}^{N} \text{match}(\tau, \mathcal{A}_{q_n}). \tag{4}$$

Retrieval-Confidence Gating Not all retrieved anchors are equally reliable. When the retrieved anchor is close to the target V-A coordinate, it can serve as a strong semantic reference. When the anchor still deviates from the target, an overly strong semantic coverage constraint may restrict residual-conditioned visual adjustments and prevent the Planner from further approaching the target affect. We therefore introduce retrieval confidence c_r:

$$c_r = \exp\left(-\frac{\|\boldsymbol{z}_r - \boldsymbol{z}^*\|_1}{\tau_r}\right), \tag{5}$$

where τ_r is a temperature coefficient. The final gated reward is:

$$R(y) = R_{\text{dir}} + c_r R_{\text{ret}}. \tag{6}$$

Planner Optimization Given the retrieved anchors and affective residual, the Planner samples multiple candidate plans that combine discrete semantics and directional controls in different ways. RACO optimizes the Planner according to their relative rewards under the same retrieval-augmented context, encouraging plans that better match the control labels while retaining retrieved semantics.

4 Experiments

4.1 Experimental Setup

Datasets We evaluate AffectRAG on three public affective datasets: **EmoSet** (Yang et al., 2023), **EMOTIC** (Kosti et al., 2019), and **FindingEmo** (Mertens et al., 2024). Since EmoSet does not provide V–A annotations, we obtain V–A labels for its source images through human annotation. For EMOTIC and FindingEmo, we use the original annotations and linearly rescale them to the common range of [0, 9]. Target V-A values are randomly sampled within this range. All methods use the same source images, source annotations, and sampled target coordinates.



Figure 4: Comparison across four V-A quadrants. AffectRAG better matches target emotions while preserving semantics. Baselines either produce limited affective changes or compromise semantics.



Figure 5: Comparison of responses to continuous V–A targets. Our method produces finer-grained affective-visual transitions, while baselines show limited changes in elements and color tone.

Metrics We use four metrics. **V-Error** and **A-Error** measure affective accuracy by computing the absolute error between the target V-A values and the predictions of a commonly used CLIP-based affective evaluator as in Appendix (Mertens et al., 2024; Dang et al., 2025). **CLIP-I** and **LPIPS** measure semantic-level consistency between the source and edited images. CLIP-I measures semantic preservation in the CLIP (Radford et al., 2021) image embedding space. LPIPS (Zhang et al., 2018) measures perceptual similarity between images. Together, these metrics evaluate whether the edited image reaches the target emotion while preserving the source semantics.

Baselines We compare AffectRAG with three groups of state-of-the-art baselines: closed-source editing models (Nano Banana Pro (Raisinghani, 2025), Seedream 5.0 (ByteDance Seed Team, 2026)), trainable general editing models (Qwen-Image-Edit (Wu et al., 2025), FLUX.1-Kontext (Batifol et al., 2025), FLUX.2-dev (Black Forest Labs, 2025)), and affective editing methods (EmoEdit (Yang et al., 2025), EmoEditor (Lin et al., 2025)). For a fair comparison, all trainable baselines are trained on the training splits of each dataset using editing pairs constructed with our AKB. Since AffectRAG optimizes the Planner rather than the image editor, it uses directional control supervision instead of edited images for training, while retaining the same dataset splits.

4.2 Main Results

Quantitative Results Table 1 reports the quantitative results on the test splits of EmoSet, EMOTIC, and FindingEmo. AffectRAG outperforms all baselines on V-Error and A-Error across all datasets, showing a stronger ability to reach target continuous V-A coordinates. This is because AffectRAG grounds abstract affective values into retrieved semantic anchors and learns residual visual control, reducing the difficulty of numeric-to-visual mapping. AffectRAG also achieves the best or second-best CLIP-I and LPIPS scores, indicating that it preserves the main content of source image.

Qualitative Results Figure 4 qualitatively compares editing results in the four V-A quadrants. AffectRAG better reflects the target emotion while maintaining source semantics. It selects visually compatible elements and adjusts lighting and color toward the target direction. In contrast, existing

Table 2: User study of affective perception.

Method	V-Error ↓	A-Error ↓	Preservation ↑
Seedream 5.0	1.613	1.684	0.837
Nano Banana Pro	1.752	1.817	0.835
Qwen-Image-Edit	1.927	1.886	0.813
FLUX.1-Kontext	1.963	1.935	0.810
FLUX.2-dev	1.901	1.902	0.852
EmoEdit	1.865	1.948	0.786
EmoEditor	1.803	1.916	0.782
AffectRAG	**1.381**	**1.503**	**0.886**

Table 3: Ablation study of AffectRAG components.

Method	RAG	Dir.	Reward	V-Error ↓	A-Error ↓
Zero-shot	✗	✗	✗	1.982	1.996
Dir.-Only	✗	✓	✗	1.738	1.915
RAG-Only	✓	✗	✗	1.632	1.778
RAG + Dir.	✓	✓	SFT	1.614	1.760
w/o R_{dir}	✓	✗	$c_r R_{\text{ret}}$	1.625	1.769
w/o R_{ret}	✓	✓	R_{dir}	1.594	1.747
w/o c_r	✓	✓	$R_{\text{dir}} + R_{\text{ret}}$	1.603	1.751
AffectRAG	✓	✓	$R_{\text{dir}} + c_r R_{\text{ret}}$	**1.568**	**1.657**

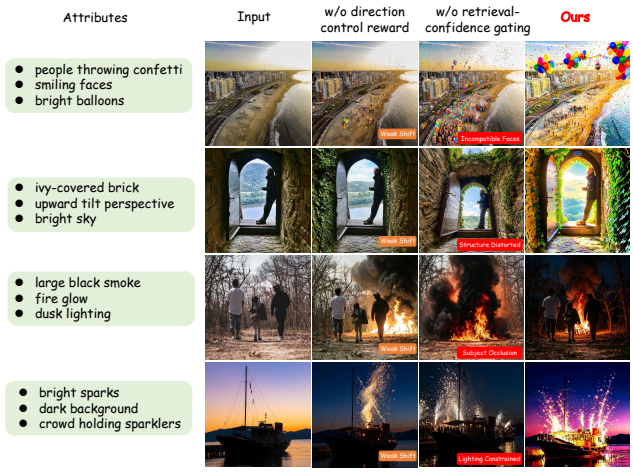

Figure 6: Ablation on RACO rewards. Removing direction reward weakens lighting-and-color adjustments. Removing gating causes over-rigid grounding to retrieved semantics.

Response to V–A Variations. To further evaluate continuous V–A perception, we select the strongest general and affective baselines, Seedream and EmoEdit, for comparison. We apply them to the same source image under gradually varying V–A targets. As shown in Figure 5, EmoEdit produces almost unchanged results, suggesting limited sensitivity to continuous values. Although Seedream reflects some affective evolution, its element and color-tone modifications remain insufficient. In contrast, AffectRAG shows finer-grained variations in background objects, brightness, color tone, and overall atmosphere, yielding emotionally coherent and visually appealing results.

methods often produce either weak affective changes or semantically mismatched edits, due to the lack of explicit knowledge connecting numerical values with visual semantics.

User Study To assess affective alignment and source-content preservation through human judgments, we conduct a human evaluation with 30 participants. All eight methods are evaluated on the same 120 source–target pairs sampled from the test splits, with 40 pairs per dataset and 10 pairs per V–A quadrant within each dataset. Method identities are hidden, and presentation order is randomized. Participants rate perceived V-A on a 0–9 scale and content preservation on a 0–1 scale, considering identity, objects, layout, and core semantics. V-Error and A-Error are computed between each output's mean human ratings and its target coordinates, then averaged across pairs. Preservation ratings are likewise averaged per output and then across pairs. As shown in Table 2, AffectRAG achieves the lowest human-rated V-Error and A-Error and the highest preservation score, supporting the improvements in affective alignment and content preservation observed in automatic evaluation.

4.3 Ablation Studies

Retrieved Anchoring and Residual Control Table 3 studies the contribution of key components of AffectRAG: knowledge retrieval (RAG) and ternary directional control (Dir.). Introducing directional control improves affective accuracy over the zero-shot Planner, indicating the effectiveness of explicit lighting-and-color control. The RAG-only variant outperforms the Dir.-only variant because retrieved semantic anchors can move the image toward a more appropriate affective region than local visual control alone. Combining both components further improves Planner performance, and RACO achieves larger gains than standard SFT. This shows that reward optimization with directional and gated retrieval rewards is more effective for learning residual control.

RACO Rewards Table 3 also examines the direction reward, retrieval reward, and confidence gating in the RACO objective. The variant without direction reward exhibits the largest increase in

Table 4: Retrieval strategy ablation. Our top-1 retrieval w/ filtering yields the best.

Retrieval setting	V-Error ↓	A-Error ↓
Shuffled coordinates	2.013	2.028
Random anchors	1.834	1.907
Top-3 anchors	1.616	1.793
w/o filtering	1.591	1.765
AffectRAG	**1.568**	**1.657**

Table 5: Cross-dataset retrieval statistics. Coverage @ 1.5 is the fraction of anchor–target distances below 1.5.

Metric	EmoSet	EMOTIC	FindingEmo
Avg. Anchor Distance ↓	0.4774	0.5374	0.5208
Avg. Retrieval Confidence ↑	0.7877	0.7644	0.7707
Avg. Compatible Candidates ↑	10350.38	9828.17	10013.25
Coverage @ 1.5 ↑	0.9715	0.9130	0.8750

Table 6: Local V–A sensitivity. Monotonicity measures adjacent-target directional consistency; Rel. Red. is error reduction relative to RAG-Only; Shift Corr. averages within-trajectory Spearman correlations between residual magnitude and predicted affective shifts along the target direction.

(a) Same-anchor residual sweep

Method	V-Error ↓	V-Monotonicity ↑	V-Rel. Red. ↑	A-Error ↓	A-Monotonicity ↑	A-Rel. Red. ↑
RAG-Only	0.406	0.566	0.0%	0.428	0.548	0.0%
AffectRAG	**0.219**	**0.842**	**46.1%**	**0.256**	**0.811**	**40.2%**

(b) Same-anchor, same-direction magnitude sweep

Method	V-Error ↓	V-Monotonicity ↑	V-Shift Corr. ↑	A-Error ↓	A-Monotonicity ↑	A-Shift Corr. ↑
RAG-Only	0.394	0.572	0.143	0.418	0.556	0.119
AffectRAG	**0.208**	**0.850**	**0.819**	**0.247**	**0.821**	**0.776**

V-A Errors, highlighting the importance of directional control for residual adjustment. Removing retrieval reward leads to a more moderate degradation, suggesting that explicit semantic coverage helps the effective use of retrieved anchors. Removing confidence gating also increases both errors. Together, these components provide complementary benefits for affective alignment. Figure 6 further illustrates the effects. Without directional reward, the Planner cannot effectively use lighting-and-color adjustments to approach the target emotion. Without confidence gating, it tends to follow retrieved content too rigidly, causing conflicts such as incompatible elements or subject occlusion.

Cross-Dataset Retrieval We assess the transferability of the AKB, constructed exclusively from the EmoSet training split, using four retrieval statistics on the test sets of EmoSet, EMOTIC, and FindingEmo (Table 5). On EMOTIC and FindingEmo, the mean number of candidates retained after compatibility filtering (Avg. Compatible Candidates) remains large, and retrieval maintains high coverage. The average distance between retrieved anchors and target V–A coordinates (Avg. Anchor Distance) and the mean retrieval-confidence score (Avg. Retrieval Confidence) are also close to those on EmoSet. These results suggest that the AKB's affective–visual associations extend beyond EmoSet and can provide relevant semantic anchors for datasets unseen during its construction.

Sensitivity to Continuous V–A Targets To examine whether AffectRAG responds to fine-grained V–A target changes beyond anchor selection, we compare AffectRAG with RAG-Only under two local target sweeps. The first varies valence or arousal around a fixed semantic anchor. The second varies residual magnitude within subsets sharing both the same anchor and the same predicted ternary direction vector. Detailed protocols and metric definitions are provided in Appendix D. As shown in Table 6, compared with RAG-Only, AffectRAG greatly improves local V–A accuracy and monotonicity with fixed anchors, and V–A shift correlations within same-direction subsets. These results indicate that AffectRAG exhibits more accurate and consistently ordered responses to small target changes, including residual-magnitude variations that leave the direction vector unchanged.

5 Conclusion

In this paper, we study continuous affective image editing (CAIE), which aims to edit an image toward a target V-A while preserving its source semantics. We propose **AffectRAG** that decomposes CAIE into semantic grounding and residual control. To translate abstract affective values into editable visual semantics, we construct a large-scale AKB to organize affective-visual associations. To improve affective editing precision, RACO optimizes the planner to map anchor-target residuals into visual-control directions. Experiments on multiple datasets show that AffectRAG achieves better V-A accuracy and semantic preservation than SOTA affective and general editing baselines.

THE USE OF LARGE LANGUAGE MODELS

In this work, we use Large Language Models (LLMs) solely to assist with manuscript preparation. Specifically, LLMs are employed to improve linguistic clarity and readability, including refining phrasing and suggesting more precise terminology. We also use LLMs to assist with LaTeX formatting for figures, tables, and mathematical expressions, thereby reducing the time and effort required for complex typesetting. All research ideas, methodological design, experimental analysis, and scientific conclusions are developed and verified by the authors.

REFERENCES

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

Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv e-prints*, pp. arXiv–2506, 2025.

Black Forest Labs. FLUX.2: Frontier visual intelligence. `https://bfl.ai/blog/flux-2`, 2025. Accessed: 2026-05-25.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. Improving language models by retrieving from trillions of tokens. In *International conference on machine learning*, pp. 2206–2240. PMLR, 2022.

Damian Borth, Tao Chen, Rongrong Ji, and Shih-Fu Chang. Sentibank: large-scale ontology and classifiers for detecting sentiment and emotions in visual content. In *Proceedings of the 21st ACM international conference on Multimedia*, pp. 459–460, 2013. doi: 10.1145/2502081.2502268.

Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18392–18402, 2023. doi: 10.1109/CVPR52729.2023.01764.

ByteDance Seed Team. Deeper thinking, more accurate generation: Introducing seedream 5.0 lite. `https://seed.bytedance.com/en/blog/deeper-thinking-more-accurate-generation-introducing-seedream-5-0-lite`, February 2026. Accessed: 2026-05-25.

Alan S Cowen and Dacher Keltner. Self-report captures 27 distinct categories of emotion bridged by continuous gradients. *Proceedings of the national academy of sciences*, 114(38):E7900–E7909, 2017.

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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15218–15228, 2025. doi: 10.1109/ICCV51701.2025.01412.

Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, volume 96, pp. 226–231, 1996.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022.

Ronak Kosti, Jose M Alvarez, Adria Recasens, and Agata Lapedriza. Context based emotion recognition using emotic dataset. *IEEE transactions on pattern analysis and machine intelligence*, 42 (11):2755–2766, 2019. doi: 10.1109/TPAMI.2019.2916866.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33: 9459–9474, 2020.

Zheng Lian, Haoyu Chen, Lan Chen, Haiyang Sun, Licai Sun, Yong Ren, Zebang Cheng, Bin Liu, Rui Liu, Xiaojiang Peng, et al. Affectgpt: A new dataset, model, and benchmark for emotion understanding with multimodal large language models. *arXiv preprint arXiv:2501.16566*, 2025.

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. doi: 10.1109/ICCV51701.2025.01519.

Da Liu, Yaxi Jiang, Min Pei, and Shiguang Liu. Emotional image color transfer via deep learning. *Pattern Recognition Letters*, 110:16–22, 2018. doi: 10.1016/j.patrec.2018.03.015.

Jana Machajdik and Allan Hanbury. Affective image classification using features inspired by psychology and art theory. In *Proceedings of the 18th ACM international conference on Multimedia*, pp. 83–92, 2010. doi: 10.1145/1873951.1873965.

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

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. doi: 10.52202/079017-0161.

Reni Paskaleva, Mykyta Holubakha, Andela Ilic, Saman Motamed, Luc Van Gool, and Danda Paudel. A unified and interpretable emotion representation and expression generation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2447–2456. IEEE, 2024.

Luiz Pessoa. On the relationship between emotion and cognition. *Nature reviews neuroscience*, 9 (2):148–158, 2008.

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.

Naina Raisinghani. Introducing nano banana pro. `https://blog.google/innovation-and-ai/products/nano-banana-pro/`, November 2025. Accessed: 2026-05-25.

Tianrong Rao, Xiaoxu Li, and Min Xu. Learning multi-level deep representations for image emotion classification. *Neural processing letters*, 51(3):2043–2061, 2020. doi: 10.1007/s11063-019-10033-9.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 3982–3992, 2019. doi: 10.18653/v1/D19-1410.

James A Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39 (6):1161, 1980. doi: 10.1037/h0077714.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022. doi: 10.52202/068431-1800.

Shuchen Weng, Peixuan Zhang, Zheng Chang, Xinlong Wang, Si Li, and Boxin Shi. Affective image filter: Reflecting emotions from text to images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10810–10819, 2023. doi: 10.1109/ICCV51070.2023.00992.

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*, 2025.

Chuan-Kai Yang and Li-Kai Peng. Automatic mood-transferring between color images. *IEEE computer graphics and applications*, 28(2):52–61, 2008. doi: 10.1109/MCG.2008.24.

Jingyuan Yang, Jie Li, Xiumei Wang, Yuxuan Ding, and Xinbo Gao. Stimuli-aware visual emotion analysis. *IEEE Transactions on Image Processing*, 30:7432–7445, 2021. doi: 10.1109/TIP.2021.3106813.

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. doi: 10.1109/ICCV51070.2023.01864.

Jingyuan Yang, Jiawei Feng, and Hui Huang. Emogen: Emotional image content generation with text-to-image diffusion models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6358–6368. IEEE, 2024.

Jingyuan Yang, Jiawei Feng, Weibin Luo, Dani Lischinski, Daniel Cohen-Or, and Hui Huang. Emoedit: Evoking emotions through image manipulation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 24690–24699, 2025. doi: 10.1109/CVPR52734.2025.02299.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2022.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3836–3847, 2023. doi: 10.1109/ICCV51070.2023.00355.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018. doi: 10.1109/CVPR.2018.00068.

Sicheng Zhao, Xingxu Yao, Jufeng Yang, Guoli Jia, Guiguang Ding, Tat-Seng Chua, Bjoern W Schuller, and Kurt Keutzer. Affective image content analysis: Two decades review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6729–6751, 2021.

Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. Large language models are human-level prompt engineers. In *The eleventh international conference on learning representations*, 2022.

APPENDIX

A ADDITIONAL QUALITATIVE RESULTS

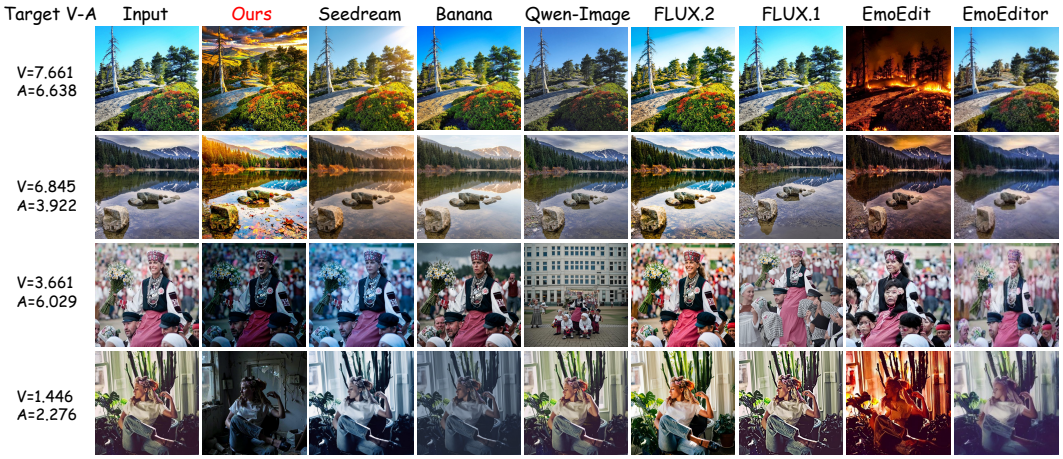

Figure 7: Comparison across four V-A quadrants. AffectRAG better matches the target emotion while preserving semantics.

B MODELS AND BASELINE METHODS

Table 7 lists the Planner and Executor backbones used in AffectRAG and the image-editing baselines, together with their official repositories or model pages.

Affective evaluator. For all automatic V–A evaluations, we use the pretrained CLIP-based affective prediction model and the released checkpoint provided by Mertens et al. (2024). The evaluator follows the architecture described in their work and predicts continuous valence and arousal values from an input image. We keep the evaluator frozen throughout all experiments and do not fine-tune it on any of our evaluation images.

Table 7: Planner and Executor backbones and image-editing baselines used in our experiments.

Model	**Official link**
AffectRAG backbones	
Qwen3-VL-4B-Instruct	https://huggingface.co/Qwen/Qwen3-VL-4B-Instruct
FLUX.2-klein-9B	https://huggingface.co/black-forest-labs/FLUX.2-klein-9B
Image-editing baselines	
FLUX.2-dev	https://huggingface.co/black-forest-labs/FLUX.2-dev
FLUX.1-Kontext-dev	https://huggingface.co/black-forest-labs/FLUX.1-Kontext-dev
Qwen-Image-Edit-2511	https://huggingface.co/Qwen/Qwen-Image-Edit-2511
Nano Banana Pro	https://ai.google.dev/gemini-api/docs/models/gemini-3-pro-image
Seedream 5.0 Lite	https://seed.bytedance.com/en/seedream5_0_lite
EmoEdit	https://github.com/JingyuanYY/EmoEdit
EmoEditor	https://github.com/ZhangLab-DeepNeuroCogLab/EmoEditor

C ANNOTATION AND DATA PREPARATION

Source-image annotations and target sampling The source-image V–A annotations for EmoSet were collected through a human annotation procedure involving 50 annotators. Source-image ratings were gathered in the same annotation sessions as the directional control labels described below. For EMOTIC and FindingEmo, we retained the original source annotations and mapped each affective dimension to [0, 9] using a linear transformation based on the endpoints of its native scale. Target

coordinates were generated by randomly sampling valence and arousal values within [0, 9]. The resulting source images, source annotations, and target coordinates were shared across methods.

Attribute-level affective annotation Fifty annotators participated in the construction of the AKB's affective coordinates. Before rating, participants received a short introduction to valence and arousal and their interpretation on the 0–9 scale. Attributes were presented as standalone textual descriptions, with their original images withheld. Five annotators separately assessed each entry, assigning one valence score and one arousal score according to the emotional impression expressed by the description. A limited number of entries appeared more than once to measure the consistency of individual raters. After incomplete and unreliable responses were removed, the retained scores were averaged separately for valence and arousal to obtain the coordinate associated with each attribute.

Directional control annotation Each annotation item contained a source image, an edited image generated from its retrieved anchor, and a sampled target V–A coordinate. Annotators compared the edited image's perceived affect with the target and specified the desired adjustment for each of three visual properties: illumination, saturation, and color temperature. Illumination and saturation were labeled as decrease, keep, or increase. Color temperature was labeled as cooler, keep, or warmer.

D Local Target Sweeps and Evaluation Metrics

Evaluation setup. We compare AffectRAG with RAG-Only on held-out source images under two local target sweeps. RAG-Only retains semantic retrieval but omits residual-conditioned directional control. Both methods use the same source images, fixed semantic anchors, and target coordinates. Output V–A are estimated using the same frozen affective evaluator as in the main experiments.

Throughout this section, i indexes an evaluation trajectory and j indexes a target within that trajectory. Each trajectory uses one fixed source image and semantic anchor, while varying one affective dimension $k \in \{V, A\}$. A source image can contribute multiple trajectories. Let $z_{r,i} = (v_{r,i}, a_{r,i})^\top$ denote the fixed anchor, and let $e_V = (1, 0)^\top$ and $e_A = (0, 1)^\top$. All local targets take the form

$$z^*_{ij} = z_{r,i} + \delta_{ij} e_k, \tag{7}$$

where δ_{ij} is the signed residual along dimension k. Only targets within $[0, 9]^2$ are retained.

Same-anchor residual sweep. For each source image and fixed anchor, we vary valence and arousal separately using

$$\delta_{ij} \in \{-0.8, -0.6, -0.4, -0.2, 0, 0.2, 0.4, 0.6, 0.8\}. \tag{8}$$

The retained targets are ordered by increasing δ_{ij}. The other target coordinate remains fixed at its anchor value. This sweep evaluates sensitivity to small target changes while holding semantic retrieval fixed. We report V-Error, V-Monotonicity, and V-Rel. Red. for valence trajectories, and the corresponding A-metrics for arousal trajectories.

Same-anchor, same-direction magnitude sweep. We further analyze subsets of the first sweep that share the same complete predicted ternary direction vector. Positive and negative residuals are considered separately:

$$\delta_{ij} = \sigma_i \rho_{ij}, \qquad \sigma_i \in \{-1, +1\}, \qquad \rho_{ij} \in \{0.2, 0.4, 0.6, 0.8\}, \tag{9}$$

where σ_i specifies the signed target direction and ρ_{ij} is the residual magnitude.

For each source image, anchor, affective dimension, and residual sign, we group targets by AffectRAG's complete predicted visual-control vector $u \in \{-1, 0, 1\}^3$. Only groups containing at least three distinct magnitudes are retained. Each group forms a trajectory ordered by increasing ρ_{ij}, and RAG-Only is evaluated on exactly the same retained targets. This analysis characterizes magnitude sensitivity within the subset where the anchor and predicted control vector remain unchanged. We report V-Error, V-Monotonicity, and V-Shift Corr., together with their arousal counterparts.

Metric notation. Metrics are computed separately for each sweep and affective dimension. Let N_k be the number of trajectories varying dimension k, and let J_i be the number of retained targets in trajectory i. For method M, let $\hat{z}^M_{ij}$ denote the evaluator's V–A prediction for the edited output. All metrics are first computed within each trajectory and then averaged across trajectories.

V-Error and A-Error (↓). Local affective error is the mean absolute deviation from the target along the varied dimension:

$$\mathrm{Error}_k^M = \frac{1}{N_k}\sum_{i=1}^{N_k}\frac{1}{J_i}\sum_{j=1}^{J_i}\left|\hat{z}_{ij,k}^M - (z_{r,i,k} + \delta_{ij})\right|. \tag{10}$$

Setting $k = V$ and $k = A$ gives V-Error and A-Error, respectively. These errors characterize the sampled anchor neighborhoods, rather than the full target distribution of the main benchmark.

V-Monotonicity and A-Monotonicity (↑). For adjacent retained targets, the requested and predicted coordinate changes are

$$\begin{aligned}\Delta_{ij}^* &= \delta_{i,j+1} - \delta_{ij},\\ \Delta_{ij,k}^M &= \hat{z}_{i,j+1,k}^M - \hat{z}_{ij,k}^M.\end{aligned} \tag{11}$$

Monotonicity measures the fraction of adjacent responses that follow the requested direction:

$$\mathrm{Monotonicity}_k^M = \frac{1}{N_k}\sum_{i=1}^{N_k}\frac{1}{J_i-1}\sum_{j=1}^{J_i-1}\mathbb{I}\left[\Delta_{ij}^*\Delta_{ij,k}^M > 0\right]. \tag{12}$$

V-Monotonicity and A-Monotonicity correspond to $k = V$ and $k = A$. The sign-product criterion handles both increasing and decreasing target coordinates. Tied predictions are not counted as successful directional responses.

V-Rel. Red. and A-Rel. Red. (↑). For the same-anchor residual sweep, relative error reduction is computed against RAG-Only:

$$\text{Rel. Red.}_k^M = 100\left(1 - \frac{\mathrm{Error}_k^M}{\mathrm{Error}_k^{\text{RAG-Only}}}\right). \tag{13}$$

V-Rel. Red. and A-Rel. Red. are reported as percentages. RAG-Only has a reference value of 0%; positive values indicate lower error on the same trajectories.

V-Shift Corr. and A-Shift Corr. (↑). For each retained magnitude-sweep trajectory, we first orient the predicted coordinate along the requested target direction $p_{ij}^M = \sigma_i \hat{z}_{ij,k}^M$. An increase in p_{ij}^M therefore corresponds to movement along the requested direction, including when the target coordinate decreases. The within-trajectory correlation is

$$r_i^M = \mathrm{Spearman}\left(\{\rho_{ij}\}_{j=1}^{J_i}, \{p_{ij}^M\}_{j=1}^{J_i}\right). \tag{14}$$

Because subtracting a trajectory-specific constant does not change ranks, this is equivalent to correlating residual magnitude with the directional affective shift relative to the zero-residual output.

Tied predictions receive average ranks. A constant-response trajectory is assigned $r_i^M = 0$ for aggregation, since its Spearman correlation is otherwise undefined. The reported metric is

$$\text{Shift Corr.}_k^M = \frac{1}{N_k}\sum_{i=1}^{N_k} r_i^M. \tag{15}$$

V-Shift Corr. and A-Shift Corr. are computed separately over valence and arousal magnitude-sweep trajectories. Higher values indicate that larger residual magnitudes produce more consistently ordered affective shifts along the requested direction.

E Planner Input and Output

With the retrieved factor f_r, attribute p_r, anchor coordinate $\boldsymbol{z}_r$, and residual $\Delta\boldsymbol{z} = \boldsymbol{z}^* - \boldsymbol{z}_r$, the Planner's input context and output plan are written as:

$$\begin{aligned}x &= (\boldsymbol{I}_s, \boldsymbol{z}^*, f_r, p_r, \boldsymbol{z}_r, \Delta\boldsymbol{z}),\\ y &= (\boldsymbol{u}, \tau) \sim \pi_\theta(\cdot \mid x).\end{aligned} \tag{16}$$

Table 8: Prompt template used by the AffectRAG Planner.

Planner Prompt

You are an affective image-editing planner. Given a source image, a target valence–arousal (V–A) coordinate, and a retrieved semantic anchor, generate an editing plan that improves alignment with the target affect while preserving the main source content.

Both valence and arousal range from 0 to 9:

- Valence describes how negative or positive the emotion evoked by the image is.
- Arousal describes how calming or stimulating the image is.

The semantic anchor contains an affective factor, a visual attribute, the attribute's V–A coordinate, and a residual representing the coordinate-wise difference between the target and the anchor.

Planning requirements:

1. Inspect the source image to understand its main subjects, scene content, and spatial layout.
2. Use the retrieved factor and attribute to guide concrete semantic edits that fit the source image.
3. Predict additional visual-control directions for illumination, saturation, and color temperature to further approach the target affect. Represent them as a three-integer vector in that order. For illumination and saturation, 1 means increase, -1 means decrease, and 0 means keep unchanged. For color temperature, 1 means warmer, -1 means cooler, and 0 means keep unchanged.
4. Generate a self-contained editing instruction based on the retrieved anchor and predicted control directions. Specify the strength and spatial scope of the visual control according to the residual magnitude.
5. Preserve the main subject, scene structure, and spatial layout. Keep the changes visually plausible.
6. Do not mention valence, arousal, coordinates, or analysis in the `"instruction"` field. Express the desired changes through concrete visual descriptions.
7. Output only a valid JSON object with exactly two keys: {`"instruction":` the editing instruction string, `"control":` the three-integer direction vector}. Do not include Markdown or additional text.

Input Fields

Retrieved affective factor: {`factor_name`}
Retrieved visual attribute: {`visual_attr`}
Target affect: V={`target_v`}, A={`target_a`}
Anchor coordinate: V={`anchor_v`}, A={`anchor_a`}
Residual: Delta V={`target_v - anchor_v`}, Delta A={`target_a - anchor_a`}

Table 9: Prompt template used for general-purpose image-editing baselines.

Baseline Editing Prompt

You are an expert in emotional image editing.

You will be given:

- the source image
- the source and target emotion positions in valence–arousal (V–A) space

Both valence and arousal range from 0 to 9:

- Valence describes how negative or positive the emotion evoked by the image is.
- Arousal describes how calming or stimulating the image is.

Your task is to edit the source image to produce an edited image with the target V–A values. Please preserve the main subject, scene structure, spatial layout, and semantic plausibility of the source image.

Input Fields

Source emotion: V={`source_v`}, A={`source_a`}
Target emotion: V={`target_v`}, A={`target_a`}

The output contains the ternary visual-control vector $\boldsymbol{u} = (u_l, u_s, u_t) \in \{-1, 0, +1\}^3$ and the editing instruction τ. The control vector is used for directional supervision, while the instruction is passed to the Executor. The prompt for AffectRAG Planner is provided in Table 8.

The prompt used for the general-purpose editing baselines in our paper is given in Table 8.

F RACO Optimization Objective

Following the group relative policy optimization framework (Shao et al., 2024), we sample G candidate plans $y_g = (\boldsymbol{u}_g, \tau_g)$ from the old policy $\pi_{\theta_{\text{old}}}(\cdot \mid x)$ for each input x. Each plan receives a RACO reward R_g. We compute its group-relative advantage as $A_g = (R_g - \bar{R})/(\sigma_R + \varepsilon_{\text{adv}})$, where $\bar{R}$ and σ_R are the group reward mean and standard deviation, and ε_{adv} is a small numerical stabilizer. The Planner minimizes

$$\mathcal{L}_{\text{RACO}} = -\mathbb{E}_{x,\{y_g\}}\left[\frac{1}{G}\sum_{g=1}^{G}\mathcal{L}_g - \beta D_{\text{KL}}\left(\pi_\theta \| \pi_{\text{ref}}\right)\right], \tag{17}$$

where the clipped surrogate objective for each plan is

$$\mathcal{L}_g = \min\Big\{r_g(\theta)A_g, \text{clip}(r_g(\theta), 1-\epsilon, 1+\epsilon)\, A_g\Big\}. \tag{18}$$

Here, $r_g(\theta) = \pi_\theta(y_g \mid x)/\pi_{\theta_{\text{old}}}(y_g \mid x)$ is the likelihood ratio of the same plan under the current and old policies. The parameter ϵ specifies the clipping threshold, β controls the KL regularization strength, and π_{ref} denotes the reference policy.

G Hyperparameters Details

Table 10 summarizes the training hyperparameters used in our experiments.

Table 10: Training hyperparameters.

Hyperparameter	**Value**
GPU Configuration	
GPU type	NVIDIA H20
Number of GPUs	8
Training Hyperparameters	
Number of sampled outputs	8
Training epochs	3
Global batch size	8
Reward Hyperparameters	
Retrieval-confidence temperature τ_r	2.0
Edit-difficulty penalty coefficient λ_d	0.5

H Metadata and Factor Prototypes in Affective Knowledge Base

The detailed candidate values of three editing metadata dimensions are listed in Table 11. We provide all affective factor prototypes in the Affective Knowledge Base in Table 12 and Table 13.

I Limitation and Future Work

Our current RACO objective explicitly supervises ternary visual-control directions, whereas residual magnitude is used as contextual information for conditioning the strength and spatial scope of editing instructions. This design provides effective continuous responses in our experiments, but does not directly optimize a dedicated magnitude-level objective. Future work could introduce finer-grained control signals for adjustment strength to further improve precision in continuous V–A editing.

Table 11: Candidate values of the three editing metadata dimensions defined for each affective factor.

Dimension	Candidate Values
edit_scope	facial body_pose character_group animal_subject object_local food_beverage decoration_prop atmosphere_effect scene_local scene_global architecture_structure symbolic_object
insertion_pattern	replace_subject_trait add_subject_adjacent add_foreground_object add_background_context global_restyle add_group_event structural_recompose
edit_difficulty	low medium high

Table 12: All factor prototypes in Affective Knowledge Base (Part 1 of 2).

#	Factor Prototype	Edit Scope	Diff.	Insertion Pattern
1	Beach Leisure	scene_global	high	global_restyle
2	Joyful & Playful Facial Expression	facial	low	replace_subject_trait
3	Relaxed Social Leisure	character_group	medium	add_group_event
4	Vibrant & Playful Beverage Presentation	food_beverage	low	add_foreground_object
5	Thrilling Aerial Amusement Ride	scene_global	high	structural_recompose
6	Vivid Festive Decorative Display	decoration_prop	medium	add_background_context
7	Tiny Creature & Flower Playful Interaction	animal_subject	medium	add_subject_adjacent
8	Joyful Group Celebration	character_group	high	add_group_event
9	Appetizing Food Presentation	food_beverage	low	add_foreground_object
10	Joyful Water Play & Recreation	character_group	high	add_group_event
11	Vibrant Swirling Motion Blur	atmosphere_effect	medium	global_restyle
12	Amusing Animal Behavior	animal_subject	medium	add_subject_adjacent
13	Playful Festive Decoration	decoration_prop	low	add_background_context
14	Expressive Dance Performance	body_pose	medium	replace_subject_trait
15	Playful Pet Interaction	animal_subject	medium	add_subject_adjacent
16	Character & Group Pose Interaction	character_group	medium	add_group_event
17	Amusement Ride & Decoration	scene_global	high	global_restyle
18	Indulgent Dessert Presentation	food_beverage	low	add_foreground_object
19	Playful Recreational Interaction	character_group	medium	add_group_event
20	Vibrant Playful Costume Performance	character_group	medium	add_subject_adjacent
21	Protest & Demonstration Action	character_group	high	add_group_event
22	Intense Destructive Fire	scene_global	high	add_background_context
23	Aggressive Threat Expression	facial	low	replace_subject_trait
24	Confrontational Protest Signage	object_local	medium	add_foreground_object
25	Physical Attack	body_pose	high	replace_subject_trait
26	Dynamic Athletic Feat	body_pose	medium	replace_subject_trait
27	Sacred Reverence & Religious Ceremony	character_group	high	add_group_event
28	Intimate Affectionate Gaze Exchange	character_group	medium	add_subject_adjacent
29	Wonder & Awe Expression	facial	low	replace_subject_trait
30	Vast Sunset Horizon Radiance	scene_global	medium	global_restyle
31	Formal Ceremonial Spectacle	character_group	high	add_group_event
32	Delicate Natural Creature Beauty	animal_subject	medium	add_subject_adjacent
33	Sunset Silhouette & Water Reflection	scene_global	medium	global_restyle
34	Microscopic Natural Intricate Structure	object_local	high	add_foreground_object
35	Vivid Color Contrast	atmosphere_effect	medium	global_restyle
36	Majestic Nature & Wildlife	scene_global	high	global_restyle
37	Wildlife Natural Wonder	animal_subject	medium	add_subject_adjacent
38	Grand Vast Natural Landscape	scene_global	high	global_restyle
39	Architectural Scale	architecture_structure	high	structural_recompose
40	Vivid Iridescent Animal Plumage	animal_subject	medium	add_subject_adjacent
41	Symmetrical Mirror Reflection	atmosphere_effect	medium	global_restyle
42	Geological Rock Formation	scene_global	high	global_restyle
43	Overwhelming Cityscape Vista	scene_global	high	global_restyle
44	Vivid Delicate Floral	object_local	low	add_foreground_object
45	Natural Harvest Abundance Beauty	object_local	medium	add_foreground_object
46	Sweet Dessert	food_beverage	low	add_foreground_object
47	Comforting Food	food_beverage	low	add_foreground_object
48	Pet Affectionate Bonding Interaction	animal_subject	medium	add_subject_adjacent

Table 13: All factor prototypes in Affective Knowledge Base (Part 2 of 2, continued from Table 12).

#	Factor Prototype	Edit Scope	Diff.	Insertion Pattern
49	Cozy Warm Beverage Leisure	food_beverage	low	add_foreground_object
50	Vibrant Floral Bloom	object_local	low	add_foreground_object
51	Festive Celebratory Decoration	decoration_prop	low	add_background_context
52	Serene Focused Activity	body_pose	medium	replace_subject_trait
53	Positive Facial Expression	facial	low	replace_subject_trait
54	Serene Aquatic Animals	animal_subject	medium	add_subject_adjacent
55	Natural Leisure & Relaxation	scene_global	medium	global_restyle
56	Vivid Floral Decorative Beauty	decoration_prop	low	add_background_context
57	Solitary Contemplative Introspection	body_pose	medium	replace_subject_trait
58	Peaceful Seaside Relaxation	scene_global	medium	global_restyle
59	Unusual Repulsive Food	food_beverage	low	add_foreground_object
60	Overwhelming Decay & Pollution	scene_global	high	global_restyle
61	Vivid Bodily Injury Wound	object_local	medium	add_subject_adjacent
62	Dynamic Performance & Joyful Energetic Expression	character_group	medium	add_group_event
63	High-Intensity Outdoor Adventure	scene_global	high	global_restyle
64	High-Speed Dynamic Motion	atmosphere_effect	medium	global_restyle
65	Panic Motion Blur	atmosphere_effect	medium	global_restyle
66	Violent Injury Threatening	object_local	medium	add_subject_adjacent
67	Fearful Facial Expression	facial	low	replace_subject_trait
68	Fire and Disaster Scene	scene_global	high	global_restyle
69	Protective Gear against Invisible Threat	object_local	medium	add_subject_adjacent
70	Abandoned Building Decay	architecture_structure	high	structural_recompose
71	Catastrophic Structural Collapse	architecture_structure	high	structural_recompose
72	Abandoned Eerie Objects	symbolic_object	medium	add_foreground_object
73	Ominous Darkness and Light	atmosphere_effect	medium	global_restyle
74	Dark Confined Space	scene_global	high	global_restyle
75	Skull Death Symbol	symbolic_object	low	add_foreground_object
76	Supernatural Horror Entity	character_group	high	add_subject_adjacent
77	Distorted Doll Face	symbolic_object	medium	add_foreground_object
78	Violent Weapon Threat	object_local	medium	add_subject_adjacent
79	Looming Threat	character_group	medium	add_subject_adjacent
80	Fearful Body Posture	body_pose	low	replace_subject_trait
81	Clown with Scary Features	character_group	medium	add_subject_adjacent
82	Threatening Arthropod	animal_subject	medium	add_foreground_object
83	Predatory Threat	animal_subject	medium	add_subject_adjacent
84	Menacing Masked Figure	character_group	medium	add_subject_adjacent
85	Horror Face	facial	low	replace_subject_trait
86	Lifeless Animal & Abandoned Object	symbolic_object	medium	add_foreground_object
87	Dim Atmosphere	atmosphere_effect	low	global_restyle
88	Natural Decay	scene_local	medium	add_background_context
89	Abandoned Ruin	architecture_structure	high	structural_recompose
90	Solitary Contemplation	body_pose	low	replace_subject_trait
91	Sorrowful Expression	facial	low	replace_subject_trait
92	Funerary Memorial	symbolic_object	medium	add_foreground_object
93	Mourning Figure	character_group	medium	add_subject_adjacent
94	Gloomy Cemetery	scene_global	high	global_restyle
95	Memorial Tribute	symbolic_object	medium	add_foreground_object