AFFECTAGENT: GROUNDED PLANNING AND MEMORY CALIBRATION FOR AFFECTIVE IMAGE EDITING

Anonymous authors
Paper under double-blind review

ABSTRACT

Affective image editing aims to align visual content with a target affect while preserving its underlying semantics. The same affective target can be realized through multiple visual modifications. However, existing end-to-end editors jointly perform affective reasoning and image generation, leaving editing strategies implicit and insufficiently grounded in affective semantics. Moreover, current optimization methods often apply task-agnostic supervision, failing to adapt learning strength to varying affective transitions. To address these limitations, we introduce AffectAgent, a planner–executor–evaluator framework that combines grounded planning with memory calibration for controllable affective image editing. Specifically, we propose Knowledge-Grounded Multi-Plan Exploration (KGME), which retrieves affective knowledge to explicitly construct multiple semantically meaningful editing plans and optimize the planner over a grounded strategy space. We further introduce Memory-Calibrated Preference Optimization (MCPO) to organize historical outcomes into an episodic memory and calibrate each sample's optimization strength against the performance distribution of tasks with similar affective transitions. Extensive experiments show that AffectAgent consistently reduces valence and arousal errors compared with state-of-the-art methods while maintaining competitive semantic consistency, demonstrating more accurate target-affect control.

1 INTRODUCTION

Affective image editing (AIE) aims to transform the perceived affect of an image toward a desired target while preserving its semantic content and visual identity. By bridging abstract affective intent and concrete visual manipulation, AIE provides a fundamental capability for human-centered visual generation, where images are expected to convey not only specified content but also desired affective states. Visual affect can be conveyed through diverse cues, including color, illumination, composition, facial expression, and scene context (Machajdik & Hanbury, 2010; Zhao et al., 2021; Yang et al., 2021). Continuous representations such as the valence–arousal space further enable fine-grained specification of affective transitions (Russell, 1980; Posner et al., 2005). Unlike edits defined by explicit operations, however, a target affect can be realized through multiple visual strategies. Effective AIE therefore requires selecting editing strategies that are appropriate for both the source image and the desired affective transition.

General-purpose generative editors have demonstrated strong instruction-following capabilities, enabling diverse visual modifications through natural-language commands and providing a basis for affect-aware image manipulation (Kawar et al., 2023; Sheynin et al., 2024). More recent AIE methods further incorporate learned emotion embeddings into generative editors to support affect-specific transformations(Zhang et al., 2026; Ye & Huang, 2025). Beyond model design, preference-based optimization further improves generation by learning to favor higher-quality outputs from pairwise or group-relative comparisons(Yang et al., 2024b; Liu et al., 2026).

Despite these advances, two limitations remain. First, existing end-to-end editors typically entangle editing-strategy selection with image generation (Weng et al., 2023; Yang et al., 2024a). Although a target affect can be realized through multiple visual strategies(Mao et al., 2026), the model implicitly resolves this one-to-many mapping within a single generation process. This entanglement leaves the intermediate editing strategy implicit, making it difficult to explicitly ground strategy selection in affective semantics. Second, existing optimization methods commonly apply a shared preference

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.

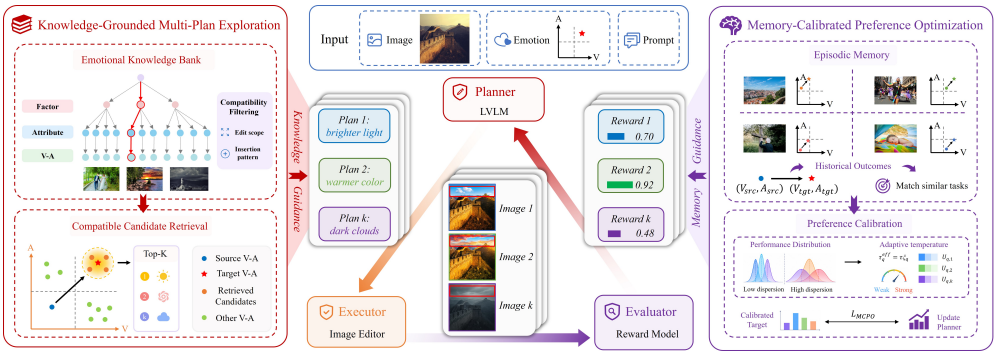

Figure 1: Overview of AffectAgent. Planner generates knowledge-grounded plans, Executor realizes them, Evaluator measures affective alignment and content consistency. KGME determines what strategies should be compared; MCPO determines how strongly they should be preferred.

scale across editing tasks. This overlooks variation in optimization characteristics across affective transitions with different directions and magnitudes: a performance level that is mediocre for one transition may already be highly competitive for another. Such variation makes preference strength difficult to interpret consistently across tasks.

To address these limitations, we introduce **AffectAgent**, a planner–executor–evaluator framework that combines *grounded planning* with *memory calibration* for controllable affective image editing. Figure 1 illustrates the overall framework. Given a source image and a target valence–arousal state, AffectAgent explores multiple editing strategies and evaluates the resulting images in terms of affective alignment and content preservation, providing preference supervision for Planner optimization. Specifically, we propose **Knowledge-Grounded Multi-Plan Exploration** (KGME), which leverages structured affective knowledge to explicitly construct source-adaptive editing plans. By exposing multiple candidate strategies, KGME provides semantically grounded plans whose effectiveness can be systematically assessed for Planner optimization. We further introduce **Memory-Calibrated Preference Optimization** (MCPO), which organizes historical editing experiences into an episodic memory and calibrates each training task against the performance distribution of tasks with similar affective transitions. This calibration enables the Planner to capture both the relative ordering of candidate strategies and the strength of their preferences, yielding task-aware optimization across varying affective transitions. Our main contributions are summarized as follows:

1) We present **AffectAgent**, a planner–executor–evaluator framework that combines *grounded planning* with *memory calibration* for controllable affective editing. Unlike end-to-end editing and task-agnostic learning, our method enables explicit multi-plan exploration and task-aware optimization.

2) We introduce Knowledge-Grounded Multi-Plan Exploration to construct semantically grounded editing plans, together with Memory-Calibrated Preference Optimization to adapt learning strength using the performance distribution of similar tasks.

3) We conduct extensive experiments on EmoSet, EMOTIC, and FindingEmo datasets. The results demonstrate that AffectAgent consistently reduces valence and arousal errors over state-of-the-art methods while maintaining competitive semantic consistency.

2 Related Work

Affective Image Editing. Affective image editing aims to modify visual content toward a target affect while preserving the source semantics. Early approaches mainly relied on global appearance manipulation, such as color and style adjustment (Lee & Park, 2011; Zhang et al., 2019). EmoEdit extends this paradigm by introducing an emotion adapter to support semantic affective modifications (Yang et al., 2025). EmoEditor further combines global emotion conditioning with local emotion-related cues in a diffusion framework (Lin et al., 2025). However, these methods primarily condition the editor on a target affect, leaving the corresponding visual strategy implicit in generation. In contrast, our method formulates affective image editing as an explicit strategy-selection problem, enabling affect-grounded candidate strategies to be directly compared and optimized.

General Image Editing. General-purpose image editing has advanced rapidly in both controllability and generation fidelity. InstructPix2Pix and MagicBrush established scalable instruction-following paradigms (Brooks et al., 2023; Zhang et al., 2023), while subsequent methods improved spatial and semantic control (Hertz et al., 2022; Couairon et al., 2022). Recent editors such as FLUX.2 (Labs, 2025) and Qwen-Image (Wu et al., 2025) have substantially advanced semantic controllability and content preservation under complex editing instructions. Multimodal large language models have also been incorporated into image editing pipelines to refine instructions, align visual content with editing intent, and coordinate visual tools (Fu et al., 2024; Huang et al., 2024; Wu et al., 2023). However, these methods mainly operate on explicitly specified editing intents, whereas affective targets are inherently abstract. AffectAgent addresses this challenge by retrieving structured affective knowledge to construct and optimize semantically grounded editing strategies.

Preference Optimization. Preference optimization aligns generative policies using relative feedback. DPO learns from pairwise preferences (Rafailov et al., 2023), while group-relative methods estimate advantages from multiple candidates associated with the same input (Shao et al., 2024; Ahmadian et al., 2024). Related objectives have also been extended to visual generation and editing through reward modeling and diffusion-specific preference optimization (Xu et al., 2023; Black et al., 2024; Prabhudesai et al., 2023; Wallace et al., 2024; Zhang et al., 2024). However, existing approaches typically interpret preference signals using shared preference scales, without accounting for how candidate quality relates to the performance distribution of similar tasks. AffectAgent instead introduces episodic memory to calibrate each task against historical experiences with similar affective transitions, enabling task-aware preference strength for Planner optimization.

3 METHOD

3.1 PROBLEM FORMULATION AND OVERVIEW

Given a source image x_q with source affect (V_q^s, A_q^s) and a target affect (V_q^t, A_q^t), our goal is to generate an edited image that matches the target affect while preserving the core semantics of the source image. To enable grounded strategy exploration and experience-aware learning, we propose **AffectAgent**, a planner–executor–evaluator framework for controllable affective image editing.

For each training task q, the planner π_θ retrieves grounding cues from an external **Emotional Knowledge Bank** $\mathcal{B}$ and generates K candidate editing plans $\{y_{q,i}\}_{i=1}^K$. Conditioned on each plan $y_{q,i}$, the executor $\mathcal{G}$ realizes the corresponding visual transformation, producing an edited image $\hat{x}_{q,i}$. The evaluator $\mathcal{E}$ provides multidimensional feedback using two in-house models for affect alignment and semantic consistency. Both are used only for training-time preference construction and are independent of the models and metrics used for final evaluation. The details of our in-house evaluators are provided in Appendix B.2. The resulting scores and task experiences are further accumulated into an **Episodic Memory** $\mathcal{M}$ to calibrate the optimization signal.

3.2 KNOWLEDGE-GROUNDED MULTI-PLAN EXPLORATION

End-to-end image editors implicitly couple affective reasoning with image generation, making intermediate strategy selection difficult to ground in affective semantics. To address this issue, we introduce **Knowledge-Grounded Multi-Plan Exploration** (KGME), which grounds the Planner with knowledge retrieved from an external Emotional Knowledge Bank (EKB) and constructs multiple grounded candidate strategies whose relative effectiveness can be evaluated for preference learning.

Emotional Knowledge Bank. The EKB serves as a structured interface between affective targets and editable visual semantics. It organizes affective knowledge in a hierarchical **element–attribute–VA** structure. An *element* denotes a visual semantic component that contributes to the affective impression of an image, while its associated *attributes* describe concrete visual manifestations in images. Each attribute is independently assigned a human-annotated valence–arousal (V–A) coordinate according to its perceived affective impression. In this way, an abstract target in the V–A space can be grounded into concrete visual cues that can be realized by an image editor.

We construct the EKB from the training split of the EmoSet dataset through three stages.

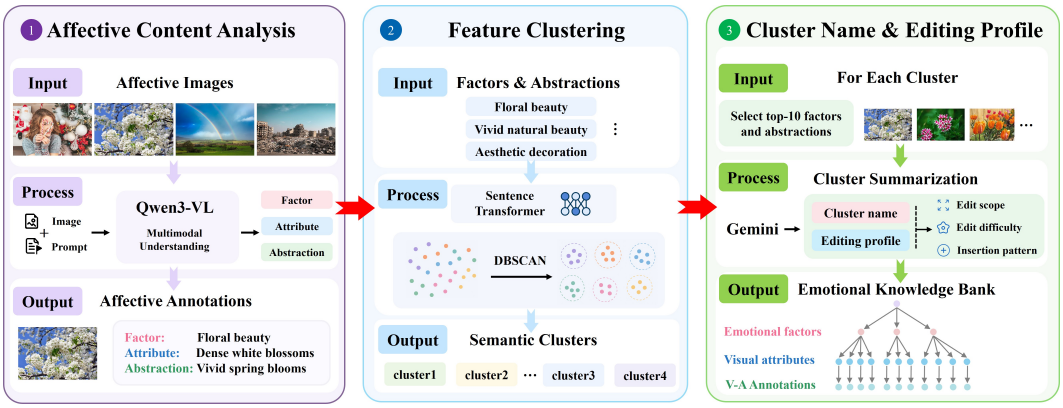

Figure 2: Construction of the EKB. Qwen3-VL extracts affective factors, attributes, and abstractions; Sentence Transformer and DBSCAN organize them into semantic clusters; and Gemini assigns cluster names and editing profiles, yielding a hierarchical bank with V-A annotations.

1) Affective Content Analysis. For each training image, we employ Qwen3-VL (Bai et al., 2025) to perform structured affective content analysis. The model describes visual elements, their corresponding visual attributes, and produces a holistic abstraction of the image-level affective semantics.

2) Feature Clustering. The extracted element descriptions are highly diverse and often semantically redundant. We therefore encode the elements together with their abstractions using Sentence Transformer (Reimers & Gurevych, 2019), and apply DBSCAN (Ester et al., 1996) to cluster the resulting representations. This step groups visually and affectively related elements into coherent clusters.

3) Cluster Naming and Editing Profile. For each cluster, we aggregate its most representative elements and abstractions, and use Gemini3 Pro (Team et al., 2023) to assign a concise element cluster name. We further associate each resulting element with an *editing profile* consisting of its **edit scope**, **edit difficulty**, and **insertion pattern**, which characterize where the edit takes effect, how difficult it is to realize, and how it modifies the image content, respectively.

The resulting EKB contains 95 affective elements and 73,262 visual attributes with V–A annotations, providing a structured pool of affect-relevant cues for editing strategy exploration.

Grounded Multi-Plan Construction. Given a training task q with source image x_q, the Planner first identifies the edit scopes $\mathcal{S}(x_q)$ and insertion patterns $\mathcal{P}(x_q)$ supported by the current image. We retain only EKB elements whose editing profiles are compatible with these source-image constraints, thereby excluding affective cues that are difficult to realize in the current scene.

For each retained element e_i, let its j-th visual attribute have an affective coordinate (v_{ij}, a_{ij}). We rank the corresponding cues by both target-affect relevance and editing difficulty:

$$S_{ij} = -\left\| \begin{bmatrix} v_{ij} \\ a_{ij} \end{bmatrix} - \begin{bmatrix} V_q^t \\ A_q^t \end{bmatrix} \right\|_1 - \lambda_d d_i, \tag{1}$$

where d_i denotes the editing difficulty of element e_i, with low, medium, and high difficulty mapped to 0, 0.5, and 1, respectively, and λ_d controls the difficulty penalty.

The top-K cues are then selected, and the Planner generates one editing plan for each cue by jointly considering the source image, target affect, and retrieved affective knowledge. In this way, KGME transforms the implicit affect-to-edit mapping into explicit, affect-grounded candidate strategies that can be systematically compared through their editing outcomes.

3.3 Memory-Calibrated Preference Optimization

Current preference optimization compares candidate quality under a shared scale, although the same reward gap can have different implications across affective transitions of varying directions and magnitudes. We introduce **Memory-Calibrated Preference Optimization** (MCPO), which builds a local performance reference from historical tasks with similar transitions and uses their outcome dispersion to derive a transition-conditioned temperature, adapting preference strength to each task.

Episodic Memory Matching. Each historical task is stored as an episode $e_j = (\Delta\mathbf{a}_j, \mathcal{Y}_j, \mathcal{R}_j)$, where $\mathcal{Y}_j = \{y_{j,i}\}_{i=1}^K$ contains the candidate plans, $\mathcal{R}_j = \{\mathbf{r}_{j,i}\}_{i=1}^K$ denotes their multidimensional evaluation outcomes, and $\Delta\mathbf{a}_j = [V_j^t - V_j^s,\ A_j^t - A_j^s]^\mathsf{T}$ describes the affective transition. All episodes form an episodic memory $\mathcal{M}$. For the current task q, we restrict the memory to episodes with the same transition direction and select the H tasks with the closest transition magnitude:

$$\mathcal{N}_q = \underset{\substack{e_j \in \mathcal{M},\, j \neq q \\ \mathrm{sgn}(\Delta\mathbf{a}_j) = \mathrm{sgn}(\Delta\mathbf{a}_q)}}{\arg\mathrm{topH}} - \left\| |\Delta\mathbf{a}_q| - |\Delta\mathbf{a}_j| \right\|_2 . \tag{2}$$

Memory-Conditioned Utility Calibration. The matched episodes $\mathcal{N}_q$ provide a local reference for interpreting the candidate outcomes of task q. For each evaluation dimension m, we estimate the mean $\mu_{q,m}$ and standard deviation $\sigma_{q,m}$ from the matched outcomes and standardize candidate i as

$$z_{q,i,m} = s_m \frac{r_{q,i,m} - \mu_{q,m}}{\sigma_{q,m} + \epsilon}, \tag{3}$$

where $s_m \in \{-1, +1\}$ aligns all metrics such that larger values indicate better performance. This standardization calibrates each metric according to its local variation: the same raw performance difference becomes more pronounced when similar historical tasks exhibit consistent outcomes and is attenuated when their outcomes are more dispersed. We then aggregate the standardized multidimensional signals into a memory-calibrated utility $U_{q,i}$. To account for dependencies among evaluation dimensions, we estimate the covariance matrix $\mathbf{C}_q$ over the corresponding standardized outcomes of the matched episodes. Given metric weights $\mathbf{w}$, we define

$$\xi_q = \sqrt{\mathbf{w}^\mathsf{T}(\mathbf{C}_q + \rho\mathbf{I})\mathbf{w}}, \qquad U_{q,i} = \frac{\mathbf{w}^\mathsf{T}\mathbf{z}_{q,i}}{\xi_q}, \tag{4}$$

where ρ is a covariance regularization coefficient. Here, $\sigma_{q,m}$ captures the metric-wise local dispersion, while ξ_q accounts for covariance-aware variation along the evaluation direction. Accordingly, $U_{q,i}$ provides a calibrated measure of candidate quality relative to historical tasks with similar affective transitions, accounting for both metric-wise dispersion and cross-metric dependence.

Task-Aware Preference Optimization. The calibrated utilities define a soft preference target over the K candidate plans. Substituting Eq. 4, we obtain

$$q_{q,i}^{\mathrm{target}} = \mathrm{softmax}_i\left(\frac{\mathbf{w}^\mathsf{T}\mathbf{z}_{q,i}}{\tau_q^{\mathrm{eff}}}\right), \qquad \tau_q^{\mathrm{eff}} = \tau\xi_q, \tag{5}$$

where τ is a global temperature and τ_q^{eff} is the transition-conditioned effective temperature. Since ξ_q rescales the global temperature, the sharpness of the target preference distribution adapts to the local outcome statistics of each task: a smaller calibrated scale produces a sharper distribution that more strongly separates candidate strategies, whereas a larger scale yields softer supervision. Together with the metric-wise standardization in Eq. 3, this allows MCPO to modulate preference strength according to the local performance distribution of similar affective transitions, rather than applying a shared preference scale uniformly across tasks.

Given a frozen reference Planner π_{ref}, we define the reference-relative score as

$$\ell_{q,i} = \log \pi_\theta(y_{q,i} \mid q) - \log \pi_{\mathrm{ref}}(y_{q,i} \mid q), \qquad p_{q,i}^{\mathrm{model}} = \mathrm{softmax}_i(\beta\ell_{q,i}), \tag{6}$$

where β controls the sharpness of the Planner-induced preference distribution. We optimize the Planner by aligning this distribution with the memory-calibrated target:

$$\mathcal{L}_{\mathrm{MCPO}} = -\frac{1}{B}\sum_{q=1}^{B}\sum_{i=1}^{K} q_{q,i}^{\mathrm{target}} \log p_{q,i}^{\mathrm{model}}. \tag{7}$$

Through this objective, MCPO preserves the relative ordering of candidate strategies while adapting preference strength to the locally calibrated performance distribution of similar historical tasks.

Training and Inference. During training, only the Planner is optimized, while the Executor and Evaluator remain fixed. At inference, neither multi-plan exploration nor episodic memory is required; the optimized Planner directly generates a single editing plan for the desired affect.

Table 1: Comparison with existing models **(Left)** and planner optimization strategies **(Right)** across three datasets. The best results are **bold**. Dataset rows left report the source-to-target V–A gaps.

(a) Comparison with Existing Models

Model	V-MAE ↓	A-MAE ↓	CLIP-I ↑	LPIPS ↓	Overall ↑
EmoSet	2.7841	2.1465	–	–	–
Qwen-Image	2.1876	2.0658	0.7015	0.5006	76.22
FLUX.2	2.7210	1.9928	**0.9054**	0.5533	75.62
EmoEdit	2.4705	1.7523	0.7988	0.3911	77.85
EmoEditor	2.4451	2.0087	0.8513	**0.1776**	79.30
Ours	**1.4603**	**1.3910**	0.7883	0.6522	**81.40**
EMOTIC	3.0559	2.1694	–	–	–
Qwen-Image	2.4799	2.0448	0.6150	0.4951	74.55
FLUX.2	3.0543	2.1100	**0.8758**	0.6325	73.17
EmoEdit	2.6555	2.0727	0.7872	0.3203	76.38
EmoEditor	2.7236	2.1030	0.8093	**0.1773**	77.54
Ours	**1.4581**	**1.1997**	0.7586	0.6584	**81.90**
FindingEmo	3.2596	2.0836	–	–	–
Qwen-Image	2.4949	2.0127	0.6141	0.5313	74.37
FLUX.2	3.0277	2.0269	**0.8811**	0.6145	73.74
EmoEdit	3.0102	1.9745	0.8352	0.2651	76.24
EmoEditor	2.8793	2.0017	0.7805	**0.1747**	77.14
Ours	**1.4763**	**1.4195**	0.7775	0.6708	**81.05**

(b) Comparison with Planner Optimization Strategies

Method	V-MAE ↓	A-MAE ↓	CLIP-I ↑	LPIPS ↓	Overall ↑
			EmoSet		
Base Planner	1.8786	1.7718	**0.8302**	0.6854	78.43
+ SFT	1.5603	1.6598	0.7809	0.7120	79.60
+ DPO	1.6692	1.6280	0.8213	0.6777	79.78
+ GRPO	1.4669	1.5625	0.7661	0.7395	80.09
+ MCPO	**1.4603**	**1.3910**	0.7883	**0.6522**	**81.40**
			EMOTIC		
Base Planner	1.8917	1.8093	0.7920	0.7204	77.77
+ SFT	1.6907	1.4661	0.7661	**0.6221**	80.22
+ DPO	1.9008	1.3463	**0.8110**	0.6257	80.18
+ GRPO	1.4715	1.6026	0.7566	0.7325	79.88
+ MCPO	**1.4581**	**1.1997**	0.7586	0.6584	**81.90**
			FindingEmo		
Base Planner	2.0727	1.7760	0.8089	0.6973	77.44
+ SFT	1.8587	1.5701	0.7847	**0.6387**	79.20
+ DPO	1.7579	1.4550	**0.8249**	0.6440	80.32
+ GRPO	1.6262	1.5871	0.7723	0.7196	79.52
+ MCPO	**1.4763**	**1.4195**	0.7775	0.6708	**81.05**

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Datasets. We train and evaluate our method on three affective image datasets with different scene distributions and emotional characteristics: EmoSet (Yang et al., 2023), EMOTIC (Kosti et al., 2017), and FindingEmo (Mertens et al., 2024). Each evaluation sample consists of a source image, its source valence–arousal state, and a target valence–arousal state. As the EKB is constructed only from the EmoSet training split, experiments on EMOTIC and FindingEmo further evaluate whether the affective knowledge transfers across datasets with different visual distributions.

Baseline and evaluation metrics. For comparison, we consider state-of-the-art general-purpose image editing models, including Qwen-Image (Wu et al., 2025) and FLUX.2 (Labs, 2025), as well as affective image editing methods EmoEdit (Yang et al., 2025) and EmoEditor (Lin et al., 2025). To ensure fairness, all baselines use the same train–test splits and source–target pairs as our method. We follow their official repositories and recommended configurations for implementation.

We assess performance using five metrics: V-MAE, A-MAE, CLIP-I (Radford et al., 2021), LPIPS (Zhang et al., 2018), and Overall Affective Editing Score (Overall). For affective accuracy, we employ the estimation model introduced by Mertens et al. (2024) to predict the valence and arousal of each edited image. V-MAE and A-MAE are computed as the absolute errors between the predicted and target V-A values. For semantic preservation, CLIP-I measures the cosine similarity between the source and edited image, while LPIPS measures their perceptual distance. Overall provides a unified measure of editing quality by aggregating affective accuracy and semantic consistency. All evaluation models and metrics are independent of the in-house evaluators during training.

4.2 MAIN RESULTS

Comparison with Existing Methods. Table 1 (Left) reports the quantitative comparison on three datasets. AffectAgent consistently achieves the lowest V-A Errors and the highest Overall scores across all datasets, indicating more accurate target-affect control. Although FLUX.2 and EmoEditor perform better on individual content-preservation metrics, their larger affective errors suggest more conservative editing. Overall, AffectAgent provides a stronger balance between affective alignment and content preservation. Figure 3 presents qualitative comparisons over diverse source–target transitions. AffectAgent produces visually salient changes toward the target affect while retaining the principal subject and scene structure. Figure 4 further evaluates continuous control by varying target valence and arousal over a two-dimensional grid. AffectAgent exhibits coherent variation across both valence and arousal dimensions, indicating controllability beyond isolated target states.

Comparison of Planner Optimization Strategies. To ensure a controlled comparison of planner optimization strategies, we keep the Executor, Evaluator, and training tasks fixed. SFT, DPO, and

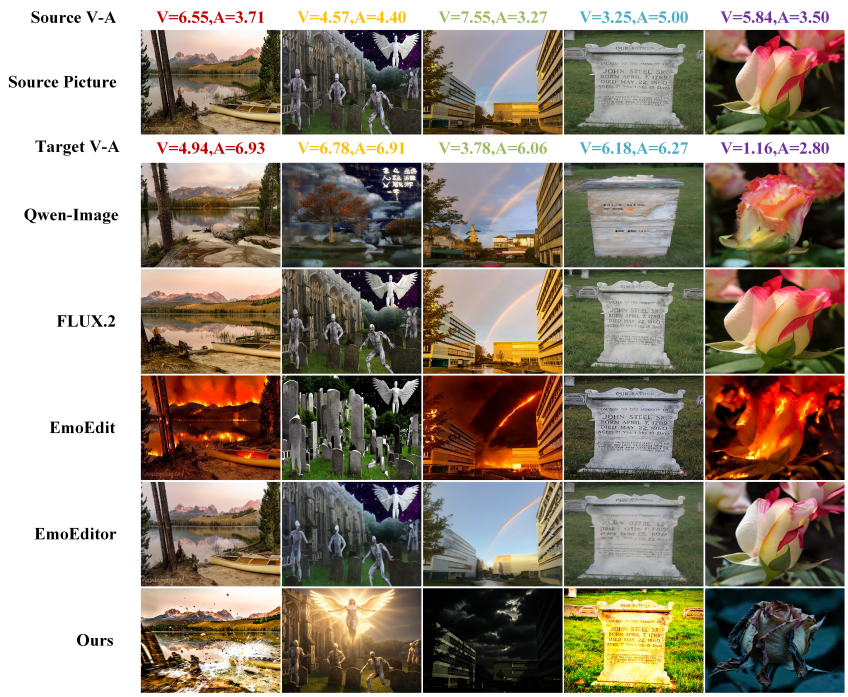

Figure 3: Qualitative comparison on diverse source–target affective transitions.

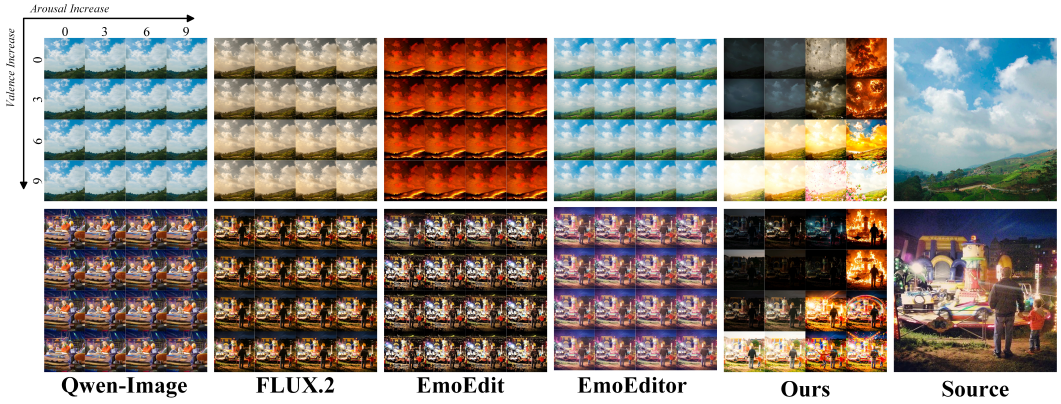

Figure 4: Continuous valence–arousal control across different methods.

MCPO operate on the same KGME-generated candidate plans, while GRPO follows its standard group-relative protocol and optimizes over plans sampled from the current policy. Table 1 (Right) compares the Base Planner with SFT, DPO, GRPO, and MCPO. MCPO achieves the lowest V-MAE and A-MAE and the highest Overall score across all three datasets. SFT and DPO obtain better CLIP-I or LPIPS scores in some cases, but their target-affect errors remain consistently higher. GRPO is inferior to MCPO across all reported metrics, indicating that group-relative optimization alone is insufficient to account for the heterogeneous preference scales across affective transitions. These results suggest that transition-conditioned calibration provides more effective supervision for planner optimization than standard supervised, pairwise, and group-relative preference learning. Additional qualitative comparisons are provided in Appendix A.

Planner Attention Analysis. To examine how planner optimization affects visual reasoning, we visualize the attention associated with the target-affect query in Figure 5. We extract attention from the penultimate language-decoder layer, using the target-affect span and its V-A values as queries and visual tokens as keys; the attention is then averaged over query tokens and heads and projected back to the image space. The attention patterns become more concentrated on affect-relevant regions

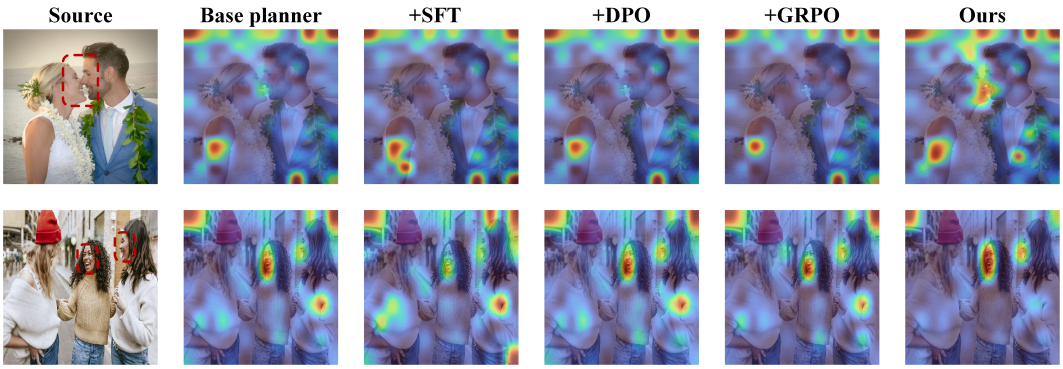

Figure 5: Planner attention visualization for target-affect queries.

Table 2: User study results.Values are reported as mean ± half-width of the 95% bootstrap confidence interval, obtained by resampling over tasks. The best results are highlighted in **bold**.

	Affect Alignment		Perceptual Quality	
Method	**Human V-MAE ↓**	**Human A-MAE ↓**	**Semantic Preservation ↑**	**Image Quality ↑**
Qwen-Image	1.98 ± 0.18	1.81 ± 0.15	4.12 ± 0.12	4.35 ± 0.08
FLUX.2	2.27 ± 0.22	1.74 ± 0.17	4.29 ± 0.10	4.38 ± 0.09
EmoEdit	2.11 ± 0.19	1.65 ± 0.17	4.10 ± 0.09	4.17 ± 0.10
EmoEditor	2.06 ± 0.20	1.70 ± 0.19	4.26 ± 0.08	4.08 ± 0.08
Ours	**1.26 ± 0.16**	**1.15 ± 0.14**	**4.32 ± 0.08**	**4.43 ± 0.07**

after optimization, with MCPO showing the clearest focus on facial expressions and interpersonal interactions. This is consistent with MCPO placing greater attention on affect-relevant visual cues.

User Study. We further conduct a human perceptual study to complement the model-based evaluation. We sample 60 source–target tasks, balanced across EmoSet, EMOTIC, and FindingEmo and covering different affective transition directions and magnitudes. For each task, we generate results from five compared methods. Thirty participants evaluate outputs under a blind and randomized protocol. A balanced assignment ensures six independent ratings per output, with each participant viewing only one method for each task. For each trial, participants are shown the source image, edited image, and target affect. They first rate the perceived V-A of the edited image, from which we compute human V-MAE and A-MAE against the target coordinates. Participants additionally rate semantic preservation and image quality on a five-point Likert scale. Semantic preservation measures whether the principal content and scene structure of the source image are retained, while image quality assesses visual naturalness, coherence, and artifacts. Model identities are hidden and trial order is randomized. As shown in Table 2, AffectAgent obtains the lowest human V-MAE and A-MAE, consistent with the automatic evaluation and confirming more accurate target-affect control from human perception. Meanwhile, its semantic-preservation and image-quality ratings are slightly higher than those of SOTA baselines, suggesting that human judgments capture aspects of semantic preservation that are not fully reflected by CLIP-I and LPIPS. Overall, the improved affective alignment does not come at the expense of perceptual quality or content preservation.

4.3 Ablation Study

Effectiveness of KGME. We assess KGME in exploring affect-grounded strategies. As shown in Figure 6 (Up), we compare Base Planner, Random Retrieval, and KGME, with the latter two optimized under the same MCPO objective. Random Retrieval already improves affective accuracy over the Base Planner, suggesting that external editing cues can facilitate strategy construction. KGME further reduces V-A Error and achieves the highest Overall score, indicating that affective grounding offers additional gains beyond knowledge retrieval alone. These results support the role of KGME in grounding the implicit affect-to-edit mapping into more effective candidate strategies.

Effectiveness of MCPO. Figure 6 (Down) compares MCPO with Groupwise, Dataset-Wide, Random-Memory, and Direction-Only calibration across increasing affective transition magnitudes. These baselines use within-task comparisons, global statistics, randomly sampled experiences, or

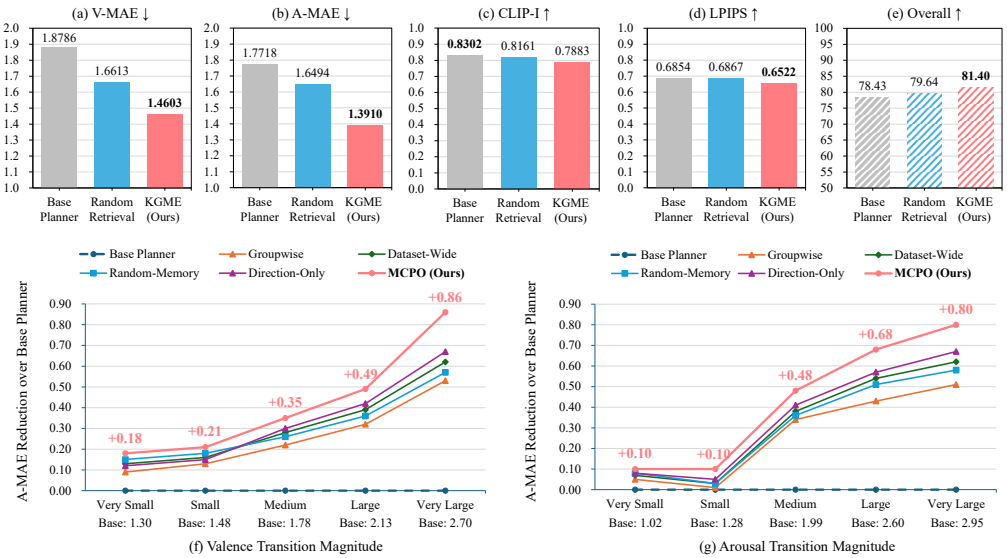

Figure 6: Ablation studies of KGME and MCPO on EmoSet. **Up** (a–e): Affective knowledge grounding improves target-affect accuracy while maintaining semantic consistency. **Down** (f–g): Memory-conditioned calibration yields larger gains as the affective transition magnitude increases.

Table 3: Component ablation of EKB on EmoSet. The best results are highlighted in **bold**.

Configuration	V-MAE ↓	A-MAE ↓	CLIP-I ↑	LPIPS ↓	Overall ↑
w/o Compatibility Filtering	1.5827	1.5239	0.7562	0.7029	79.90
w/o Difficulty Penalty	1.5545	1.4313	0.7714	0.6738	80.64
w/o Element Level	1.6283	1.5890	0.7937	0.6595	79.98
w/o Attribute Level	1.7118	1.6325	**0.8036**	**0.6438**	79.65
Full EKB (Ours)	**1.4603**	**1.3910**	0.7883	0.6522	**81.40**

direction-matched experiences, respectively. MCPO consistently achieves the lowest V-MAE and A-MAE, and its advantage widens as the transition magnitude increases. This suggests that task-aware calibration becomes increasingly important for more challenging affective transitions. The gains over Random-Memory and Direction-Only further show that the benefit comes not merely from using historical experiences, but from leveraging references from more closely matched tasks.

Component Analysis of EKB. Table 3 isolates the contribution of individual components in the EKB. Removing the difficulty penalty causes a moderate performance drop, indicating that editing difficulty helps prioritize realizable strategies. Excluding compatibility filtering leads to a large degradation in content preservation, suggesting that incompatible elements introduce edits that substantially alter the source semantics. Removing either level of the hierarchical EKB further degrades performance, with the largest affective error observed without attribute-level cues. This highlights the importance of fine-grained attributes, which enable more precise alignment with the target affect. With all components enabled, the full EKB achieves the best affective accuracy and Overall score, supporting the roles of compatibility filtering, hierarchical structure, and difficulty penalty.

5 Conclusion

We presented AffectAgent, a planner–executor–evaluator framework for controllable affective image editing. KGME grounds multi-plan exploration in structured affective knowledge, while MCPO calibrates reward gaps using the outcome distribution of historical tasks with similar valence–arousal transitions. Across EmoSet, EMOTIC, and FindingEmo, AffectAgent achieves the lowest V-MAE and A-MAE and the highest Overall score, while retaining competitive semantic consistency. These results prove the value of combining grounded strategy exploration with transition-aware preference optimization for accurate affective control. Future work will investigate richer human preference signals, more expressive affective representations, and extensions to video and interactive editing.

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

Arash Ahmadian, Chris Cremer, Matthias Gallé, Marzieh Fadaee, Julia Kreutzer, Olivier Pietquin, Ahmet Üstün, and Sara Hooker. Back to basics: Revisiting reinforce-style optimization for learning from human feedback in llms. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12248–12267, 2024.

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.

Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. In *International Conference on Learning Representations*, volume 2024, pp. 4965–4987, 2024.

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

Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Diffedit: Diffusion-based semantic image editing with mask guidance. *arXiv preprint arXiv:2210.11427*, 2022.

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.

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

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control.(2022). *URL https://arxiv.org/abs/2208.01626*, 3(3), 2022.

Yuzhou Huang, Liangbin Xie, Xintao Wang, Ziyang Yuan, Xiaodong Cun, Yixiao Ge, Jiantao Zhou, Chao Dong, Rui Huang, Ruimao Zhang, et al. Smartedit: Exploring complex instruction-based image editing with multimodal large language models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8362–8371. IEEE, 2024.

Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6007–6017. IEEE, 2023.

Ronak Kosti, Jose M Alvarez, Adria Recasens, and Agata Lapedriza. Emotion recognition in context. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 1960–1968. IEEE, 2017.

Black Forest Labs. FLUX.2: Frontier Visual Intelligence. `https://bfl.ai/blog/flux-2`, 2025.

Joonwhoan Lee and EunJong Park. Fuzzy similarity-based emotional classification of color images. *IEEE Transactions on Multimedia*, 13(5):1031–1039, 2011.

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

Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-grpo: Training flow matching models via online rl. *Advances in neural information processing systems*, 38:40783–40818, 2026.

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.

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

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

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

Mihir Prabhudesai, Anirudh Goyal, Deepak Pathak, and Katerina Fragkiadaki. Aligning text-to-image diffusion models with reward backpropagation. 2023.

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

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

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.

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.

Shelly Sheynin, Adam Polyak, Uriel Singer, Yuval Kirstain, Amit Zohar, Oron Ashual, Devi Parikh, and Yaniv Taigman. Emu edit: Precise image editing via recognition and generation tasks. In *2024 ieee/cvf conference on computer vision and pattern recognition (cvpr)*, pp. 8871–8879. IEEE, 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.

Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8228–8238. IEEE, 2024.

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, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. Visual chatgpt: Talking, drawing and editing with visual foundation models. *arXiv preprint arXiv:2303.04671*, 2023.

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

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

Zitong Xu, Huiyu Duan, Zhongpeng Ji, Xinyun Zhang, Yutao Liu, Xiongkuo Min, Ke Gu, Jian Zhang, Shusong Xu, Jinwei Chen, et al. Edithf-1m: A million-scale rich human preference feedback for image editing. In *European Conference on Computer Vision*, pp. 278–298. Springer, 2026.

Jingyuan Yang, Xinbo Gao, Leida Li, Xiumei Wang, and Jinshan Ding. Solver: Scene-object interrelated visual emotion reasoning network. *IEEE Transactions on Image Processing*, 30:8686–8701, 2021.

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

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

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

Kai Yang, Jian Tao, Jiafei Lyu, Chunjiang Ge, Jiaxin Chen, Weihan Shen, Xiaolong Zhu, and Xiu Li. Using human feedback to fine-tune diffusion models without any reward model. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8941–8951. IEEE, 2024b.

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

Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. Magicbrush: A manually annotated dataset for instruction-guided image editing. *Advances in neural information processing systems*, 36:31428–31449, 2023.

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

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

Shu Zhang, Xinyi Yang, Yihao Feng, Can Qin, Chia-Chih Chen, Ning Yu, Zeyuan Chen, Huan Wang, Silvio Savarese, Stefano Ermon, et al. Hive: Harnessing human feedback for instructional visual editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9026–9036, 2024.

Wei Zhang, Xuanyu He, and Weizhi Lu. Exploring discriminative representations for image emotion recognition with cnns. *IEEE Transactions on Multimedia*, 22(2):515–523, 2019.

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.

APPENDIX

A QUALITATIVE COMPARISON OF PLANNER OPTIMIZATION STRATEGIES

As shown in Figures 7 and 8, we provide qualitative comparisons for both isolated source–target transitions and continuous valence–arousal control. The same source images and target affective states are used across the compared planner variants.

B FURTHER DETAILS ON THE EXPERIMENT SETUP

B.1 ATTRIBUTE-LEVEL V–A ANNOTATION PROTOCOL

The V–A coordinates of the visual attributes in the EKB were obtained through human annotation. We recruited 50 annotators and provided them with a brief instruction on the valence–arousal representation before annotation. Each attribute was independently rated by 5 annotators on a 0–9 scale for both valence and arousal, based on the affective impression conveyed by the attribute itself, without access to its original source image. A small subset of attributes was repeated to assess intra-annotator consistency, and incomplete or unreliable responses were excluded before aggregation. The final V–A coordinate was obtained by averaging the valid ratings.

B.2 TRAINING-TIME IN-HOUSE EVALUATORS

During preference construction, we employ two fixed in-house evaluators for affective alignment and semantic consistency. Both evaluators are trained independently of the Planner and remain frozen throughout preference optimization. They are used only to construct training-time supervision and are not involved in the final evaluation.

Affect Evaluator. The affect evaluator uses a CLIP image encoder followed by a regression head to estimate continuous valence and arousal scores. It is trained on a disjoint subset of the EmoSet training split, with no overlap with the samples used for AffectAgent training or final evaluation. The model outputs a two-dimensional V–A estimate for each edited image.

Semantic Evaluator. The semantic evaluator is built on Qwen3-VL and fine-tuned using the attribute-preservation annotations from EditHF-1M (Xu et al., 2026), a large-scale human-feedback dataset for image editing. It takes the source image, edited image, and editing instruction as input and outputs a scalar score measuring the preservation of content and semantics.

B.3 MODELS AND BASELINE METHODS

Table 4 lists the models and affective image-editing baselines used in our experiments, together with their publicly available links.

Table 4: Models and baseline methods used in our experiments.

Model	**Link**
Qwen3-VL-8B-Instruct	https://huggingface.co/Qwen/Qwen3-VL-8B-Instruct
Qwen-Image	https://huggingface.co/Qwen/Qwen-Image
FLUX.2	https://huggingface.co/black-forest-labs/FLUX.2-klein-9B
EmoEdit	https://github.com/JingyuanYY/EmoEdit
EmoEditor	https://github.com/ZhangLab-DeepNeuroCogLab/EmoEditor

B.4 DEFINITION OF THE OVERALL METRIC

We define Overall as a weighted aggregation of affective accuracy and semantic consistency:

$$\text{Overall} = 100\,[(1-\lambda)S_{\text{sem}} + \lambda S_{\text{aff}}]\,, \tag{8}$$

where

$$S_{\text{sem}} = \frac{1}{2}\left(\frac{\text{CLIP-I}+1}{2} + \frac{1}{1+\text{LPIPS}}\right),$$
$$S_{\text{aff}} = 1 - \frac{\text{V-MAE}+\text{A-MAE}}{18}. \tag{9}$$

Here, S_{sem} and S_{aff} denote normalized semantic consistency and affective accuracy. We set $\lambda = 0.7$, assigning greater weight to affective accuracy because the primary objective of affective image editing is to reach the target affect while preserving source semantics as a complementary constraint. The value of λ is fixed across all methods and datasets and is not tuned on the test results.

B.5 Hyperparameters Specification

For reproducibility, AffectAgent is initialized from Qwen3-VL-8B-Instruct and trained with LoRA on two GPUs using distributed training. The planner generates four candidate plans per task. The complete training configuration is summarized in Table 5.

Table 5: Training hyperparameters of AffectAgent.

Category	**Setting**
Base model	Qwen3-VL-8B-Instruct
Training GPUs	2 GPUs (distributed training)
Candidate plans (K)	4
Preference coefficient (β)	0.05
Preference temperature (τ)	0.5
Calibration neighbors	100
Neighbor-search chunk size	512
Covariance ridge	10^{-6}
Calibration ϵ	10^{-6}
LoRA rank / α	64 / 128
Learning rate	10^{-5}
Training epochs	4
Per-device batch size	1
Gradient accumulation steps	8
Maximum sequence length	102,400
Maximum image pixels	1,048,576

B.6 Computational Efficiency

We evaluate the practical computational cost of each method using per-image inference time and peak GPU memory. All methods were evaluated on a single NVIDIA RTX PRO 6000 Blackwell GPU (96 GiB) with NVIDIA driver 580.95.05 and CUDA 13.0. Qwen-Image used Python 3.12.13 and PyTorch 2.13.0; FLUX.2 and AffectAgent used Python 3.12.13 and PyTorch 2.12.1; and EmoEdit and EmoEditor used Python 3.10.20 and PyTorch 2.13.0. All PyTorch environments were built with CUDA 13.0. Table 6 reports the results.

C Qualitative Results of the Ablation Study

We provide additional qualitative comparisons to complement the quantitative ablations in Figure 6. We examine both discrete source-to-target transitions and continuous control over the V-A space.

C.1 Effectiveness of KGME

Figures 9 and 10 compare target-aware knowledge grounding with the Base Planner and Random Retrieval. Random Retrieval uses the same MCPO objective but replaces target-aware retrieval with randomly sampled affective knowledge. The results show that KGME produces more transition-relevant and coherent edits across both discrete and continuous affective targets, supporting the importance of grounding rather than retrieval alone.

Table 6: Computational efficiency on a single NVIDIA RTX PRO 6000 Blackwell GPU.

Metric	Qwen-Image	FLUX.2	EmoEdit	EmoEditor	Ours
Inference time (s) ↓	11.082	2.454	2.558	0.655	3.658
Peak GPU memory (GiB) ↓	57.96	33.60	5.29	2.33	33.60

C.2 Effectiveness of MCPO

Figures 11 and 12 compare MCPO with alternative planner optimization and reward-calibration strategies. MCPO responds more consistently to the specified affective transitions across both evaluation settings, agreeing with its lower V-MAE and A-MAE and higher Overall score in Figure 6.

D Planner Prompts

D.1 Prompt for Plan Construction

Table 7 presents the prompt used by KGME to construct the candidate editing plans. For each retrieved factor–exemplar pair, the planner is provided with the corresponding affective factor and exemplar attributes. The placeholders are dynamically populated for each source–target pair and retrieved knowledge item.

Table 7: Prompt used by KGME for candidate-plan construction. The values enclosed by angle brackets are dynamically populated.

Prompt type	Prompt content
System prompt	You are an expert planner for emotional image editing. Your task is to write ONE natural, concise, and executable English editing instruction for an image editing model. You will be given: (1) the source image; (2) a retrieved emotional factor; (3) retrieved exemplar attributes; and (4) the source and target emotion positions in valence–arousal space from 0 to 9. Your job: (1) understand the source image content from the image itself; (2) use the retrieved factor and exemplar attributes as references, not as rigid text to copy; (3) produce an instruction that is semantically plausible for the source image; (4) preserve the main identity, scene structure, and layout unless emotional editing requires mild local change; (5) do NOT mention valence, arousal, numbers, coordinates, or analysis in the final instruction; (6) do NOT output JSON, bullets, explanation, or multiple options; and (7) output only a single English instruction sentence or short paragraph. The instruction should be suitable as input to an image editing model.
User prompt	<SOURCE_IMAGE> Retrieved emotional factor: <FACTOR_NAME> Retrieved exemplar attributes: <EXEMPLAR_ATTRIBUTES> Source emotion: V $= \langle V_s \rangle$, A $= \langle A_s \rangle$ Target emotion: V $= \langle V_t \rangle$, A $= \langle A_t \rangle$ Desired relative change: Valence should ⟨move higher/move lower/stay close⟩ relative to the source image; Arousal should ⟨move higher/move lower/stay close⟩ relative to the source image. Please generate the final image editing instruction.

D.2 Prompt for Evaluation

At evaluation time, the trained planner receives only the source image and the source and target valence–arousal values. As shown in Table 8, retrieved knowledge and continuous controls are omitted, allowing the evaluation to measure whether the optimized planner has internalized grounded affective editing strategies.

Table 8: Prompt used by the optimized instruction planner during evaluation. The values enclosed by angle brackets are dynamically populated for each source–target affective pair.

Prompt type	**Prompt content**
System prompt	You are an expert planner for emotional image editing. Your task is to write ONE natural, concise, and executable English editing instruction for an image editing model. You will be given: (1) the source image; and (2) the source and target emotion positions in valence–arousal space from 0 to 9. Your job is to: (1) understand the source image content; (2) infer visually plausible changes that move the perceived emotion from the source position toward the target position; (3) preserve the main identity, objects, scene structure, and layout unless affective editing requires mild local changes; (4) express the intended emotional change through concrete and executable visual attributes, such as lighting, color, atmosphere, facial expression, weather, or local scene details, when appropriate; (5) avoid mentioning valence, arousal, numerical values, coordinates, or analysis in the final instruction; (6) output no JSON, bullets, explanations, or multiple alternatives; and (7) output only one English editing instruction as a single sentence or short paragraph. The instruction should be suitable as input to an image editing model.
User prompt	<SOURCE_IMAGE> Source emotion: V $= \langle V_s \rangle$, A $= \langle A_s \rangle$ Target emotion: V $= \langle V_t \rangle$, A $= \langle A_t \rangle$ Desired relative change: Valence should ⟨increase/decrease/remain close⟩ relative to the source image; Arousal should ⟨increase/decrease/remain close⟩ relative to the source image. Please generate the final image editing instruction.