EMOINSIGHT: LEARNING FINE-GRAINED VISUAL AFFECT FROM HOMOLOGOUS VARIATIONS AND PREFERENCE-GUIDED SEMANTIC SUPERVISION

Anonymous authors
Paper under double-blind review

ABSTRACT

Visual Emotion Assessment (VEA) aims to model human affective perception elicited by visual content. In real-world scenarios, emotion perception is inherently subjective, arises from reasoning over multiple interacting visual cues, and can vary substantially with subtle changes in visual context. However, existing VEA methods largely treat emotion assessment as numerical scoring. This limits their capacity for deep semantic reasoning over diverse cues and capturing fine-grained visual-affective correlations. To bridge this gap, we introduce a new benchmark dataset (**ContiEmo**) constructed from homologous image groups with densely annotated affective variations and human-annotated multi-cue reasoning labels. Building upon ContiEmo, we develop **EmoInsight**, a preference-guided VEA model. Specifically, a *Multi-Teacher Synergistic Affective Reasoning* mechanism is proposed to enhance semantic reasoning capabilities by transferring diverse affective knowledge with instance-specific dynamic weighting derived from human preferences. To further capture fine-grained visual-affective correlations, we introduce a *Coarse-to-Fine Factor-Guided Reinforcement* strategy that jointly optimizes V-A estimation and consistency with affective factor cues. Extensive experiments on ContiEmo and public real-world VEA benchmarks demonstrate that our method consistently improves fine-grained affect estimation and multi-factor recognition.

1 INTRODUCTION

Emotions fundamentally shape human perception, modulating the way we interpret and interact with the world (Vega, 1996; Tyng et al., 2017; Li et al., 2020a). Among diverse emotion carriers, visual modality stands out for its intuitiveness and semantic richness. (Chainay et al., 2012; Yang et al., 2018). Visual Emotion Assessment (VEA) studies the alignment between visual content and human emotions, holding great potential for computer vision tasks such as aesthetic assessment (Li et al., 2020b) and stylized captioning, (Guo et al., 2019) as well as human-centric applications including opinion mining (Li et al., 2019) and mental monitoring (Wieser et al., 2012). Recently, VEA has emerged as a critical dimension of image understanding, attracting considerable research attention.

Early studies in VEA primarily employed handcrafted features grounded in artistic and psychological theories (Machajdik & Hanbury, 2010; Zhao et al., 2014), which exhibit limited capacity to capture the full complexity of human affect. Although deep learning techniques have markedly improved emotion evaluation performance (Xu et al., 2022), they still encounter critical challenges regarding human-aligned perception: **(1) Deficiency in Reasoning:** Current models lack the capacity for deep semantic reasoning over diverse emotional cues, constraining their ability to interpret the inherent subjectivity and multi-cue nature of affective content. **(2) Insufficient Fine-Grained Visual-Affective Supervision:** Existing approaches optimize solely based on independently collected images, providing limited supervision for learning how subtle visual changes correspond to fine-grained emotional variations. Compounding these algorithmic limitations is **a fundamental deficit in data resources**: existing VEA datasets predominantly comprise isolated emotional images (Yang et al., 2023; Kosti et al., 2017a), devoid of human reasoning annotations and homologous samples that preserve semantic content while varying affective appearance. Consequently, models trained on such data struggle to align effectively with human emotional understanding.

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.

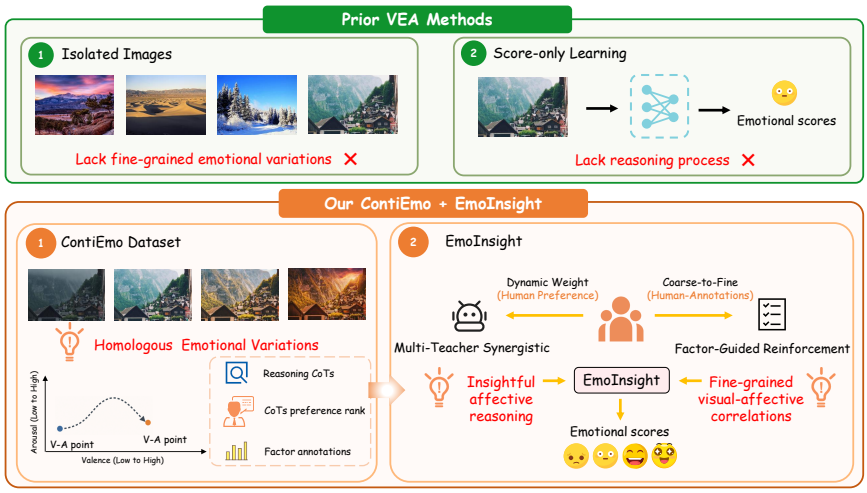


Figure 1: Distinctions between EmoInsight and prior methods.

To address these challenges, we propose ContiEmo, a large-scale VEA dataset featuring rich reasoning annotations and densely sampled homologous affective variations. Building upon this, we present EmoInsight, a novel evaluation model designed to improve semantic reasoning and fine-grained emotion assessment. Figure 1 illustrates the key distinctions between EmoInsight and prior methods. Specifically, we employ Qwen3-VL-4B as the backbone architecture. To elevate the model's reasoning ability, we propose a Multi-Teacher Synergistic Affective Reasoning mechanism. This mechanism enables our compact vision-language model (VLM) to assimilate emotional insights from multiple state-of-the-art large VLMs (LVLMs), thereby acquiring high-performance Chain-of-Thought (CoT) reasoning while maintaining computational efficiency. These multi-teacher CoTs are synergistically distilled via a preference-guided dynamic weighting mechanism, allowing the student to learn from diverse affective interpretations. Furthermore, to capture fine-grained visual-affective associations, we develop a Coarse-to-Fine Factor-Guided Reinforcement strategy. The hierarchical factors encourage the model to look beyond mere evaluation accuracy and instead focus on pivotal visual cues associated with emotional variations. To construct ContiEmo, we utilized image editing and first-and-last frame generation techniques to create densely sampled homologous transition sequences. Compared to existing datasets, ContiEmo presents distinct advantages: large-scale human annotations, densely distributed emotional values, and diverse reasoning labels, including CoT texts, factor importance rankings and detailed factor content. In summary, the contributions of this work are as follows:

- To enhance semantic reasoning ability, we propose a *Multi-Teacher Synergistic Affective Reasoning* mechanism that transfers diverse affective semantic knowledge into a compact VLM through preference-guided dynamic weighting, enabling the model to integrate multiple affective cues.
- To capture fine-grained visual-affective correlations, we develop a *Coarse-to-Fine Factor-Guided Reinforcement* strategy that jointly optimizes V-A estimation and consistency with coarse-to-fine affective factor supervision, thereby enhancing the sensitivity to subtle affective variations.
- We introduce **ContiEmo**, a large-scale VEA dataset built from homologous image groups with densely annotated affective variations and rich reasoning supervision. Extensive experiments on ContiEmo, EMOTIC, and FindingEmo demonstrate the effectiveness and generalization of the proposed framework in fine-grained visual affect assessment and multi-factor recognition.

2 Related Work

Visual Emotion Assessment Method. Visual Emotion Assessment (VEA) aims to evaluate the emotions conveyed or evoked by images, effectively bridging the gap between visual content and subjective human perception. Early endeavors in VEA primarily focused on classifying discrete emotional categories (Borth et al., 2013; Yang et al., 2020; Xu et al., 2022). However, given the complexity of human affect, these category-based approaches often fail to fully capture the emotional diversity (Gunes et al., 2011). Consequently, research interest has shifted towards score-based assessment methods (Kosti et al., 2017b; Toisoul et al., 2021; Kollias, 2022; Lian et al., 2025). For

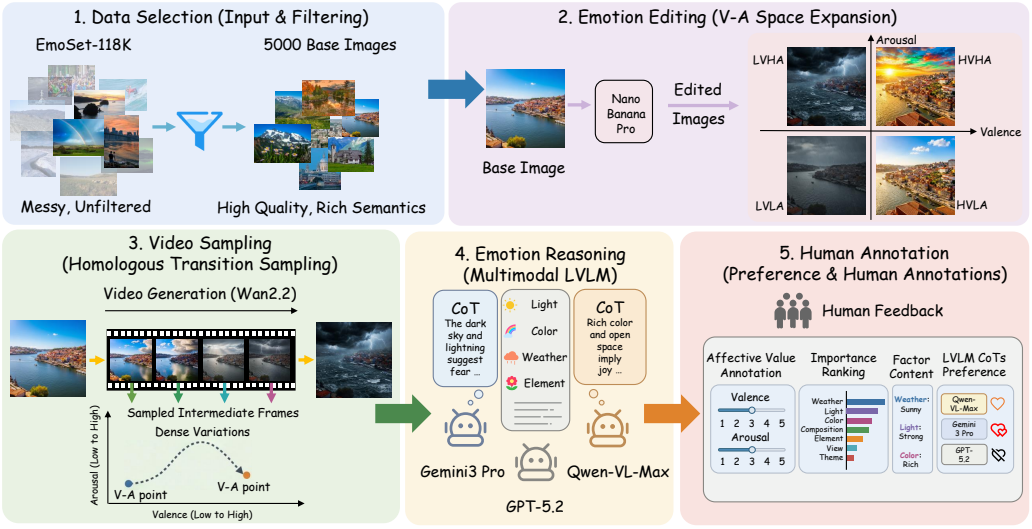


Figure 2: The construction pipeline of ContiEmo dataset.

instance, FindingEmo (Mertens et al., 2024) evaluated multiple backbones for valence and arousal score prediction, demonstrating the intrinsic correlation between affective scores and visual cues across diverse structures. Beyond score-based approaches, another prominent paradigm is distribution-based VEA, which maps images to a probability distribution of emotions (Xiong et al., 2019; Wang & Geng, 2021; Jing et al., 2023; Xu et al., 2025). For instance, StyleEDL (Jing et al., 2023) present a style-guided high-order attention network designed to model the intensity distributions and intrinsic correlations among multiple emotions. To further mitigate emotion-irrelevant redundancy, MFRN (Xu et al., 2025) introduced the Multi-Feature Refinement Network, enhancing the model's sensitivity to subtle variations in co-existing emotions within the distribution.

Visual Emotion Assessment Dataset. To explore human affective responses to visual stimuli, a variety of large-scale datasets have been established (Mollahosseini et al., 2017; Peng et al., 2015; You et al., 2016). According to the paradigm of VEA, these benchmarks can be categorized into category-based, distribution-based, and score-based datasets. For category-based datasets, EmoSet (Yang et al., 2023) comprises 118,102 images annotated with discrete emotion labels and a rich set of attributes. Leveraging these cues, an attribute-aware fusion model was developed to recognize emotions. To address the inherent ambiguity of emotions, distribution-based datasets like Flickr-LDL and Twitter-LDL were introduced (Yang et al., 2017). These datasets aggregate annotations from multiple viewers to form a label distribution. Beyond discrete labels, score-based datasets provide a more granular measurement of affect (Zhang et al., 2025). For instance, EMOTIC (Kosti et al., 2017a) combines 26 emotion categories with valence, arousal, and dominance scores for over 23,000 images, focusing primarily on the emotional states of people within the images rather than the overall scene. In contrast, FindingEmo (Mertens et al., 2024) targets complex natural scenes by annotating emotions of the entire image across valence and arousal, offering a holistic perspective on visual sentiment.

3 Method

3.1 ContiEmo Dataset Construction

ContiEmo is a large-scale VEA dataset designed to provide dense supervision over fine-grained visual-affective variations. It comprises 105K images, each accompanied by continuous affective scores, reasoning texts explaining the emotional content, preference rankings, and a suite of hierarchical factor annotations. Following the emotional definitions in EMOTIC and FindingEmo datasets, we employ the Valence-Arousal (V-A) score to quantify the emotions conveyed by images:

- Valence (V): The positivity or negativity of the emotion, ranging from 1 (negative) to 5 (positive).
- Arousal (A): The intensity or activation level of the emotion, ranging from 1 (calm) to 5 (excited).

To build ContiEmo, we follow a five-stage pipeline as illustrated in Figure 2: data selection, emotion editing, video sampling, emotion reasoning, and human annotation.

Data Selection. We begin by selecting 5,000 base images suitable for emotion editing from EmoSet-118K (Yang et al., 2023). The selection criteria include: high image quality, naturalistic scenery, distinct subjects, and semantic richness to support affective editing and interpretation.

Emotion Editing. To explicitly control the direction of emotional variation, we utilize the Nano Banana Pro model (Team et al., 2023) to edit each base image into four affective targets, corresponding to the four quadrants of V–A space (High/Low V-A): HVHA, HVLA, LVHA, and LVLA. This yields four (base, edited) pairs per base image, each representing a distinct affective transition direction.

Video Sampling. Given the four affective directions, we construct densely sampled homologous transition sequences using the temporal coherence of generated videos. Specifically, we employ Wan2.2 (Wan et al., 2025) with a first–last-frame conditioned video generation technique to synthesize videos that start from the base image and end at the corresponding edited image. These videos feature smooth visual transitions with varying emotions in V–A space. From each video, we uniformly sample 4 intermediate frames to obtain image sequences. We impose no constraint that the emotions of frames follow linear interpolation or correspond to intermediate emotional states between endpoints. Instead, the emotion values of all sampled frames are independently annotated by human annotators. This stage converts discrete edits into dense homologous image groups with visual-affective variations.

Emotion Reasoning. To provide diverse emotion reasoning processes, we employ a set of state-of-the-art LVLMs to analyze each image: Gemini 3 Pro (Team et al., 2023), GPT-5.2 (Singh et al., 2025), and Qwen-VL-Max (Bai et al., 2025). The models identify semantic associations between visual elements (e.g., lighting, weather cues) and implied affective content, and we retain their CoT reasoning texts as candidates for knowledge distillation and preference alignment.

Human Annotation. We recruit 50 annotators to provide multi-dimensional labels for all images. Under a balanced allocation protocol, each image is assigned to five annotators randomly selected from the pool. The annotations include:

(1) Affective Value Annotation: Annotators assign valence and arousal scores on a 1–5 scale based on their subjective understanding of the emotion conveyed by the image.

(2) Coarse-Grained Factor Importance Ranking: We summarize seven major factors that commonly influence perceived image emotion: weather, lighting, color, composition, elements, perspective, and theme. Annotators rank these factors from most to least important according to which cues dominate their affective judgment for each image.

(3) Fine-Grained Factor Content Annotation: For fine-grained supervision, we select weather, lighting, and color as three scene-level factors that are broadly applicable and visually observable across our heterogeneous image collection, and ask annotators to categorize specific attributes: Weather type (8 categories, e.g., sunny, rainy), Lighting intensity (5 levels from strong to weak), Color richness (5 levels from rich to monotone).

(4) Preference Ranking over LVLM Reasoning CoTs: For each image, annotators review the three candidate reasoning texts produced by the LVLMs and rank them based on convincingness, consistency with the image content, and alignment with their own interpretation and analytical logic.

The final V-A values are obtained by averaging the five ratings, while factor annotations are determined by majority voting. Through this pipeline, ContiEmo provides large-scale homologous image groups with dense transitions, enriched with diverse human annotations. In this work, VEA remains a single-image assessment task. Homologous images are not treated as sequence inputs; instead, each image is independently used for training, exposing the model to fine-grained visual-affective variations.

3.2 Multi-Teacher Synergistic Affective Reasoning

Given the intrinsic abstract and multi-cue nature of real-world emotions, effective assessment necessitates sophisticated common-sense knowledge and deep semantic-logical reasoning, which small-scale models typically lack. To enhance the model's reasoning ability while capturing the inherent subjectivity and ambiguity of human emotional perception, we propose a novel Multi-Teacher Synergistic Affective Reasoning mechanism, as shown in Figure 3 (a). Specifically, we leverage the high-quality

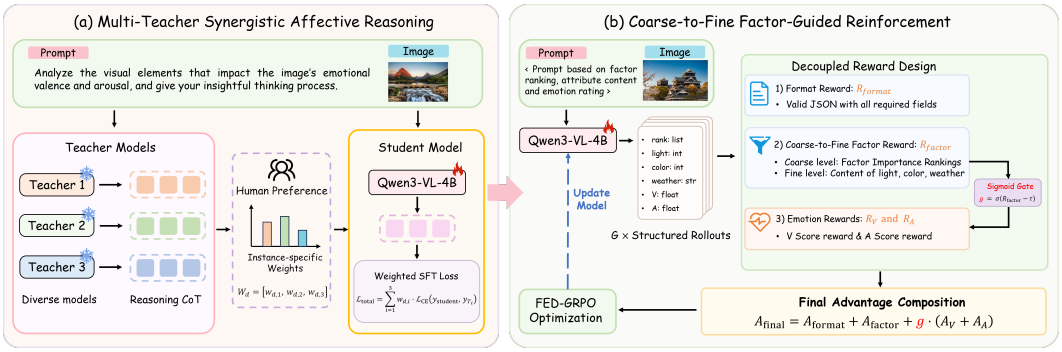


Figure 3: The overall architecture of EmoInsight.

CoT reasoning texts generated by LVLM teachers (denoted as T_1, T_2, T_3) as the synergistic distillation targets, allowing our lightweight Qwen3-VL-4B (Bai et al., 2025) backbone to inherit complex semantic insights and logical reasoning through Supervised Fine-Tuning (SFT). To adapt the contribution of different teacher CoTs across instances, we leverage human preference rankings collected from a pool of 50 annotators, with each image independently ranked by five annotators, to obtain instance-specific teacher weights. Formally, let $R_{d,p,i} \in \{1, 2, 3\}$ represent the rank assigned by the p-th annotator to the CoT generated by model T_i for image d. Our strategy consists of two stages:

Dynamic Weighting. The raw rankings are converted into preference scores and normalized into a distillation weight vector $W_d = [w_{d,1}, w_{d,2}, w_{d,3}]$ for each image d. This vector represents the probability distribution of human preferences over the three reasoning styles rather than absolute correctness, reflecting the consensus or divergence of human judgment. The weight is calculated as:

$$C_{d,i} = \sum_{p=1}^{5} (4 - R_{d,p,i}),$$
$$w_{d,i} = \frac{C_{d,i}}{\sum_{j=1}^{3} C_{d,j}}. \tag{1}$$

Weighted Loss SFT. Rather than converging toward a single fixed target, the student model learns from a weighted distribution. We minimize a dynamically weighted Cross-Entropy (CE) loss between the student output $y_{student}$ and teacher reasoning texts y_{T_i} derived from all three teachers:

$$\mathcal{L}_{total} = \sum_{i=1}^{3} w_{d,i} \cdot \mathcal{L}_{CE}(y_{student}, y_{T_i}). \tag{2}$$

This multi-teacher approach with soft-target supervision effectively models the uncertainty and diversity inherent in emotional interpretations, enabling the model to learn from diverse teacher CoTs while adapting their contributions to human preferences.

3.3 Coarse-to-Fine Factor-Guided Reinforcement

To further enhance the model's sensitivity to fine-grained visual–affective associations, we introduce a Coarse-to-Fine Factor-Guided Reinforcement strategy in Figure 3 (b). This strategy introduces hierarchical auxiliary supervision through affective factors closely related to emotion variations.

Specifically, instead of constraining only the final emotion assessment results, we impose a series of factor-level rewards, ranging from coarse to fine granularity, to jointly supervise factor prediction and affect estimation. In total, we incorporate four rewards for reinforcement learning (RL) as follows:

Format Reward. This reward enforces structured outputs by requiring each response to be a JSON object with six fields: rank, weather, light, color, valence and arousal. The reward is 1 if the response satisfies all conditions; otherwise, it is 0. At inference time, EmoInsight generates a CoT describing affective cues, followed by a structured JSON output containing factor predictions and V-A scores.

Coarse-to-Fine Factor Reward. We encourage the model's predictions to agree with human-annotated affective factors. At the coarse level, the model is required to rank the importance of the

Table 1: Performance comparison across various methods on three datasets. Best scores are in blue, second-best in green.

	ContiEmo							EMOTIC					FindingEmo					
	Emotion				Factor			Emotion				Factor	Emotion				Factor	
Method	V-Err↓	A-Err↓	V-SR↑	A-SR↑	W-ACC↑	L-ACC↑	C-ACC↑	V-Err↓	A-Err↓	V-SR↑	A-SR↑	S-ACC↑	V-Err↓	A-Err↓	V-SR↑	A-SR↑	S-ACC↑	D-ACC↑
StyleEDL	0.652	0.429	0.591	-0.402	-	-	-	0.509	0.549	0.180	0.532	-	0.765	0.635	0.374	0.232	-	-
EmoSet	0.430	0.366	0.721	0.433	0.363	0.497	0.544	0.466	0.589	0.188	0.565	0.483	0.584	0.593	0.412	0.173	0.250	0.537
VA-CLIP	0.721	0.315	0.605	-0.191	-	-	-	0.480	0.637	0.474	0.221	-	0.798	0.606	0.639	0.345	-	-
MFRN	0.486	0.349	0.710	0.429	-	-	-	0.483	0.617	0.021	0.321	-	0.755	0.650	0.285	0.198	-	-
Qwen3-VL-4B	0.611	0.419	0.696	0.436	0.619	0.503	0.563	0.524	0.790	0.339	0.350	0.587	0.494	0.665	0.590	0.333	0.279	0.610
Qwen3-VL-8B	0.535	0.360	0.715	0.454	0.653	0.519	0.578	0.501	0.746	0.353	0.430	0.631	0.484	0.618	0.627	0.372	0.309	0.625
EmoInsight	0.371*	0.246*	0.767*	0.500*	0.934*	0.606*	0.625*	0.351*	0.430*	0.549*	0.706*	0.850*	0.417*	0.533*	0.693*	0.486*	0.544*	0.676*

Note: * indicates a statistically significant improvement over the best-performing baseline for each metric ($p < 0.05$). – indicates that the method does not support the factor prediction.

$n = 7$ major factors. For each factor i, its ranking reward is based on the positional discrepancy between the predicted ranking $\hat{P}_i$ and the ground-truth ranking P_i^{gt}, with $w_i = n - i$ assigning greater emphasis to higher-ranked factors. We normalize the weighted discrepancy using a position-dependent upper bound $\text{d_max} = [6, 5, 4, 3, 4, 5, 6]$. At the fine level, we further supervise the classification of three scene-level factors. For each factor f in $\mathcal{F} = \{\text{weather}, \text{light}, \text{color}\}$, its reward is defined as a binary indicator function $\mathbb{I}(\cdot)$, taking value 1 when the predicted category $\hat{c}_f$ matches the ground-truth category c_f^{gt}, and 0 otherwise. The complete formulation is:

$$\begin{aligned} R_{\text{coarse}} &= 1 - \frac{\sum_{i=0}^{n-1} w_i \left| \hat{P}_i - P_i^{\text{gt}} \right|}{\sum_{i=0}^{n-1} w_i \cdot \text{d_max}_i}, \\ R_{\text{fine}} &= \frac{1}{3} \sum_{f \in \mathcal{F}} \mathbb{I}\left(\hat{c}_f = c_f^{\text{gt}} \right), \\ R_{\text{factor}} &= R_{\text{coarse}} + R_{\text{fine}}. \end{aligned} \tag{3}$$

Valence and Arousal Reward. These rewards guide the model to reason about the degree of emotion expressed along the V-A dimensions. The predicted values $(\hat{V}, \hat{A})$ are compared with the ground-truth values (V_{gt}, A_{gt}). If the discrepancy falls within a predefined threshold ϵ, the reward is assigned to 1. This design allows predictions with acceptable deviations to receive positive feedback without requiring exact matches. The formulations of the V reward R_{V} and A reward R_{A} are as follows:

$$R_V = \mathbb{I}\left(\left| \hat{V} - V_{\text{gt}} \right| \leq \varepsilon \right), \quad R_A = \mathbb{I}\left(\left| \hat{A} - A_{\text{gt}} \right| \leq \varepsilon \right). \tag{4}$$

In our Coarse-to-Fine Factor-Guided Reinforcement strategy, the Coarse-to-Fine Factor Reward reflects the consistency between predicted factors and human annotations. Based on this, we design a sigmoid-based gating mechanism that weights the advantages of V-A Rewards according to the quality of the inferred factors. When factor predictions show lower agreement with human preference, the confidence in the final emotion estimation is correspondingly low, and thus a smaller weight is assigned. The gated weight is formulated as $1/(1 + e^{-(R_{\text{factor}} - \tau)})$, where τ is a gating threshold. The final advantage is computed as the sum of the format, factor and the gated V-A advantages.

Most commonly used RL paradigms, such as PPO (Schulman et al., 2017) and GRPO (Shao et al., 2024), directly sum multiple reward components and then apply normalization. Under our process-outcome reward setting, this often causes subtle differences among reward combinations to collapse into identical advantage values, preventing the model from distinguishing whether the factor reasoning or emotion assessment requires further optimization. Therefore, we design a factor-emotion decoupled paradigm termed FED-GRPO, which performs group-wise normalization for factor and emotion reward independently, preserving more informative learning signals from the two reward types.

4 Experiment

4.1 Baselines and Datasets

To rigorously benchmark the affective understanding capabilities of our proposed method, we compare it against 6 representative baselines, categorized into two groups:

Figure 4: Attention visualization of Qwen3-VL-4B and EmoInsight.

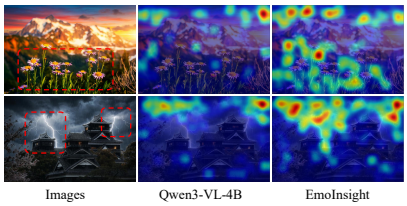


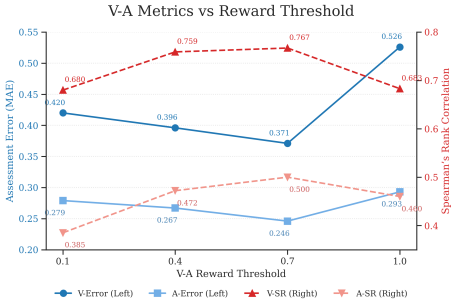
Table 2: Ablation of reward function design in Coarse-to-Fine Factor-Guided Reinforcement strategy. Best in blue.

Method	V	A	Factor	Guide	V-Err ↓	A-Err ↓	V-SR ↑	A-SR ↑
Without RL	✗	✗	✗	✗	0.614	0.309	0.554	0.156
Outcome-only	✓	✓	✗	✗	0.408	0.270	0.713	0.469
Naive Combine	✓	✓	✓	✗	0.397	0.254	0.739	0.481
Ours	✓	✓	✓	✓	0.371	0.246	0.767	0.500

Table 3: Ablation of Multi-Teacher Synergistic Affective Reasoning. Best scores are in blue.

Study	Method	V-Err ↓	A-Err ↓
Teacher Selection	Original Qwen3-VL-4B	0.888	0.512
	Single: Qwen-VL-Max	0.651	0.337
	Single: GPT-5.2	0.642	0.335
	Single: Gemini3 Pro	0.634	0.385
	Multi-Teacher	0.614	0.309
Teacher Weighting	Uniform Weighting	0.711	0.349
	Fixed Weighting	0.701	0.327
	Dynamic Weighting	0.614	0.309

Figure 5: Performance variation across reward thresholds.



Specialized VEA Methods: StyleEDL (Jing et al., 2023), EmoSet (Yang et al., 2023), VA-CLIP (Mertens et al., 2024), MFRN (Xu et al., 2025). They are designed for VEA.

General-Purpose LVLMs: Qwen3-VL-4B and Qwen3-VL-8B (Bai et al., 2025), with strong general understanding capabilities.

To evaluate the affective assessment and factor recognition capabilities of different methods, we conduct experiments on ContiEmo and two public benchmarks, EMOTIC and FindingEmo. All methods are independently trained and evaluated on the same splits of each dataset. Both public datasets contain human-annotated V-A values along with certain factor labels for isolated images and are used to evaluate the generalization. The training objectives are adapted to the annotations in each dataset. Specifically, EMOTIC provides labels for the emotional states of persons in images, while FindingEmo includes annotations for both emotional states and the decision factor for V-A judgment.

4.2 Implementation Details and Evaluation Metrics

We implement all models on 8 NVIDIA H20 GPUs (96 GB each) under Linux with Python 3.10, PyTorch 2.5.1, and CUDA 12.4. For fair comparison, all methods follow a unified experimental protocol with identical training-test splits. To prevent information leakage, we partition the dataset at the base-image level, ensuring that each base image and all its derived intermediate frames and endpoint images belong to the same split. Hyperparameters for baselines follow the settings recommended in their original papers. EmoInsight is initialized with Qwen3-VL-4B, trained with a batch size of 16 and the AdamW optimizer at a learning rate of 1e-4. The V-A reward threshold ϵ is set to 0.7 and the gating threshold τ is set to 1. Detailed LVLM settings are offered in the Appendix.

We use several metrics for evaluation: **V-Err and A-Err** use Mean Absolute Error to measure the emotion assessment precision, while **V-SR and A-SR** compute Spearman's Rank Correlation Coefficient to evaluate the consistency between model predictions and human-annotated V-A. For factor reasoning, we adopt **W-ACC, L-ACC, and C-ACC** on ContiEmo to quantify the accuracy of weather, light, and color factors, **S-ACC** on EMOTIC to measure the accuracy of the state factor, and **S-ACC and D-ACC** on FindingEmo to assess the accuracy of state and decision factors. Statistical significance is evaluated in Table 1 using a two-sided paired bootstrap test with 10,000 resamples. ContiEmo is resampled at the base-image level to preserve dependencies among homologous samples. The p-values are adjusted using Holm's correction, with $p < 0.05$ considered statistically significant.

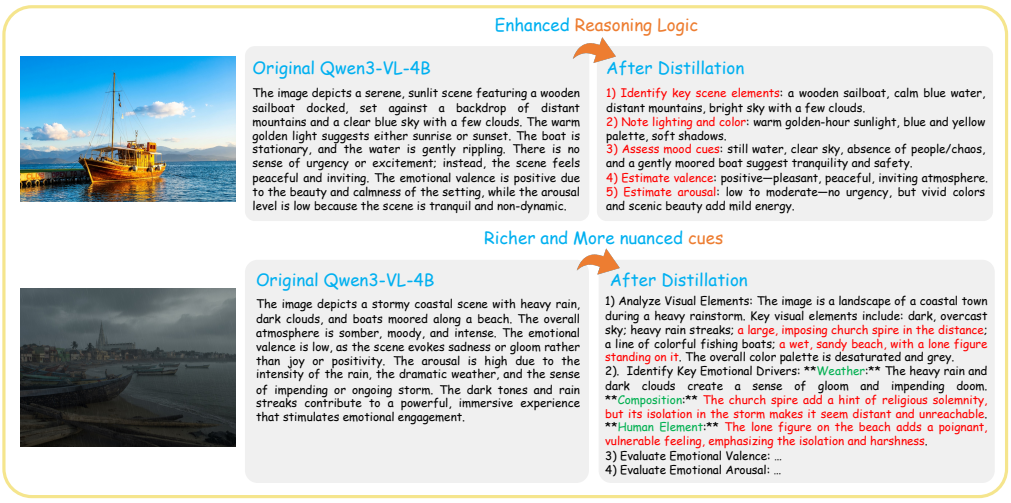


Figure 6: Qualitative comparison of affective CoTs before and after multi-teacher distillation.

4.3 Comparative Analysis

Quantitative Results. Table 1 presents the quantitative comparison between EmoInsight and the baselines. Our model achieves state-of-the-art performance, recording the best affect estimation accuracy and rank-order consistency while securing the highest accuracy across all factor-level metrics. Its improvements over the best-performing baseline for each metric are statistically significant. These results demonstrate that our method excels in both affective assessment and multi-factor recognition. For specialized VEA methods, they suffer from a deficiency of deep semantic reasoning. While general LVLMs have reasoning ability, they struggle to map visual semantics to precise emotional values. As a result, they remain less sensitive to subtle affective differences across similar images.

Qualitative Visualization. To investigate the model's perception, we compare the attention maps of EmoInsight and Qwen3-VL-4B baseline in Figure 4. The backbone model only attends to coarse affective cues such as warm sunsets or dark clouds, while often overlooking fine-grained details relevant to human emotion perception, e.g., foreground flowers inducing pleasantness or lightning that increases arousal. In contrast, EmoInsight consistently focuses on these key visual regions and correctly infers their affective implications, suggesting greater sensitivity to fine-grained cues.

4.4 Ablation Studies

4.4.1 Effectiveness of Multi-Teacher Synergistic Affective Reasoning Mechanism

We evaluate the impact of our synergistic distillation strategy by comparing the original Qwen3-VL-4B backbone without task-specific training against versions fine-tuned via single-teacher distillation and our proposed Multi-Teacher mechanism. To validate the effect of dynamic weighting, we compare our method with two variants: uniform weighting without human preferences and fixed weighting with human preferences averaged across all samples. All teacher-based variants are trained solely through CoT distillation without V-A supervision. In Table 3, all models utilizing reasoning-based SFT exhibit consistent performance gains over the original backbone, confirming that teacher-generated CoTs transfer affective reasoning knowledge that benefits downstream emotion assessment. Notably, the Multi-Teacher setting achieves the best overall performance. Moreover, fixed weighting yields only marginal gains over uniform weighting, while dynamic instance-specific weighting yields greater improvements. This suggests that adapting teacher contributions to instance-specific preferences better accommodates diverse affective interpretations.

To provide a more intuitive visualization of the contribution of CoT distillation to the model's reasoning capabilities, we compare the thinking processes of Qwen3-VL-4B baseline and the distilled model. As illustrated in Figure 6, our distillation mechanism enhances the model in both reasoning logic and detail perception. The distilled model not only generates clear and well-structured CoT, but also identifies subtle cues vital for emotional judgment that are overlooked by baseline.

Table 4: Ablation of reward decoupling.

Method	V-Err ↓	A-Err ↓	V-SR ↑	A-SR ↑
w/o FED	0.432	0.258	0.752	0.476
Ours	0.371	0.246	0.767	0.500

Table 5: Ablation of the sampled intermediate frames.

Method	V-Err ↓	A-Err ↓	V-SR ↑	A-SR ↑
Only Endpoint Images	0.466	0.272	0.698	0.435
Add Random Images	0.428	0.261	0.724	0.457
Add Intermediate Frames	0.371	0.246	0.767	0.500

4.4.2 Effectiveness of Coarse-to-Fine Factor-Guided Reinforcement strategy

We further analyze the impact of our Reinforcement Learning (RL) strategy on the distilled model.

Reward Function Design. We compare four reward configurations: (a) the distilled Qwen3-VL-4B without RL; (b) outcome-only V-A rewards; (c) naive combination of V-A and Coarse-to-Fine Factor rewards; and (d) our proposed rewards. The results in Table 2 indicate that incorporating Factor rewards improves agreement with factor recognition while also benefiting V-A estimation. Furthermore, our Factor-Guided mechanism achieves the best performance, suggesting that gating affective rewards with visual factors enables more effective joint optimization of affect estimation and factor-level agreement with human judgments.

RL Paradigm. To validate the effect of FED-GRPO, we conduct a comparative analysis against the standard GRPO paradigm. As presented in Table 4, model with FED yields a distinct improvement in affective assessment performance, underscoring the critical role of reward decoupling. Isolating reward signals enables the model to explicitly attend to the distinct characteristics inherent in affective and factor rewards, facilitating more targeted optimization. In contrast, GRPO with simple summation hinders model's convergence toward a state that satisfies all optimality objectives.

V-A Reward Threshold. In Equation (4), we introduce a threshold parameter ϵ to determine the allocation of reward. We investigate the impact of various thresholds. Figure 5 illustrates the variation of V-Err, A-Err, V-SR, and A-SR. As observed, performance is constrained when the threshold is relatively small. This limitation stems from overly stringent reward conditions. Reward sparsity makes it difficult for the model to converge effectively. As the threshold increases, the performance improves consistently. However, a further increase beyond a certain point results in performance degradation. This is because an excessively large threshold allows the model to receive rewards even with imprecise predictions. Thus, we selected 0.7 as the final threshold.

4.4.3 Effectiveness of Dense Homologous Frames

To disentangle the benefit of dense homologous supervision, we compare training with endpoint images only against two augmented settings: adding intermediate frames and adding an equal number of random images. The random images are sampled to match the V-A distribution of the homologous frames while originating from different base-image groups. As shown in Table 5, adding random images yields only marginal improvements, whereas adding intermediate frames substantially improves both emotion assessment precision and consistency. These results indicate that the gains arise not merely from the increased training-set size. Instead, preserving semantic correspondence while varying affective appearance provides more effective information for fine-grained VEA.

5 Conclusion

We present ContiEmo, a large-scale visual emotion assessment (VEA) dataset built from homologous image groups with densely annotated affective variations and rich reasoning supervision. Built upon ContiEmo, we propose EmoInsight, a fine-grained VEA model that combines semantic reasoning distillation with factor-aware reinforcement learning. Specifically, we introduce a Multi-Teacher Synergistic Affective Reasoning mechanism to transfer diverse affective knowledge using instance-specific dynamic weighting, enhancing the model's semantic reasoning capability. To further capture fine-grained visual–affective correlations, we propose a Coarse-to-Fine Factor-Guided Reinforcement strategy to jointly optimizes V-A estimation and agreement with human-annotated visual factors. Extensive experiments on ContiEmo and public real-world VEA datasets demonstrate consistent improvements in affect estimation and factor recognition, while ablations suggest the benefit of dense homologous supervision, validating the effectiveness of both our benchmark and framework.

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, 2025. *URL https://arxiv.org/abs/2511.21631*, 2025.

Damian Borth, Rongrong Ji, Tao Chen, Thomas Breuel, and Shih-Fu Chang. Large-scale visual sentiment ontology and detectors using adjective noun pairs. In *Proceedings of the 21st ACM international conference on Multimedia*, pp. 223–232, 2013.

Hanna Chainay, George A Michael, Mélissa Vert-Pré, Lionel Landré, and Amandine Plasson. Emotional enhancement of immediate memory: Positive pictorial stimuli are better recognized than neutral or negative pictorial stimuli. *Advances in Cognitive Psychology*, 8(3):255, 2012.

Hatice Gunes, Björn Schuller, Maja Pantic, and Roddy Cowie. Emotion representation, analysis and synthesis in continuous space: A survey. In *2011 IEEE international conference on automatic face & gesture recognition (FG)*, pp. 827–834. IEEE, 2011.

Longteng Guo, Jing Liu, Peng Yao, Jiangwei Li, and Hanqing Lu. Mscap: Multi-style image captioning with unpaired stylized text. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4204–4213, 2019.

Peiguang Jing, Xianyi Liu, Ji Wang, Yinwei Wei, Liqiang Nie, and Yuting Su. Styleedl: Style-guided high-order attention network for image emotion distribution learning. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 853–861, 2023.

Dimitrios Kollias. Abaw: Valence-arousal estimation, expression recognition, action unit detection & multi-task learning challenges. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2328–2336, 2022.

Ronak Kosti, Jose M Alvarez, Adria Recasens, and Agata Lapedriza. Emotic: Emotions in context dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 61–69, 2017a.

Ronak Kosti, Jose M Alvarez, Adria Recasens, and Agata Lapedriza. Emotion recognition in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1667–1675, 2017b.

Jiansheng Li, Chuanlan Luo, Qi Zhang, and Rustam Shadiev. Can emotional design really evoke emotion in multimedia learning? *International Journal of Educational Technology in Higher Education*, 17(1):24, 2020a.

Leida Li, Hancheng Zhu, Sicheng Zhao, Guiguang Ding, and Weisi Lin. Personality-assisted multi-task learning for generic and personalized image aesthetics assessment. *IEEE Transactions on Image Processing*, 29:3898–3910, 2020b.

Zuhe Li, Yangyu Fan, Bin Jiang, Tao Lei, and Weihua Liu. A survey on sentiment analysis and opinion mining for social multimedia. *Multimedia Tools and Applications*, 78(6):6939–6967, 2019.

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. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025.

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.

Laurent Mertens, Elahe Yargholi, Hans Op 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.

Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1):18–31, 2017.

Kuan-Chuan Peng, Tsuhan Chen, Amir Sadovnik, and Andrew C Gallagher. A mixed bag of emotions: Model, predict, and transfer emotion distributions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 860–868, 2015.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

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.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

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.

Antoine Toisoul, Jean Kossaifi, Adrian Bulat, Georgios Tzimiropoulos, and Maja Pantic. Estimation of continuous valence and arousal levels from faces in naturalistic conditions. *Nature Machine Intelligence*, 3(1):42–50, 2021.

Chai M Tyng, Hafeez U Amin, Mohamad NM Saad, and Aamir S Malik. The influences of emotion on learning and memory. *Frontiers in psychology*, 8:235933, 2017.

Manuelde Vega. The representation of changing emotions in reading comprehension. *Cognition & emotion*, 10(3):303–322, 1996.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Jing Wang and Xin Geng. Label distribution learning by exploiting label distribution manifold. *IEEE transactions on neural networks and learning systems*, 34(2):839–852, 2021.

Matthias J Wieser, Elisabeth Klupp, Peter Weyers, Paul Pauli, David Weise, Daniel Zeller, Joseph Classen, and Andreas Mühlberger. Reduced early visual emotion discrimination as an index of diminished emotion processing in parkinson's disease?–evidence from event-related brain potentials. *Cortex*, 48(9):1207–1217, 2012.

Haitao Xiong, Hongfu Liu, Bineng Zhong, and Yun Fu. Structured and sparse annotations for image emotion distribution learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 363–370, 2019.

Liwen Xu, Zhengtao Wang, Bin Wu, and Simon Lui. Mdan: Multi-level dependent attention network for visual emotion analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9479–9488, 2022.

Qinfu Xu, Shaozu Yuan, Yiwei Wei, Jie Wu, Leiquan Wang, and Chunlei Wu. Multiple feature refining network for visual emotion distribution learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 8924–8932, 2025.

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.

Jufeng Yang, Ming Sun, and Xiaoxiao Sun. Learning visual sentiment distributions via augmented conditional probability neural network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

Jufeng Yang, Dongyu She, Yu-Kun Lai, Paul L Rosin, and Ming-Hsuan Yang. Weakly supervised coupled networks for visual sentiment analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7584–7592, 2018.

Xiaocui Yang, Shi Feng, Daling Wang, and Yifei Zhang. Image-text multimodal emotion classification via multi-view attentional network. *IEEE Transactions on Multimedia*, 23:4014–4026, 2020.

Quanzeng You, Jiebo Luo, Hailin Jin, and Jianchao Yang. Building a large scale dataset for image emotion recognition: The fine print and the benchmark. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30, 2016.

Zhicheng Zhang, Weicheng Wang, Yongjie Zhu, Wenyu Qin, Pengfei Wan, Di Zhang, and Jufeng Yang. Videmo: Affective-tree reasoning for emotion-centric video foundation models. In *Advances in Neural Information Processing Systems*, 2025.

Sicheng Zhao, Yue Gao, Xiaolei Jiang, Hongxun Yao, Tat-Seng Chua, and Xiaoshuai Sun. Exploring principles-of-art features for image emotion recognition. In *Proceedings of the 22nd ACM international conference on Multimedia*, pp. 47–56, 2014.

APPENDIX

A DETAILED IMPLEMENTATION

We implement EmoInsight on eight NVIDIA H20 GPUs. The training pipeline consists of two stages: Multi-Teacher Synergistic Affective Reasoning based on LoRA fine-tuning, followed by full-parameter Coarse-to-Fine Factor-Guided Reinforcement. The detailed settings are as below.

A.1 MULTI-TEACHER SYNERGISTIC AFFECTIVE REASONING

We initialize the student model from Qwen3-VL-4B and implement the multi-teacher distillation stage using LoRA. LoRA adapters are applied to all linear layers, while the vision encoder and vision-language aligner remain frozen. We enable sequence packing and gradient checkpointing to improve training efficiency. With eight GPUs, a per-device batch size of 16 results in 128 packed sequences per optimization step. Table 6 summarizes the main hyperparameters.

Table 6: Configuration of the Multi-Teacher Synergistic Affective Reasoning stage.

Configuration	**Value**
Base model	Qwen3-VL-4B
Training strategy	LoRA-based SFT
Target modules	All linear layers
Frozen modules	Vision encoder and aligner
LoRA rank / alpha	64 / 128
Training epochs	2
Per-device batch size	16
Gradient accumulation	1
Optimizer	AdamW
Learning rate	1×10^{-4}
Warmup ratio	0.05
Maximum prompt length	4096
Sequence packing	Enabled
Gradient checkpointing	Enabled

A.2 COARSE-TO-FINE FACTOR-GUIDED REINFORCEMENT

The reinforcement-learning stage is initialized from the multi-teacher SFT checkpoint and optimized with full-parameter training. For each prompt, the policy samples four completions for group-relative advantage estimation. In FED-GRPO, the global training batch size is defined at the completion level. A per-device batch size of 16 yields 128 completions per optimization step. The main training configurations are summarized in Table 7.

A.3 INFERENCE CONFIGURATION

During inference, we enable stochastic decoding with `do_sample=True`. The temperature, top-k, and top-p are set to 1.0, 50, and 0.95, respectively.

B PROMPT TEMPLATES AND OUTPUT SCHEMA

Table 8 presents the prompt templates used for affective image editing, first-last-frame video generation, teacher reasoning, Multi-Teacher SFT, and factor-guided reinforcement learning. For image editing, the target valence and arousal levels are instantiated as either *high* or *low*, yielding the four V-A quadrants: HVHA, HVLA, LVHA, and LVLA. The same visual emotion reasoning prompt is applied to all three teachers (Gemini3 Pro, GPT-5.2, Qwen-VL-Max), and as the input prompt for student model during Multi-Teacher SFT.

Tables 9 and 10 present two examples of the reasoning processes generated by the three teachers. The models exhibit distinct analytical patterns for the same image, reflecting the diversity of affective

Table 7: Configuration of the Coarse-to-Fine Factor-Guided Reinforcement stage.

Configuration	**Value**
Initialization	Multi-Teacher checkpoint
Training strategy	Full-parameter FED-GRPO
Training epochs	2
Generations per prompt G	4
Per-device batch size	16
Gradient accumulation	1
Optimizer	AdamW
Learning rate	1×10^{-4}
Maximum prompt length	4096
Maximum completion length	256
KL coefficient β	0.001
V-A reward threshold ϵ	0.7
Gating threshold τ	1.0
Gradient checkpointing	Disabled

reasoning. By incorporating instance-specific human preferences, our approach effectively aligns these diverse reasoning patterns with human judgments.

Figures 7 shows additional visualization results of our Multi-Teacher Synergistic Affective Reasoning.

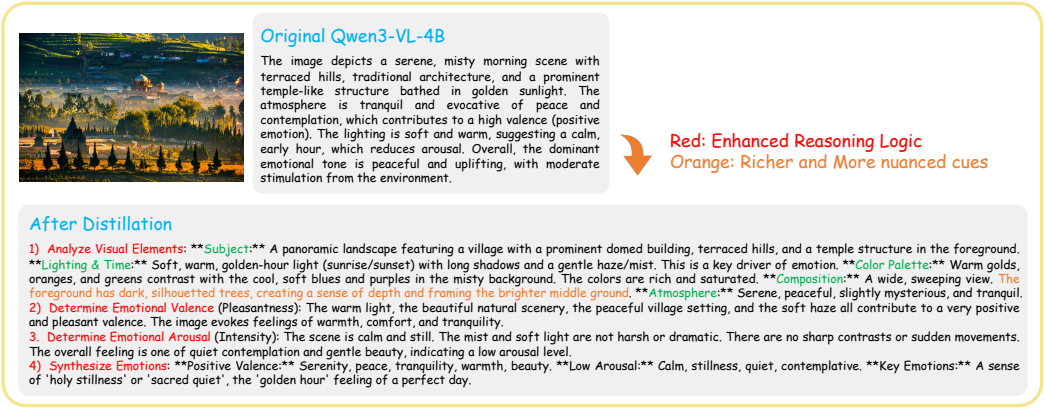


Figure 7: Qualitative comparison of affective CoTs before and after multi-teacher distillation.

C ContiEmo Construction Details

C.1 Dataset V–A Distribution

Figure 8 compares the V-A distributions of the edited endpoint images and the sampled intermediate frames. The endpoint images form separated clusters across the four target V–A quadrants, with relatively sparse coverage around the neutral center $(V = 3, A = 3)$. This distribution reflects the four affective editing directions: HVHA, HVLA, LVHA, and LVLA. In contrast, the intermediate frames are densely distributed around the central transition region and extend toward different affective directions. Their V-A values are independently annotated, providing fine-grained samples between affective endpoints. Therefore, combining endpoint and intermediate images improves the coverage density of the V-A space, yielding homologous images with continuous visual-affective variations.

C.2 Annotation Protocol

We recruit a pool of 50 annotators and adopt a balanced allocation protocol. Each image is independently assigned to five annotators randomly selected from the pool, while the overall annotation

Table 8: Prompt templates used in the data-construction and training pipeline of EmoInsight.

Affective Image Editing: Nano Banana Pro

You are an expert in affective image editing with a strong understanding of emotional Valence and Arousal. Edit the input image to convey **[VALENCE LEVEL]** Valence and **[AROUSAL LEVEL]** Arousal. Preserve the original composition, scene structure, and silhouettes of major objects, including mountains and buildings when present. The resulting image should exhibit cinematic lighting and photorealistic quality.

Instantiation: **[VALENCE LEVEL]** and **[AROUSAL LEVEL]** are set to *high* or *low*, corresponding to the HVHA, HVLA, LVHA, and LVLA editing directions.

First–Last-Frame Video Generation: Wan2.2

Generate a smooth and temporally coherent video from the start frame to the end frame. Keep the camera viewpoint fixed and preserve the scene geometry and object identities. Ensure that the color palette and lighting evolve continuously throughout the video. The output should exhibit cinematic lighting and photorealistic quality.

Teacher Reasoning and Multi-Teacher SFT

You are an expert in visual emotion reasoning with a clear understanding of emotional Valence and Arousal. Analyze the provided image and generate an insightful reasoning trace explaining the semantic associations between its visual elements and implied affective content. Highlight the key visual elements, scene context, and situations that substantially influence the perceived Valence and Arousal.
Return the result strictly as a single JSON object using the following structure:

```
{
  "thinking process":  "<your reasoning process>"
}
```

Do not include any additional text outside the JSON object.

Coarse-to-Fine Factor-Guided Reinforcement

You are an expert in visual emotion assessment. Analyze the provided image according to the following instructions:
1. Weather Condition: Identify the weather using one of the following categories: *Sunny, Cloudy, Overcast, Rainy, Foggy, Snowy, Thunderstorm,* or *Indoor*.
2. Lighting Intensity: Evaluate the lighting intensity using one of the following levels: *Very Strong, Strong, Moderate, Weak,* or *Very Weak*.
3. Color Richness: Evaluate the color richness using one of the following levels: *Very Rich, Rich, Moderate, Monotonous,* or *Very Monotonous*.
4. Factor Ranking: Rank the following seven factors according to their importance in determining the image's Valence and Arousal, from most important to least important: *Weather, Light, Color, Composition, Elements, Perspective,* and *Theme*. Each factor must appear exactly once.
5. Emotion Rating: Rate Valence and Arousal on a continuous scale from 1.0 to 5.0. For Valence, 1.0 represents very negative affect and 5.0 represents very positive affect. For Arousal, 1.0 represents very calm affect and 5.0 represents very active affect. Report both values as floating-point numbers rounded to one decimal place.
Output Requirement: Return the result strictly and exclusively as a single JSON object enclosed by `<answer>` and `</answer>` tags using the following structure:

```
<answer>
{
  "weather":  "str",
  "light":  "str",
  "color":  "str",
  "rank":  ["factor1", "factor2", "factor3", "factor4", "factor5", "factor6",
"factor7"],
  "valence":  float,
  "arousal":  float
}
</answer>
```

Do not include conversational text, Markdown code blocks, or explanations inside or outside the answer tags.

workload is balanced across annotators. Each annotator provides a pair of continuous V–A ratings, a complete importance ranking of seven affective factors, fine-grained factor categories, and a preference ranking over the three teacher-generated CoTs.

C.2.1 Affective Values.

Annotators independently rate Valence and Arousal on a five-point continuous scale. The final V–A annotation of each image is obtained by averaging the five individual ratings.

C.2.2 Coarse-Grained Factor Importance Ranking.

For coarse-grained annotation, annotators rank the following seven factors from most to least important: *Weather, Light, Color, Composition, Elements, Perspective,* and *Theme*. Ties are not allowed within an individual ranking.

Table 9: Reasoning processes generated by the three teacher models.

Gemini3 Pro: 1. **Analyze the image content:** The image shows a night-time scene of a violent storm over a body of water. Key elements include multiple bright lightning bolts, dark and heavy storm clouds, heavy rain, rough and choppy water with whitecaps, a small rocky island with trees, and a large forested mountain range in the background. The overall color scheme is dark blue and black, with bright white light from the lightning. 2. **Determine Emotional Valence:** The valence is clearly **negative**. The scene is dark, chaotic, and threatening. The storm, the rough water, and the lightning all signify danger and a hostile environment, evoking feelings of fear, anxiety, or foreboding. There are no positive or calming elements. 3. **Determine Emotional Arousal:** The arousal is **high**. The scene is dynamic and energetic. The multiple lightning strikes act as intense, sudden bursts of energy. The crashing waves, the falling rain, and the dramatic contrast between the bright flashes and the deep shadows all contribute to a sense of high tension, urgency, and power. 4. **Synthesis:** The image is a powerful depiction of a dangerous storm. The lightning and rough water are the main drivers of the high arousal, creating a sense of immediate threat and raw energy. The overall darkness, heavy clouds, and isolated landscape contribute to the negative valence, evoking feelings of fear, awe, and powerlessness against the forces of nature.
GPT-5.2: - Key scene: a nighttime thunderstorm over a lake with a dark mountain and dense forest. - Valence cues: cold blue-gray palette, heavy clouds, rain streaks, and rough waves suggest danger and discomfort (negative valence). - Arousal cues: multiple bright lightning bolts, high contrast flashes, and turbulent water create intensity, unpredictability, and urgency (high arousal). - Composition/mood: the small rocky outcrop with trees feels isolated and vulnerable against vast, violent nature, adding tension and awe.
Qwen-VL-Max: I analyze the image by focusing on key visual elements that contribute to emotional valence (the positivity or negativity of emotion) and arousal (the intensity of emotion). The scene is dominated by a stormy night with multiple lightning strikes illuminating dark clouds, which creates high arousal due to the dynamic and intense nature of lightning. The darkness and heavy rain contribute to a negative valence, evoking feelings of danger, unease, or awe. The mountainous silhouette and dense forest add a sense of isolation and grandeur, enhancing the dramatic tension. The churning water in the foreground reflects the storm's energy, further increasing arousal. The contrast between the bright lightning and the dark environment emphasizes the unpredictability and power of nature, leading to an overall feeling of awe mixed with fear. This combination results in a high-arousal, low-valence emotional state.

C.2.3 Fine-Grained Factor Content.

Annotators select one weather category from *Sunny, Cloudy, Overcast, Rainy, Foggy, Snowy, Thunderstorm,* and *Indoor*; one lighting level from *Very Strong, Strong, Moderate, Weak,* and *Very Weak*; and one color-richness level from *Very Rich, Rich, Moderate, Monotonous,* and *Very Monotonous*.

C.2.4 Annotation Aggregation.

The coarse-grained factor rankings are aggregated according to their mean positions across the five annotators, with lower mean positions indicating higher importance. When two factors have the same mean position, we sequentially compare their numbers of first-, second-, and subsequent-position votes. Any remaining exact ties are resolved using a fixed random seed. Fine-grained categorical annotations are aggregated through plurality voting. In the rare case that multiple categories receive the same number of votes, one of the tied categories is selected using the same fixed random seed. This deterministic procedure ensures that the aggregated labels are reproducible.

C.2.5 CoT Preference Ranking.

For each image, the three reasoning texts generated by the teacher LVLMs are presented to the annotators in a randomly shuffled order to reduce positional bias. Annotators rank the three CoTs according to their consistency with the image, reasoning quality, and agreement with their own affective interpretation. Ties are not permitted. The five rankings are then converted into instance-specific weights using the scoring rule defined in the main text.

Table 10: Reasoning processes generated by the three teacher models.

Gemini3 Pro: 1. **Visual Elements:** I see a dramatic sunrise or sunset over a calm body of water. A long wooden pier extends from the muddy foreground towards the sun. The sky is filled with clouds illuminated in intense shades of orange, yellow, and red. There's a soft mist over the distant hills. The light is golden and warm, casting long reflections on the water and the wet sand. 2. **Emotional Valence (Positivity):** The image is highly positive. The sunrise/sunset is a classic symbol of beauty, hope, and new beginnings. The warm, golden light is comforting and inviting. The calm water and natural setting evoke feelings of peace and serenity. 3. **Emotional Arousal (Intensity):** The arousal level is moderate to high. While the scene is peaceful (calm water, still pier), the visual intensity of the colorful sky and the bright sun is very strong. The dramatic lighting creates a sense of awe and wonder, which is an activating emotion. 4. **Synthesis:** The image combines the high positive valence of a beautiful natural event with the moderate-to-high arousal of intense, dramatic lighting. It conveys feelings of awe, tranquility, warmth, and hope.
GPT-5.2: 1) Identify core scene and cues: a calm body of water at sunset/sunrise, a long pier extending into the lake, low hills in the distance, dramatic cloud cover, warm golden-orange light. 2) Estimate emotional valence (pleasantness): warm colors, tranquil water, and scenic composition suggest positive/soothing feelings; no visible threat or conflict. 3) Estimate arousal (intensity/activation): mostly low-to-moderate due to still water and quiet setting; slightly elevated by the intense glowing sun, strong contrast, and dramatic clouds. 4) Link visual elements to emotions: warm glow + reflections -¿ comfort/hope; empty pier + vast open space -¿ solitude/contemplation; cloud drama + sun flare -¿ awe and poignancy. 5) Summarize: serene, reflective atmosphere with a touch of grandeur and melancholy from the empty, fading light.
Qwen-VL-Max: The image captures a sunset over a calm body of water with a wooden pier extending into it. The sun is low on the horizon, casting a golden-orange glow across the sky and reflecting vividly on the water's surface. The clouds are illuminated from below, creating a dramatic contrast between warm tones and darker areas. The stillness of the water enhances the reflection, contributing to a sense of tranquility. The silhouette of the pier adds structure and leads the viewer's eye toward the sun, emphasizing depth and perspective. The surrounding landscape features gentle hills partially veiled in mist, adding to the serene atmosphere. There are no visible human figures or movement, reinforcing a feeling of solitude and peace. The overall color palette—dominated by warm hues of orange, red, and gold—evokes warmth and comfort, while the soft lighting reduces emotional arousal, promoting relaxation. The scene balances beauty with stillness, resulting in high positive valence and low arousal.

C.3 Complete FED-GRPO Formulation

Given a mini-batch containing B prompts $\{x_i\}_{i=1}^{B}$, we sample G responses $\{y_{i,j}\}_{j=1}^{G}$ for each prompt x_i. For every response, we compute four reward components: the format reward, factor reward, valence reward, and arousal reward. We denote the raw value of the k-th reward component for response $y_{i,j}$ as $R_{i,j}^{(k)}$, where

$$k \in \mathcal{K} = \{\text{format}, \text{factor}, V, A\}.$$

Reward-wise Group Normalization. Unlike standard GRPO, which first combines all reward components and then performs group normalization, FED-GRPO normalizes each reward objective independently. For the k-th reward component, its group-wise mean and standard deviation under prompt x_i are computed as

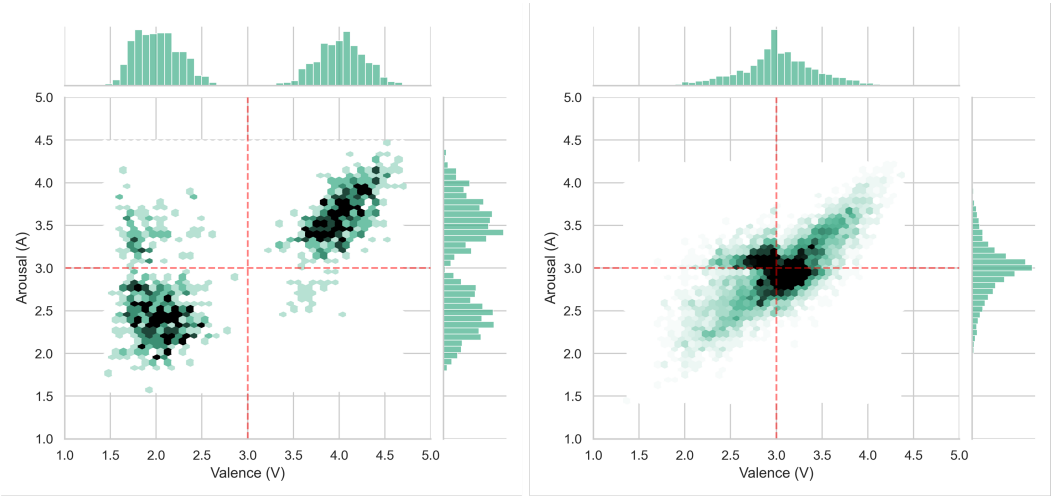
$$\mu_i^{(k)} = \frac{1}{G}\sum_{j=1}^{G} R_{i,j}^{(k)}, \tag{5}$$

and

$$\sigma_i^{(k)} = \sqrt{\frac{1}{G}\sum_{j=1}^{G}\left(R_{i,j}^{(k)} - \mu_i^{(k)}\right)^2}. \tag{6}$$

The corresponding normalized advantage is

$$A_{i,j}^{(k)} = \frac{R_{i,j}^{(k)} - \mu_i^{(k)}}{\sigma_i^{(k)} + \delta}, \tag{7}$$



(a) (b)

Figure 8: V-A distributions of edited endpoint images and sampled intermediate frames. The left panel shows that endpoint images form separated clusters across the four target quadrants, whereas the right panel shows that intermediate frames densely cover the central transition region around $(V = 3, A = 3)$. The marginal histograms show the corresponding distributions along the Valence and Arousal dimensions.

where δ is a small constant for numerical stability. When all responses within a group receive the same reward, the corresponding normalized advantages become zero and therefore produce no optimization signal for that objective.

This decoupled normalization prevents differences among factor reasoning and emotion assessment rewards from collapsing into an identical scalar advantage, allowing each objective to provide an independent relative learning signal.

Factor-Guided Advantage Aggregation. We use the factor reward before normalization to determine the reliability of the predicted affective factors. Specifically, the factor-guided gate for response $y_{i,j}$ is

$$g_{i,j} = \text{sigmoid}\left(R_{i,j}^{(\text{factor})} - \tau\right) = \frac{1}{1 + \exp\left[-\left(R_{i,j}^{(\text{factor})} - \tau\right)\right]}, \tag{8}$$

where τ is the gating threshold.

The independently normalized advantages are then aggregated as

$$\widetilde{A}_{i,j} = A_{i,j}^{(\text{format})} + A_{i,j}^{(\text{factor})} + g_{i,j}\left(A_{i,j}^{(V)} + A_{i,j}^{(A)}\right). \tag{9}$$

Therefore, the valence and arousal advantages receive larger weights when the model produces reliable factor reasoning, whereas inaccurate factor predictions suppress the corresponding emotion-level optimization signals.

Policy Optimization. After obtaining the final normalized advantage A_j^{final}, we optimize the policy using the standard GRPO objective:

$$\begin{aligned}\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{\substack{q\sim P(\mathcal{Q}),\\ \{o_j\}_{j=1}^{G}\sim\pi_{\text{old}}(\cdot|q)}}\Bigg[\frac{1}{G}\sum_{j=1}^{G}\min\Big(\rho_j(\theta)A_j^{\text{final}},\\ \text{clip}\left(\rho_j(\theta), 1-\epsilon_{\text{clip}}, 1+\epsilon_{\text{clip}}\right)A_j^{\text{final}}\Big) - \beta D_{\text{KL}}(\pi_\theta \,\|\, \pi_{\text{ref}})\Bigg],\end{aligned} \tag{10}$$

where

$$\rho_j(\theta) = \frac{\pi_\theta(o_j \mid q)}{\pi_{\text{old}}(o_j \mid q)} \tag{11}$$

Table 11: Dataset-specific affective and factor supervision. All V–A values are represented in the unified $[1, 5]$ space.

Dataset	V–A Supervision	Coarse-Grained Factor Supervision	Fine-Grained Dataset-Specific Factor Supervision
ContiEmo	Valence and Arousal are originally annotated in $[1, 5]$.	Importance ranking of seven emotional factors.	Weather condition, lighting intensity, and color richness.
EMOTIC	Both Valence and Arousal are mapped from $[1, 10]$ to $[1, 5]$.	Not available.	Emotional state.
FindingEmo	Valence is mapped from $[-3, 3]$ and Arousal from $[0, 6]$ to $[1, 5]$.	Not available.	Emotional state and decision factor.

is the importance ratio for the sampled response. Here, $P(\mathcal{Q})$ denotes the training-question distribution, G is the number of responses sampled for each question, and π_θ, π_{old}, and π_{ref} denote the trainable policy, rollout policy, and frozen reference policy, respectively. ϵ_{clip} controls the clipping range, while β controls the KL regularization.

C.4 Dataset-Specific Training Objectives

Since ContiEmo, EMOTIC, and FindingEmo provide different affective-value ranges and factor annotations, we adapt the training targets to the annotations available in each dataset. All models are independently trained on the corresponding training split without transferring labels across datasets. To maintain a unified output space, the original V–A labels of the public datasets are linearly mapped to $[1, 5]$. Given an original value $x \in [a, b]$, the normalized value is computed as

$$\mathcal{N}_{[a,b]\to[1,5]}(x) = 1 + 4\frac{x-a}{b-a}. \tag{12}$$

Accordingly, both Valence and Arousal in EMOTIC are transformed from $[1, 10]$ to $[1, 5]$. For FindingEmo, Valence is transformed from $[-3, 3]$ to $[1, 5]$, and Arousal is transformed from $[0, 6]$ to $[1, 5]$.

Table 11 summarizes the dataset-specific supervision. When a factor annotation is unavailable, its corresponding output JSON field and reward term are removed, and the output schema is adapted accordingly.

C.5 Attention Visualization Procedure

To visualize the image regions involved in affective reasoning, we extract the attention weights from the penultimate Transformer layer. Specifically, we select the attention from all generated response tokens to the visual tokens, excluding the input-prompt and special tokens. We average these weights across all selected query tokens and attention heads to obtain a holistic visualization of the visual evidence used throughout the affective reasoning process.

Let $\mathbf{A}^{(\ell,h)} \in \mathbb{R}^{T_q \times T_k}$ denote the attention matrix of head h in layer ℓ, where ℓ is the penultimate Transformer layer. For each visual token $m \in \mathcal{I}$, its aggregated attention score is computed as

$$s_m = \frac{1}{H|\mathcal{Q}|} \sum_{h=1}^{H} \sum_{q\in\mathcal{Q}} A_{q,m}^{(\ell,h)}, \qquad m \in \mathcal{I}, \tag{13}$$

where H is the number of attention heads, $\mathcal{Q}$ denotes the response tokens, and $\mathcal{I}$ denotes the visual-token positions.

The resulting visual-token scores are reshaped according to the spatial grid of the image tokens and resized to the original image resolution. Finally, we adopt min-max normalization and the normalized map is overlaid on the original image for visual comparison. The maps are used only for qualitative visualization and should not be interpreted as causal attribution.

D THE USE OF LARGE LANGUAGE MODELS

In this work, we use Large Language Models (LLMs) to aid and polish writing. We utilize LLMs to refine the quality of our manuscript by suggesting more precise terminology. Additionally, we use LLMs to optimize LaTeX templates for figures, tables, and mathematical expressions, reducing the time and effort required for typesetting complex layouts.

E LIMITATIONS AND FUTURE WORK

Despite improving human alignment, EmoInsight primarily captures aggregated affective patterns. It therefore cannot account for differences in how specific individuals perceive and interpret the same visual content. Such individual variation may arise from personal experience, cultural background, and affective preferences, making a single population-level prediction insufficient for personalized applications.

In future work, we plan to extend the current framework toward personalized visual emotion assessment. Specifically, user-level preference data and historical affective annotations could be used to adapt the model through parameter-efficient fine-tuning or preference alignment. This would enable the model to produce individual-specific reasoning and V-A predictions that more closely reflect each user's perception.