electronics

MDPI

Article

LLaVA-Emo: Interpretable Affective Image Stylization via Chain-of-Thought Reasoning

Kaichen Tang and Qi Xu *

College of Information Engineering, Shanghai Maritime University, Shanghai 201306, China
* Correspondence: qixu@shmtu.edu.cn

Abstract

Affective Image Stylization (AIS) converts an emotional intent into executable artistic visual styles. Existing methods are often limited to discrete label settings and provide limited interpretability of how target emotions are realized. We propose LLaVA-Emo, an interpretable AIS framework built on multimodal Chain-of-Thought (CoT) reasoning. Our method decouples generation into two structured outputs: <reasoning> provides visual–affective causal explanations grounded in the input image evidence, and <style_prompt> expresses actionable, renderer-ready style instructions that directly condition a frozen diffusion renderer. We constructed a training set by screening ArtEmis' sentiment interpretations and fine-tune LLaVA-1.5-7B with LoRA, where SFT mainly supervises the structured intermediate <reasoning> (and output format), while the true executability of <style_prompt> is enforced by our DPO stage via render-and-reward feedback. The rendering stage remains training-free, and we further apply DPO for preference optimization to align candidate outputs with both emotion fidelity and instruction executability. Experiments on the EmoEdit inference set demonstrate that LLaVA-Emo improves emotion alignment while providing stronger process interpretability.

Keywords: affective image stylization; interpretable generation; multimodal chain-of-thought; reasoning-to-prompt decoupling; diffusion rendering; direct preference optimization (DPO)

1. Introduction

"Art is a method of opening up areas of feeling rather than merely an illustration of an object."—Francis Bacon.

Visual art has long served as a primary vehicle for human emotional expression. Across diverse artworks, viewers experience resonance through a sophisticated interplay of color, lighting, and brushwork [1,2]. With the rapid progress of generative AI, Affective Image Stylization (AIS) has emerged as a specialized task in computer vision. Given a natural photograph (e.g., a street view or portrait) and a target emotion (e.g., "sadness" or "excitement"), an AIS model aims to generate a stylized image that preserves the semantic structure of the input while endowing it with artistic beauty and emotional depth.

Unlike Neural Style Transfer, which mainly focuses on texture migration, AIS must satisfy two requirements simultaneously: the stylization should precisely correspond to nuanced emotional intent rather than producing a generic artistic look, and the generation process should be logically grounded in artistic principles instead of relying on superficial color or texture overlays. However, many existing approaches simplify emotions into discrete labels and treat the generation process as a black box, see Figure 1, which weakens

Check for updates

Academic Editor: George Angelos Papadopoulos

Received: 18 May 2026
Revised: 8 June 2026
Accepted: 9 June 2026
Published: 13 June 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

both the fidelity of emotion control and the interpretability needed for verifying how a style realizes the desired affect.

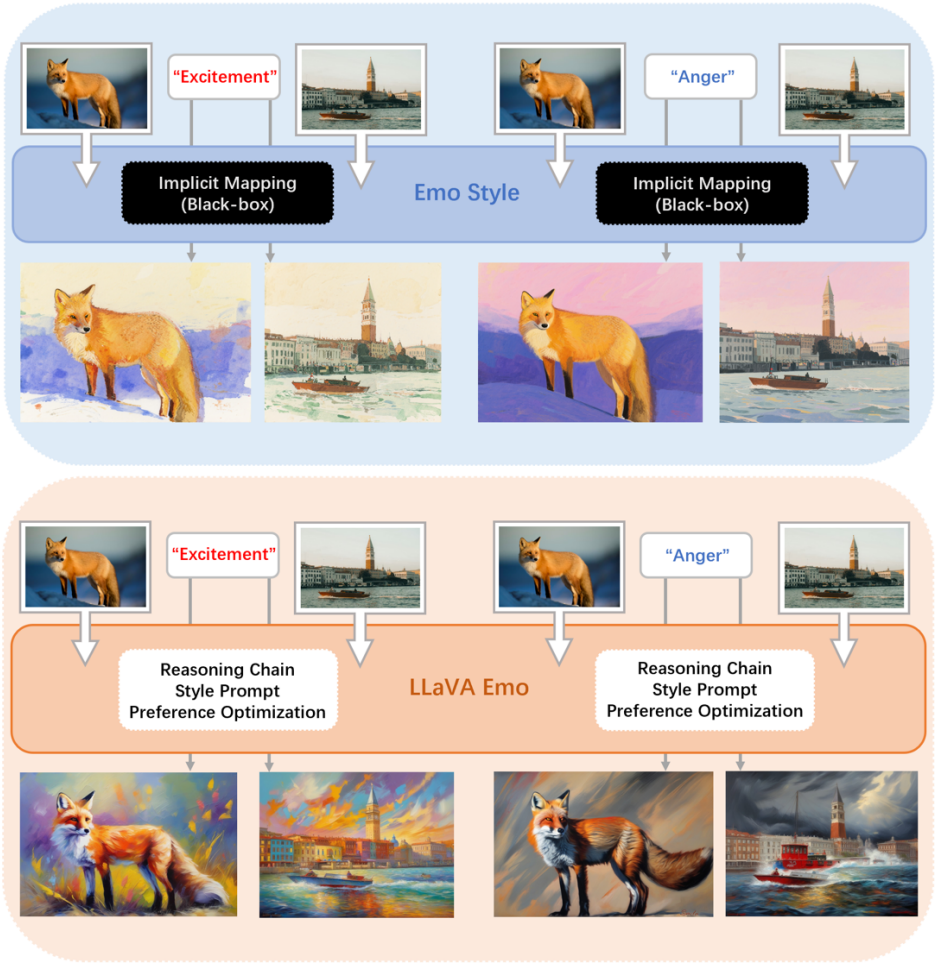


Figure 1. Interpretable AIS via reasoning-to-prompt decoupling. LLaVA-Emo first generates <reasoning> to explain the visual–emotion causal rationale, and then outputs <style_prompt> as an executable artistic instruction. The renderer is frozen and the diffusion process is driven by the style prompt. Preference optimization (DPO) further encourages emotion-matching and actionable instructions.

Existing AIS methods typically follow two technical trajectories, yet both face fundamental drawbacks. First, end-to-end frameworks learn an implicit mapping from affective labels to pixel distributions. However, this formulation provides limited interpretability: users can observe the final image but cannot understand the causal artistic logic that leads to the target emotion. Furthermore, constructing the required training triplets can introduce degradation: to decouple content from style, prior pipelines may rely on imperfect de-stylization procedures such as UnZipLoRA, which can compromise structural integrity.

Second, with the rise of large language models, zero-shot prompt engineering becomes a natural alternative. Nevertheless, applying general multimodal models to AIS reveals a domain-knowledge gap. Their outputs may hallucinate visual elements absent from the input or translate emotions into poetic but non-actionable descriptions. As a result, downstream renderers receive control signals that are not reliably grounded in the visual evidence of the input image. Finally, although large affective datasets like ArtEmis contain abundant human explanations, existing methods often discard the most valuable component—the reasoning process bridging visual elements to emotional experience—and

retain only coarse categorical labels. This creates a data bottleneck that limits the model's ability to produce controllable, verifiable, and explanation-consistent emotional stylization.

Figure 2 provides representative individual cases. To evaluate LLaVA-Emo, we follow the same inference protocol as EmoEdit: given a content image and an emotion condition, we generate stylized outputs and report a task-aligned quantitative comparison using a reduced metric set that covers semantic/content consistency, structural preservation, perceptual deviation, and emotion alignment (CLIP, DINO, LPIPS, Emo-A, and Emo-S). We further provide qualitative comparisons on representative samples under diverse emotion conditions.



Figure 2. Stylization results on diverse content. Performance of LLaVA-Emo on various subjects, including forests, architecture, and animals. The model demonstrates robust generalization across different image structures and emotional prompts.

Beyond the test inputs, our method can be applied to any user-provided content at inference time. We compare LLaVA-Emo against state-of-the-art style transfer, image editing, and affective image manipulation baselines, and validate its user-perceived effectiveness with a preference study that measures aesthetics, emotion fidelity, and their balance.

To verify that multimodal CoT is translated into actionable generation control rather than "pseudo-reasoning", we conduct targeted baselines and analyses based on reasoning-before-generation/editing paradigms (e.g., [3,4]). In addition, interpretability and causal-alignment evaluations explicitly connect the intermediate <reasoning> with the executable <style_prompt>, highlighting how the reasoning-to-prompt pipeline steers affective stylization through observable output changes.

In summary, our contributions are as follows.

1. Methodological innovation: reasoning-to-prompt decoupling. We reformulate AIS as a two-stage process. The model first generates explicit semantic reasoning to ground the visual–emotional causal rationale, and then produces an executable style instruction that directly conditions a frozen diffusion renderer. During training, SFT mainly supervises the structured <reasoning> (and its output structure), while the true executability of <style_prompt> is ensured through DPO with a render-and-reward objective. This decoupling improves interpretability and enables training-free control during rendering.
2. Explainable generation: from knowing what to knowing why. In addition to producing stylized images, the model provides art-theory-aligned explanations of how the target emotion is realized. This connects the intermediate reasoning with the resulting visual attributes, making the control signal more verifiable.

3. Domain expertise: parameter-efficient instruction refinement. We fine-tune a multimodal backbone with LoRA to obtain denser and less verbose style instructions, improving the utility of generated prompts for diffusion rendering without modifying the renderer.
4. System generality: plug-and-play transfer. Because the style instruction is decoupled from the renderer, the framework can be adapted to different diffusion backbones without retraining the reasoning module.

2. Related Work

2.1. Artwork Emotion Analysis

Artwork emotion analysis is a foundational pillar of Affective Computing. The existing literature has predominantly focused on "recognition and understanding" rather than "generation" For instance, leveraging large-scale art datasets such as WikiArt and ArtEmis [5], researchers have sought to learn fine-grained emotional representations and linguistic descriptions [5–7]. While these studies provide an essential "affective prior" for emotion-driven generation, a significant gap remains: most works prioritize emotion classification or captioning [8–10] instead of investigating how emotions can be procedurally translated into artistic styles during the image synthesis process. Motivated by this, we formulate Affective Image Stylization (AIS) as a controllable generation task steered by "text-level interpretable guidance".

2.2. Text-Driven Image Stylization and Editing

Text-driven stylization [11–14] and editing methods typically employ natural language as a condition to rewrite visual appearances. These methods generally fall into two categories: (1) reference-based style transfer or heavy-conditioned editing [15–17], and (2) zero-shot text-guided generation [14,18,19]. Despite their ability to produce aesthetically pleasing results, they face two primary challenges in affective tasks: first, the understanding of emotional requests remains at a coarse-grained semantic level; second, they lack an explicit mechanism to explain why specific visual alterations correspond to the target emotion, leading to insufficient interpretability of control signals. In contrast, we define affective stylization as a "Reasoning-before-Execution" task, where a Vision-Language Model (VLM) generates interpretable affective reasoning and actionable instructions, which are then rendered by a frozen diffusion model to ensure superior emotional alignment.

2.3. Affective Image Stylization and Manipulation

Affective Image Stylization (AIS) and Affective Image Manipulation (AIM) aim to evoke target emotions through visual transformations. Existing works typically model emotions in two ways: (1) discretizing emotions into fixed category labels and learning a mapping from "labels to style/pixels" [20,21], or (2) injecting textual affective constraints into generative models via affective mappers or condition vectors [22–25]. A representative state-of-the-art method, EmoStyle [26], models the coupling of emotion and content as an end-to-end generation process, further implementing interpretable emotion-style prototype learning through style quantization. However, end-to-end black-box mappings still struggle to explicitly explain the causal creative logic prior to generation, and the expressive capacity is often limited by the granularity of discrete categories. Consequently, we propose LLaVA-Emo, which upgrades affective stylization from "implicit mapping" to "explicit reasoning and decoupled execution". Our model generates not only the stylized output but also structured, interpretable reasoning (<reasoning>) and actionable style prompts (<style_prompt>), establishing a verifiable bridge between emotional intent and visual synthesis.

2.4. VLMs and Chain-of-Thought for Generation

Multimodal LLMs such as the LLaVA series and the CoT paradigm have been shown to improve both the accuracy and interpretability of multimodal reasoning tasks [27–29]. However, when CoT is not translated into an actionable intermediate representation, it may remain "pseudo-reasoning" rather than directly steering generation. MindOmni-style methods [3] further emphasize an intent or mind representation, but it often serves mainly as a conceptual guide.

Notably, GoT [4] also adopts a "reasoning-before-generation/editing" paradigm by generating explicit reasoning chains and using a diffusion model guided by semantic-spatial constraints. The key difference is that GoT primarily targets visual composition and spatial relations, whereas our approach decouples CoT into explanatory reasoning and style instructions that are explicitly rendered by a downstream diffusion model for affective stylization control, making the control both interpretable in language and verifiable through observable changes in the generated artwork.

3. Materials and Methods

3.1. Framework Architecture

As shown in Figure 3, our proposed framework, LLaVA-Emo, consists of four components: (1) an *Affective Art Director*, (2) a *Decoupled Output* with <reasoning> and <style_prompt>, (3) a *Training-Free Renderer*, and (4) *Emotion-driven Preference Optimization* (DPO). The core goal is to learn, during training, a mapping from (*image, emotion*) to (*reasoning, executable style instructions*), while keeping the pixel renderer frozen and fully controlled by the generated style prompt.

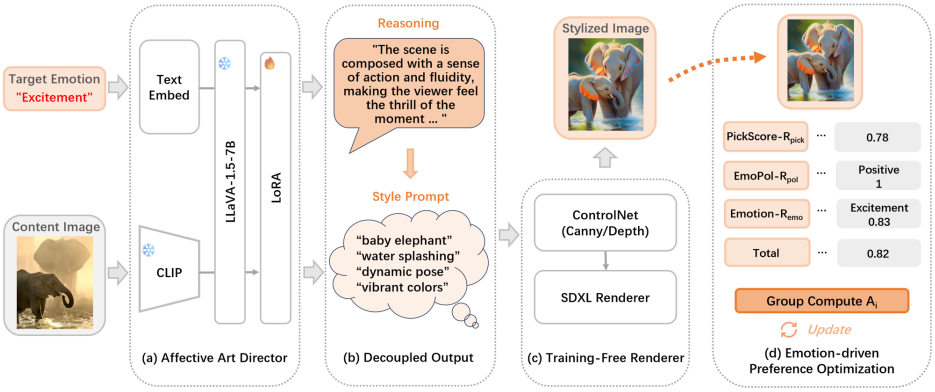


Figure 3. LLaVA-Emo learns (image, emotion) $\rightarrow$ ($\langle$reasoning$\rangle$, $\langle$style_prompt$\rangle$). The renderer is frozen and generates stylized images by executing the generated style prompt under a structural constraint. DPO improves the preference for higher-reward generations.

3.1.1. Input and Multimodal Encoding

Given a content image $X_v \in \mathbb{R}^{H\times W\times 3}$ and a target emotion description e (discrete emotion names), we first extract visual features using a CLIP vision encoder: $Z_v = E_v(X_v)$. The emotion text is encoded into a semantic embedding, and the multimodal context is formed by concatenating visual and emotional representations.

3.1.2. Affective Art Director: LLaVA with LoRA

We adopt LLaVA-1.5-7B as the backbone and perform parameter-efficient fine-tuning with LoRA. The trainable parameters θ_{LoRA} account for approximately 2% of the total parameters (with $r = 64$ and $\alpha = 128$), while the remaining backbone weights are frozen. Conditioned on the multimodal input, the model produces two structured outputs: (1) <reasoning>, which explains how visual cues (e.g., color, lighting, and composi-

tion) relate to the target emotional intent; and (2) <style_prompt>, which is converted into comma-separated visual keywords as an executable instruction for downstream diffusion rendering.

3.1.3. Training-Free Renderer

Unlike end-to-end training, we keep the rendering module entirely frozen. A spatial anchor is obtained from ControlNet under a structural condition $C(X_v)$ (e.g., Canny or Depth). The final stylized image is then generated by a pre-trained diffusion renderer using the generated style prompt and the structural constraint: $X_s = \text{Renderer}(\text{style_prompt}, C(X_v))$. This decoupled design enables seamless transfer when replacing the renderer backbone (e.g., SDXL to FLUX) without retraining the reasoning/prompt generation module.

3.1.4. Emotion-Driven Preference Optimization (DPO)

During training, we further optimize LLaVA-Emo using DPO to improve preference over sampled candidates. DPO is chosen because it directly optimizes multimodal generation via preference learning, rather than relying on independent token-level likelihood or contrastive losses that may not reflect rendering outcomes. In our language-to-renderer alignment setting, executability and emotion controllability are ultimately verified in the rendered images; DPO leverages render-and-reward feedback to update only the multimodal instruction generator toward candidates that causally yield better stylizations. This makes the optimization objective explicitly consistent with the downstream diffusion renderer behavior, strengthening the alignment between <style_prompt> and the generated artwork.

For each training instance (image and emotion), the model generates multiple candidate outputs, which are rendered into images. We compute a reward via task-aligned metrics (e.g., PickScore, EmoPol, Emo-A), and construct a preference pair (y^+, y^-), where y^+ is the higher-reward output (chosen) and y^- is the lower-reward output (rejected). DPO then updates only θ_{LoRA}, strengthening the causal consistency between <reasoning> and the executable <style_prompt>.

3.2. Dataset Construction

As shown in Figure 4, we screen Dataset Construction from ArtEmis through four sequential stages—consensus-based deduplication and one-image multi-emotion preservation, semantic filtering via comment–artwork embedding similarity, manual quality inspection on low-similarity subsets, and long-tail emotion balancing via reweighting—to obtain a reliable and emotion-balanced instruction dataset for interpretable AIS.

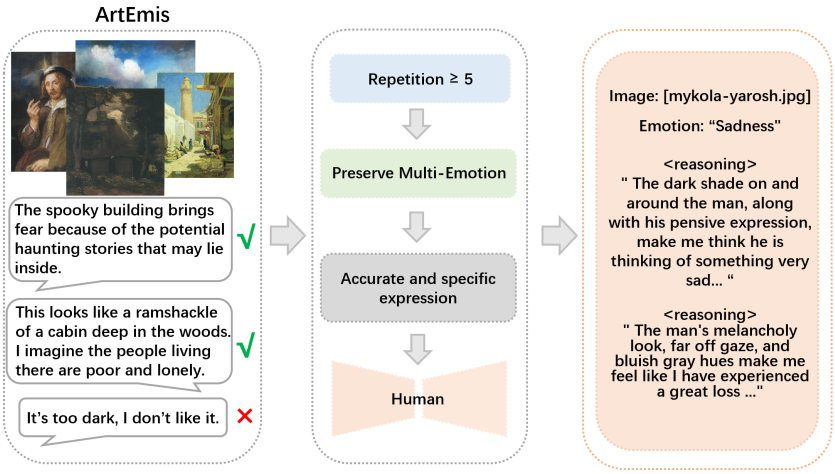


Figure 4. Construction pipeline of the dataset. We keep high-consensus annotations with `repetition` $\geq$ `5` and deduplicate by `(painting, emotion)` to preserve *one-image multi-emotion*.

3.2.1. Data Cleaning: Consensus Deduplication and One-Image Multi-Emotion

ArtEmis contains 454,684 usable comments (excluding the table header) over 80,031 artworks. We improve label reliability and reduce redundancy by: (i) consensus deduplication: we keep annotations with `repetition` $\geq$ `5` to filter out low-consensus and subjective noise. (ii) Multi-emotion preservation: instead of filtering by artwork alone (i.e., one-image one-emotion), we deduplicate by the key `(painting, emotion)`. This preserves multiple emotional views for the same artwork. After deduplication and reordering, we obtain around 80,000 filtered candidates for the next stage (the final training size is determined by subsequent cleaning and balancing).

3.2.2. Semantic Filtering: Keeping Semantically Faithful Comments

To improve training reliability, we perform semantic filtering by computing the cosine similarity between sentence embeddings of (a) human comments and (b) textual descriptions of the corresponding artworks. Specifically, for each candidate artwork, we first generate an English caption using BLIP, which serves as the artwork description for filtering. We then compute cosine similarity between the human comment and the BLIP caption, and keep samples with a similarity threshold of 0.4596. After this semantic filtering stage, we obtain 53,516 samples. For the retained set, the median cosine similarity is 0.6402, indicating strong semantic faithfulness. Low-similarity cases are discarded, especially when the comment is extremely short, ambiguous, or missing key elements that are not visually supported. We also mark the problematic subset during this stage for follow-up manual inspection.

3.2.3. Manual Sampling Inspection

We further conduct manual sampling inspection on the low-similarity subset (10%) to verify that the retained human comments remain aligned with the corresponding images and preserve emotional intent, rather than relying on template-like phrasing. After manual inspection, the number of retained samples becomes 53,151. Details of the manual inspection, as well as examples of retained and discarded samples, are provided in Appendix D.

3.2.4. Emotion Balancing: Long-Tail Reweighting

After semantic filtering and cleaning, the emotion distribution may still be imbalanced. We apply an add-only class balancing strategy by *reweighting* instead of generating additional samples. Specifically, we treat emotions whose proportion is below a threshold of 10% as minority classes and adjust the retention probability to control the number of retained samples per emotion. This ensures the ratio of numbers between emotional samples is as close to uniform as possible.

After reweighting, the final balanced dataset contains 50,330 training samples across eight emotions, with the following distribution: contentment (11,104; 20.08%), awe (7755; 15.41%), disgust (5629; 11.18%), excitement (5650; 11.23%), fear (5618; 11.16%), anger (5584; 11.09%), sadness (5072; 10.08%), and amusement (4918; 9.77%). Overall, the eight emotions sum to 50,330 samples (100%).

We further convert each retained sample into the standard LLaVA multimodal instruction format, where the image path and the comments field encode the target emotion and the corresponding supervision text from ArtEmis.

3.3. Training Strategy and Objective

3.3.1. Parameter-Efficient Fine-Tuning with LoRA

We train LLaVA-Emo to map the input pair $x = (X_v, e)$ into a structured output: $Y = [\langle \text{reasoning} \rangle; \langle \text{style_prompt} \rangle]$. The visual encoder and the multimodal projector are kept frozen. Only LoRA parameters are updated.

For an attention weight W_0, we apply low-rank adaptation

$$W = W_0 + \Delta W = W_0 + BA, \quad B \in \mathbb{R}^{d\times r},\ A \in \mathbb{R}^{r\times k}.$$

where d and k denote the input and output dimensions of the weight matrix, respectively, and r represents the LoRA rank. $A \in \mathbb{R}^{r\times k}$ is a matrix initialized with a random Gaussian distribution, and $B \in \mathbb{R}^{d\times r}$ is initialized to zero to ensure that ΔW is zero at the onset of training. In our implementation, we set the rank $r = 64$.

During SFT, we mainly supervise the model to generate the structured intermediate output <reasoning> that explains the visual–emotional causal rationale and follows our decoding template. Specifically, the human-authored comments within the ArtEmis dataset, which contain rich analysis and explanations of why an artwork evokes a specific emotion, serve as the direct supervision for this reasoning stage. The corresponding <style_prompt> is learned primarily in a format sense (i.e., producing renderer-ready instructions) during SFT, relying on the inherent zero-shot generative capabilities of the pre-trained LLaVA backbone. Its true executability and emotion controllability are subsequently enforced by our DPO stage: we render parsed candidates with a fixed diffusion renderer and compute rewards to form preference pairs, optimizing the model's core prompt-generation capabilities and ensuring that the final style prompts causally lead to the desired stylization outcomes.

3.3.2. Stage I: Self-Supervised Autoregressive SFT

We use the standard autoregressive language modeling objective and optimize the token likelihood in Y

$$\mathcal{L}_{\mathrm{SFT}}(\theta) = -\sum_{i=1}^{L} \log P_\theta(y_i \mid X_v, e, y_{<i}),$$

where θ denotes the trainable LoRA parameters, L is the total sequence length, X_v is the input content image, e is the target emotion condition, and y_i is the i-th token in the structured output Y. During SFT, we primarily supervise the generated <reasoning>; the loss on <style_prompt> tokens is masked (i.e., excluded from the gradient update). The term $y_{<i}$ represents all previous tokens before position i, and $P_\theta(\cdot)$ is the autoregressive token distribution predicted by the model.

This design keeps SFT focused on learning the reasoning structure and format, while the true executability and emotion controllability of <style_prompt> are enforced by our DPO stage through a render-and-reward objective.

3.3.3. Stage II: Emotion-Driven Preference Optimization (DPO)

After SFT, we apply *Emotion-driven Preference Optimization* (DPO) to improve preference toward candidates with both higher emotion fidelity and better generation quality. DPO updates only the LoRA parameters; the vision backbone, projection layers, and renderer remain frozen.

Reward Computation

For each rendered candidate, we compute a scalar reward from three components. Let x_i be the i-th rendered candidate image conditioned on (X_v, e_t), where e_t is the target emotion label of the painting. Let s_i^{emo} denote the classifier score (i.e., the softmax probability) of the target emotion e_t for image x_i, and let s_i^{pick} denote the PickScore similarity for that candidate.

We define the polarity component as *EmoPol*. Following Mikels' emotion model [30], we use eight discrete emotions and group them into positive polarity (amusement, awe, contentment, excitement) and negative polarity (anger, disgust, fear, sadness). For a target emotion e_t, let $\mathcal{E}_{\mathrm{same}}$ be the set of emotions with the same polarity as e_t, and let $\mathcal{E}_{\mathrm{opp}}$ be the set of emotions with the opposite polarity.

In our implementation, *EmoPol* corresponds to the continuous polarity score $\widehat{s}_i^{\mathrm{pol}}$ computed from the emotion classifier probabilities

$$\hat{s}_i^{\text{pol}} = \alpha \sum_{e \in \mathcal{E}_{\text{same}}} p(e \mid x_i) - \beta \sum_{e \in \mathcal{E}_{\text{opp}}} p(e \mid x_i), \tag{1}$$

where $p(e \mid x_i)$ are the softmax probabilities produced by the same ArtEmis-trained emotion classifier used for s_i^{emo}, and $\alpha = \beta = 1$.

The final per-candidate reward is

$$R_i = \omega_1 \hat{s}_i^{\text{emo}} + \omega_2 \hat{s}_i^{\text{pol}} + \omega_3 \hat{s}_i^{\text{pick}}, \tag{2}$$

where $\hat{s}_i^{\text{emo}}$ and $\hat{s}_i^{\text{pick}}$ are obtained by applying the same within-sample normalization $\text{Norm}(\cdot)$ to $\{s_i^{\text{emo}}\}_{i=1}^k$ and $\{s_i^{\text{pick}}\}_{i=1}^k$, respectively, using either min–max or z-score normalization. The polarity term $\hat{s}_i^{\text{pol}}$ (i.e., EmoPol) is not normalized across candidates. We set the reward weights to $\omega_1 = 1.0$, $\omega_2 = 0.5$, $\omega_3 = 0.2$ for $\hat{s}_i^{\text{emo}}$, $\hat{s}_i^{\text{pol}}$ and $\hat{s}_i^{\text{pick}}$. When any log-based polarity shaping is computed in auxiliary code paths, we use the same numerical constant $\varepsilon = 1 \times 10^{-12}$ for stability.

Rewards are computed on the rendered images, so they supervise the executability of the parsed <style_prompt> under the fixed renderer.

Preference Pair Construction and DPO Loss

We set y^+ to the candidate with the largest R_i and y^- to the candidate with the smallest R_i, or optionally the second-smallest R_i (`-reject_mode second_min`; default is `min`). DPO minimizes

$$\mathcal{L}_{\text{DPO}}(\theta) = -\log \sigma\Big(\beta\big[(\log \pi_\theta(y^+ \mid x) - \log \pi_\theta(y^- \mid x)) - (\log \pi_{\text{ref}}(y^+ \mid x) - \log \pi_{\text{ref}}(y^- \mid x))\big]\Big), \tag{3}$$

where σ is the logistic sigmoid, $\pi_\theta(\cdot \mid x)$ is the LoRA-updated policy, $\pi_{\text{ref}}(\cdot \mid x)$ is the reference policy, and β = 0.1, temperature = 0.7 by default.

3.4. Inference and Rendering

3.4.1. Two-Stage Generation: CoT Reasoning and Style Prompt

Given a content image X_v and a target emotion description e, LLaVA-Emo generates: (i) a structured <reasoning> (typically 15–20 tokens) and (ii) a structured <style_prompt> formatted as comma-separated visual keywords. We then parse the content inside the <style_prompt> tag using a regular expression and normalize it into the required comma-separated format.

3.4.2. Training-Free Renderer with Structural Anchoring

The rendering stage is fully frozen and decoupled from the language model. We first compute a structural condition $C(X_v)$ using ControlNet (Canny/Depth). We set the ControlNet scale to 0.5 to balance structure preservation and style stylization. The final stylized image is generated by a diffusion renderer (SDXL/FLUX) conditioned on the parsed style prompt

$$X_s = \text{Renderer}(\text{style_prompt}, C(X_v)).$$

We use diffusion sampling steps of 30 and a CFG scale of 7.0, producing the final output X_s.

4. Results

4.1. Dataset and Evaluation

4.1.1. Datasets

We evaluate LLaVA-Emo under the EmoEdit inference protocol [25]. Each test instance provides a content image and a target emotion label. For every (content, emotion) pair,

we generate a structured ⟨`style_prompt`⟩ with our model, parse it into executable visual directives, and render the final stylized image using a frozen diffusion generator with the same structural conditioning as in Section 3.4. We generate stylizations for eight target emotions based on Mikels' wheel [30], yielding 3240 outputs in total.

To verify generalization beyond the EmoEdit setting, we further conduct experiments on two additional official public datasets: ArtEmis [5] and EmoSet [31]. For both datasets, we maintained the same inference pipeline as EmoEdit. This allows us to assess whether the proposed model preserves content while improving emotion controllability across both artistic and photographic domains.

4.1.2. Evaluation Metrics

Since AIS is a new task, we adopt a complementary set of evaluation metrics to measure performance from three key aspects: content consistency, visual/style quality, and emotion alignment. Specifically, CLIP [28] quantifies semantic consistency between the input content and the stylized output, ensuring that strong stylization does not break the underlying meaning. DINO [32] captures structural and geometry-level preservation, discouraging solutions that achieve similarity mainly through superficial texture changes. LPIPS [33] further measures perceptual differences, providing an additional and more reliable view of how much the representation is altered at the feature level. For emotion control, Emo-A [5] is computed with an independent emotion classifier to directly assess whether the target affect is achieved, matching the task objective. Since AIS requires both preserving structure and evoking emotions, we use Emo-S [25] to evaluate the increase in predicted scores for the desired emotion type.

4.2. Implementation Details

Our experiments are implemented in PyTorch 2.1.2 with CUDA 12.1. All training and inference are conducted on a single machine equipped with two NVIDIA GeForce RTX 4090D GPUs.

We fine-tune the model using LoRA with a learning rate set to 2×10^{-4}. The batch size is 32, and we train for 3 epochs. We use the AdamW optimizer throughout, with standard default momentum parameters and weight decay as configured in our implementation. For all methods, the input content images are resized to 1024×1024 before being fed into the pipeline. After SFT, we perform DPO for preference optimization with a learning rate set to 5×10^{-5}, where the DPO temperature (beta) is fixed to 0.1. We train for 1 epoch with batch size 16, using AdamW as the optimizer. The same 1024×1024 resized inputs are used for all baselines to ensure a consistent evaluation protocol.

For a fair and rigorous comparison, we re-implemented all baseline methods under a unified protocol without altering their intrinsic architectural designs. Specifically, we standardized the rendering engine, sampling hyper-parameters, and prompt templates across all evaluated models; the comprehensive configurations are detailed in Appendix A.

4.3. Cross-Dataset Evaluation

To test the generalization of LLaVA-Emo beyond the EmoEdit setting, we evaluate the same inference protocol on three official affective image stylization datasets: ArtEmis, EmoSet, and EmoEdit. For each dataset, the model is conditioned on the content image and a target emotion label, and the stylized result is generated with the same diffusion generator and structural conditioning as in Section 3.4. Table 1 reports the quantitative results.

Across datasets, our model peaks on ArtEmis while showing slightly lower scores on EmoSet, confirming that transferring affective stylization to real photographs poses a greater domain shift challenge. EmoEdit delivers consistently robust performance, validating the cross-distribution efficacy of our emotion-aware styling mechanism. Notably,

while structural measures (CLIP/DINO/LPIPS) remain stable, emotion-specific scores exhibit greater sensitivity to domain variations. Consequently, to establish a standardized benchmark and perform comprehensive comparative evaluations against existing baselines, the subsequent experiments are conducted on the EmoEdit dataset.

Table 1. Cross-dataset quantitative comparison of affective image stylization. ↑ and ↓ indicate higher and lower values are preferred, respectively.

Dataset	CLIP ↑	DINO ↑	LPIPS ↓	Emo-A ↑	Emo-S ↑
ArtEmis	0.7645	0.6288	0.6012	0.3752	0.3540
EmoSet	0.7315	0.5912	0.6294	0.2985	0.263
EmoEdit	0.7594	0.6194	0.6173	0.3473	0.327

4.4. Comparisions

To evaluate LLaVA-Emo, we use EmoEdit public dataset as the test dataset and compare it against state-of-the-art methods across several categories of affective generation and control: (1) style transfer: LSAST [34] and InST [14]; (2) image editing: IP2P [35] and LEDITS++ [36]; (3) text-driven affective stylization and filtering: IDAST [11] and AIF [37]; (4) reasoning-before-generation approaches: GoT [4] and MindOmni [3]; (5) Affective Image Manipulation (AIM): EmoEditor [22] and EmoEdit [25]; and (6) Affective Image Stylization (AIS): EmoStyle [26].

4.4.1. Quantitative Comparisons

As shown in Table 2, LLaVA-Emo achieves a strong overall performance. It attains the best CLIP score (0.7594), suggesting effective semantic preservation, and the second-best DINO score (0.6194), indicating good structural and local visual consistency. More importantly, Emo-A reaches 0.3473 (the highest), reflecting the best emotion alignment among all methods. Note that LPIPS is not always the absolute minimum (0.6173 vs. 0.5159 for LEDITS++), and it remains higher than some non-emotion baselines. This is consistent with our design goal: we prioritize emotion alignment and semantic preservation, trading off the lowest perceptual deviation in favor of more reliable emotion expression.

Table 2. Quantitative comparison of affective image stylization. The best results are bolded, and the second best are underlined. ↑ and ↓ indicate higher and lower values are preferred, respectively.

Method	CLIP ↑	DINO ↑	LPIPS ↓	Emo-A ↑	Emo-S ↑
LSAST	0.7407	0.6140	0.6077	0.1313	0.016
InST	0.6873	0.4189	0.6885	0.1242	0.014
IP2P	0.7284	0.4928	0.5861	0.2336	0.232
LEDITS++	0.7328	0.5379	**0.5159**	0.1695	0.094
IDAST	<u>0.7514</u>	0.5183	<u>0.5465</u>	0.1195	0.041
AIF	0.7372	0.5895	0.6545	0.1266	0.069
GoT	0.7364	0.6049	0.5581	0.1508	0.083
MindOmni	0.7215	0.4590	0.6353	0.1820	0.126
EmoEditor	0.6367	0.3765	0.6797	0.1437	0.103
EmoEdit	0.6936	0.4502	0.5755	0.1859	0.165
EmoStyle	0.7102	**0.6615**	0.6436	<u>0.3278</u>	<u>0.274</u>
LLaVA-Emo (Ours)	**0.7594**	<u>0.6194</u>	0.6173	**0.3473**	**0.327**

4.4.2. Qualitative Comparisons

We present qualitative results in Figure 5. Style transfer baselines (e.g., LSAST, InST) can produce visually stylized artworks, but they often fail to consistently convey the target

emotion; across different emotions, the overall tone changes are limited, leading to emotionally ambiguous results. Image-editing approaches (e.g., IP2P, LEdits++) introduce plausible artistic elements, yet the emotion-to-style mapping remains less reliable, sometimes relying on a small number of coarse cues (such as global brightness or saturation) rather than emotion-specific style transformations.

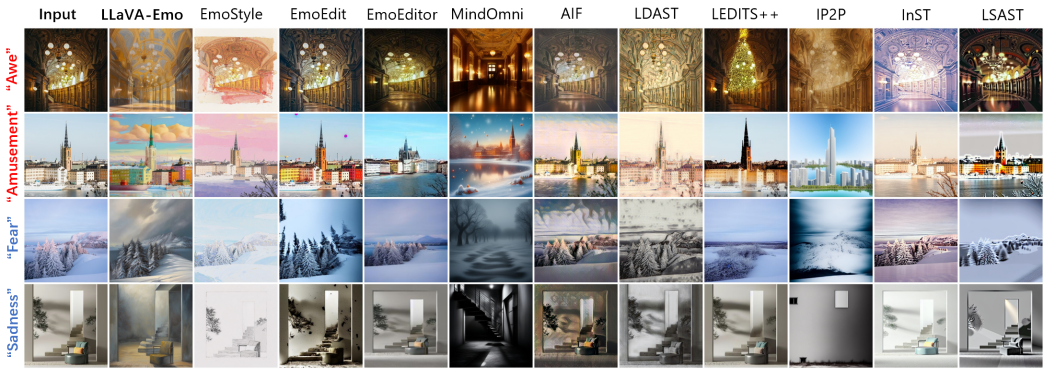


Figure 5. Comparison with the state-of-the-art methods, where LLaVA-Emo surpasses others on emotion fidelity and aesthetic appeal.

Affective image manipulation methods (e.g., EmoEdit, EmoEditor, AIF, LDAST, MindOmni) generally capture the coarse affective tendency (e.g., warmer vs. colder palettes), but the style differences across emotions can still appear subtle, and the generated stylization may drift away from the intended artistic mood. In contrast, our LLaVA-Emo produces more clearly separated, emotion-dependent visual characteristics while maintaining input semantic structure, yielding outputs that are both aesthetically appealing and emotionally evocative across all shown emotion conditions.

4.4.3. User Study

We conduct a user preference study to evaluate perceived aesthetics, emotion fidelity, and the balance between them. The study includes 67 participants with diverse ages and backgrounds, and each session lasts about 15 min. To ensure reliability, we randomized the panel positions and added trap questions; responses selecting trap options and completed in under 10 min were excluded. The significance test reports p-value $= 3.946 \times 10^{-12}$.

For each evaluation set, participants view stylized results from InST, AIF, EmoStyle, and our LLaVA-Emo under the same content input and target emotion. Participants answer three forced-choice questions: Aesthetic Perception, Emotion Fidelity, and Balance. Vote shares are reported as mean $\pm$ standard deviation.

As shown in Table 3, LLaVA-Emo achieves the highest preference across all aspects: 69.10% for Aesthetic Perception, 54.91% for Emotion Fidelity, and 67.72% for Balance. EmoStyle ranks second in all questions, while InST and AIF receive lower vote shares, supporting that grounded, instruction-executable generation better matches human perception.

Table 3. User preference study. The numbers show the percentage of participants who voted for the result. ↑ indicate higher values are preferred.

Method	Aesthetic Perception ↑	Emotion Fidelity ↑	Balance ↑
InST	2.40 ± 4.26%	15.63 ± 3.56%	1.34 ± 5.53%
AIF	9.08 ± 9.42%	5.09 ± 2.22%	7.76 ± 9.91%
EmoStyle	19.42 ± 17.24%	24.37 ± 5.57%	23.18 ± 13.48%
LLaVA-Emo (Ours)	69.10 ± 18.16%	54.91 ± 4.35%	67.72 ± 15.68%

4.4.4. Renderer Comparison

We replace the diffusion renderer with FLUX.1 and SDXL, keeping a constant seed of 42, a guidance scale of 7.0, 30 inference steps and the same prompt in all experiments. This evaluation is performed under the same EmoEdit inference inputs and emotion conditions. Table 4 show that LLaVA-Emo remains effective across different renderers, with performance depending on the renderer. Overall, this supports its *plug-and-play* property without retraining.

Table 4. Quantitative comparison of base generative models on affective image synthesis. We evaluate models across semantic alignment (CLIP), structural consistency (DINO, LPIPS), and emotional accuracy (Emo-A, Emo-S). ↑ and ↓ indicate higher and lower values are preferred, respectively.

Method	CLIP ↑	DINO ↑	LPIPS ↓	Emo-A ↑	Emo-S ↑
FLUX.1	0.7344	0.5396	0.6530	0.3266	0.318
SDXL	0.7594	0.6194	0.6173	0.3473	0.327

4.5. Ablation Study: Verifying the Necessity of Key Components

Table 5 validates the necessity of each component under consistent rendering and sampling.

The ablation study results in Figure 6 confirm the complementary roles of our core components. Without domain fine-tuning (generic_llava), the model lacks AIS-specific reasoning, yielding the weakest emotion control and degraded structural consistency. Removing inference-time reasoning (wo_reasoning) by forcing the model to directly output comma-separated style prompts significantly harms affective alignment; without reasoning-grounded cues, the generated keywords become less accurate. When bypassing LLaVA entirely at conditioning time (wo_style_prompt) and using only the raw emotion label, the lack of image-specific textual content relies too heavily on lexical priors, weakening the stylistic transformation.

Table 5. Ablation study on the core components of LLaVA-Emo. ↑ and ↓ indicate higher and lower values are preferred, respectively. The best results are bolded, and the second best are underlined.

Method	CLIP ↑	DINO ↑	LPIPS ↓	Emo-A ↑	Emo-S ↑
generic_llava	0.7036	0.4757	0.6678	0.1120	0.031
wo_reasoning	0.7516	0.6097	**0.6156**	0.1650	0.188
wo_style_prompt	0.7097	0.5615	0.6381	<u>0.2219</u>	<u>0.240</u>
wo_dpo	<u>0.7552</u>	<u>0.6145</u>	<u>0.6167</u>	0.1950	0.137
wo_controlnet	0.7148	0.4387	0.6531	0.2059	0.142
wo_dpo & wo_controlnet	0.7105	0.4295	0.6612	0.1420	0.095
LLaVA-Emo	**0.7594**	**0.6194**	0.6173	**0.3473**	**0.327**



Figure 6. Ablation on LLaVA-Emo components. Reasoning, Style_prompt, DPO, and ControlNet are shown to be necessary.

Furthermore, disabling preference optimization (wo_dpo) eliminates the render-and-reward supervision that ensures prompt executability, making the instructions less causally effective for emotion-faithful stylization. Operating without structural anchoring (wo_controlnet) causes severe degradation in spatial preservation, as the style drifts away from geometric constraints. Finally, a joint ablation removing both preference alignment and structural guidance (wo_dpo & wo_controlnet) triggers a severe performance bottleneck. The pipeline simultaneously loses its optimization path for affective precision and its geometric anchoring, confirming that these two modules serve as orthogonal pillars for semantic and structural sanity. Overall, these results validate that SFT-guided reasoning and DPO-enforced instruction control are both indispensable for high-fidelity affective stylization.

4.6. Interpretability and Causal Alignment Analysis

We validate interpretable and causally consistent generation by evaluating (i) hierarchical explainability metrics between reasoning and visual realization, (ii) internal semantic coupling between the intermediate reasoning <reasoning> and executable style prompt <style_prompt>, and (iii) counterfactual directional consistency under emotion polarity flips.

4.6.1. Hierarchical Explainability Metrics

We compare our method with GLM-5 to examine whether instruction generation becomes more actionable when specialized for an affective stylization control pipeline, rather than relying on a top-level general-purpose vision-language model. The results are shown in Table 6.

Table 6. Quantitative comparison of reasoning quality and generation diversity. ↑ indicate higher values are preferred. The best results are bolded.

Method	Actionability ↑	Seg/Token ↑	TTR ↑	Input Alignment ↑
GLM-5	0.3236	0.199	0.765	**0.5240**
LLaVA-Emo (Ours)	**0.5175**	**0.517**	**0.983**	0.4593

(1) Input Alignment (A_{in}).

We compute A_{in} to measure whether the cues explicitly mentioned in the generated <reasoning> are actually realized in the rendered image. This rules out cases where the model produces visually generic explanations that do not correspond to observable low-level visual properties. First, we extract a cue set $C(R)$ from the reasoning text by matching tokens in a small attribute lexicon $\mathcal{A}$. Only cues explicitly mentioned in R are included, hence $|C(R)| = |C|$. For each cue $c \in C$, we define a deterministic predicate $\text{pass}(c) \in \{0, 1\}$ based on low-level statistics computed on the rendered image $\hat{I}$ (all values are normalized to $[0, 1]$). The alignment score is

$$A_{in} = \begin{cases} 0.5, & \text{if } C = \emptyset, \\ \frac{1}{|C|} \sum_{c \in C} \text{pass}(c), & \text{otherwise.} \end{cases}$$

Low-level features used by $\text{pass}(\cdot)$.

Given $\hat{I}$, we compute normalized RGB channel means $\mu_R, \mu_G, \mu_B \in [0, 1]$. After converting RGB to HSV (with $h, s, v \in [0, 1]$), we compute global mean saturation and value

$$\mu_S(\hat{I}) = \mathbb{E}_p[s(p)], \qquad \mu_V(\hat{I}) = \mathbb{E}_p[v(p)].$$

To quantify "cool/warm", we define hue distances to red and yellow

$$d_R(h) = \min(|h|,\ |h-1|), \qquad d_Y(h) = \left|h - \tfrac{1}{6}\right|.$$

We convert distances into a clipped score

$$\phi(h) = \mathrm{clip}_{[0,1]}\left(1 - \frac{\min(d_R(h), d_Y(h))}{0.25}\right), \qquad W(\hat{I}) = \mathbb{E}_p[\phi(h(p))].$$

For "muted/monochrome" we also define a simple RGB spread

$$D(\hat{I}) = |\mu_R - \mu_G| + |\mu_G - \mu_B| + |\mu_R - \mu_B|.$$

Per-cue pass rules.

We then use indicator predicates $\mathbf{1}[\cdot]$

$$\begin{aligned}
\mathrm{pass}(\textit{dark}) &= \mathbf{1}[\mu_V(\hat{I}) \le 0.45],\\
\mathrm{pass}(\textit{bright}) &= \mathbf{1}[\mu_V(\hat{I}) \ge 0.55],\\
\mathrm{pass}(\textit{cool}) &= \mathbf{1}[W(\hat{I}) \le 0.45],\\
\mathrm{pass}(\textit{warm}) &= \mathbf{1}[W(\hat{I}) \ge 0.55],\\
\mathrm{pass}(\textit{muted}) &= \mathbf{1}[\mu_S(\hat{I}) \le 0.35],\\
\mathrm{pass}(\textit{vibrant}) &= \mathbf{1}[\mu_S(\hat{I}) \ge 0.45],\\
\mathrm{pass}(\textit{monochrome}) &= \mathbf{1}\big[D(\hat{I}) \le 0.25 \ \wedge\ \mu_S(\hat{I}) \le 0.30\big].
\end{aligned}$$

where R is the generated <reasoning>; $C(R)$ is the extracted cue set with size $|C|$; $\hat{I}$ is the rendered image. μ_R, μ_G, μ_B are normalized mean RGB channels; $\mu_S(\hat{I})$ and $\mu_V(\hat{I})$ are global mean saturation and value in HSV; $h, s, v \in [0,1]$ are per-pixel HSV components, and $\mathbb{E}_p[\cdot]$ denotes expectation over pixels. $d_R(h)$ and $d_Y(h)$ measure hue proximity to red and yellow, $\phi(h)$ converts hue distance into a clipped score, and $W(\hat{I})$ aggregates it across pixels. $D(\hat{I})$ measures RGB spread as a proxy for "muted/monochrome". $\mathbf{1}[\cdot]$ is the indicator function, returning 1 if the condition holds and 0 otherwise.

(2) Actionability (A_{out})

We compute actionability to measure whether the intermediate <reasoning> produces *visually verifiable* instructions. Specifically, if a term mentioned in the reasoning is truly actionable, it should be recognized as present in the generated image by an independent multimodal evaluator. From R, we extract a term set $T = \{t_k\}_{k=1}^{K}$. For each generated image, we query two multimodal evaluators, Kimi-K2.5 and GPT-4o, using a strict JSON protocol. Each evaluator outputs a binary decision $\mathrm{present}^{(m)}(t_k) \in \{0,1\}$ indicating whether term t_k is visually realized in the image, where $m \in \{\mathrm{Kimi}, \mathrm{GPT4o}\}$. We average the binary decisions over both terms and evaluators

$$A_{out} = \frac{1}{2K} \sum_{k=1}^{K} \sum_{m \in \{\mathrm{Kimi,GPT4o}\}} \mathbf{1}\Big[\mathrm{present}^{(m)}(t_k)\Big].$$

where K is the number of extracted terms, t_k is the k-th term, m indexes the evaluator model, and $\mathrm{present}^{(m)}(t_k)$ denotes whether term t_k is visually realized according to evaluator m. A higher A_{out} indicates more actionable and visually grounded terms in <reasoning>.

(3) Instruction Density and Diversity (on P)

We use instruction density and diversity to verify that <style_prompt> is not only interpretable, but also information-efficient. If the prompt contains more actionable content per unit length and avoids excessive repetition, it is more likely to provide strong and stable guidance to the diffusion renderer. Let P be the generated <style_prompt>. We parse P as a comma-separated list of segments. Let N_{seg} denote the number of non-empty

comma-separated segments and let N_{tok} denote the token count of P. We define instruction density as

$$\rho_{seg} = \frac{N_{seg}}{N_{tok}}.$$

We also measure lexical diversity by counting the number of unique tokens in P. Let N_{type} be the number of unique tokens and define the type–token ratio

$$\mathrm{TTR} = \frac{N_{type}}{N_{tok}}.$$

where P is the generated style prompt, N_{seg} counts non-empty comma-separated segments, N_{tok} is the token length of P, and N_{type} is the number of distinct tokens appearing in P. A higher ρ_{seg} indicates more instruction content per token, while a higher TTR indicates fewer repetitions and richer lexical coverage.

4.6.2. Internal Semantic Coupling

We verify that the intermediate <reasoning> is semantically coupled with the executable <style_prompt> at the sample level (rather than being independent or template-driven). For each sample i, let R_i and P_i denote the generated <reasoning> and <style_prompt>, respectively. We embed them using $f(\cdot)$: $\mathbf{v}(R_i) = f(R_i)$ and $\mathbf{v}(P_i) = f(P_i)$. Paired semantic similarity is computed as

$$S_{ca}^{\text{paired}} = \frac{1}{N}\sum_{i=1}^{N} \cos\big(\mathbf{v}(R_i), \mathbf{v}(P_i)\big).$$

where N is the number of evaluated paired samples, $\mathbf{v}(\cdot)$ denotes the embedding produced by $f(\cdot)$, and $\cos(\cdot,\cdot)$ is the cosine similarity between two embeddings. The gap between S_{ca}^{paired} and S_{ca}^{shuffle} represents how strongly the reasoning and prompt are semantically coupled at the sample level. To obtain a chance baseline, we shuffle $\{P_i\}_{i=1}^{N}$ across samples while keeping $\{R_i\}_{i=1}^{N}$ fixed, yielding S_{ca}^{shuffle}.

Across $N = 3240$ valid samples, we obtain $S_{ca}^{\text{paired}} = 0.5413$ and $S_{ca}^{\text{shuffle}} = 0.403$. A paired t-test on per-sample cosine values (two-sided) yields $t = 30.577$ and p-value $= 1.398 \times 10^{-154}$, supporting that $S_{ca}^{\text{paired}} > S_{ca}^{\text{shuffle}}$ and that the paired reasoning–prompt semantics are not independent.

Keyword-Level Attribute Alignment

We further test whether the semantic connection appears in concrete artistic attributes. Using an art lexicon L (size $|L| = 255$), we extract keyword sets $A_R(i) \subseteq L$ from R_i and $A_P(i) \subseteq L$ from P_i, and compute Jaccard overlap

$$J_i = \mathrm{Jaccard}\big(A_R(i), A_P(i)\big) = \frac{|A_R(i) \cap A_P(i)|}{|A_R(i) \cup A_P(i)|}.$$

where L is the art lexicon, and $A_R(i)$ and $A_P(i)$ are the matched keyword sets extracted from R_i and P_i, respectively. A higher S_{attr} indicates that artistic attributes referenced in reasoning are systematically transcribed into the corresponding executable prompt.

The mean attribute alignment is $S_{\text{attr}} = \frac{1}{N}\sum_{i=1}^{N} J_i$. We obtain $S_{\text{attr}} = 0.2088$ with standard deviation 0.1812, indicating that reasoning-level artistic attributes are systematically transcribed into instruction-level keywords.

4.6.3. Image-Dependency Validation (SSI): "Look-At-Image" Reasoning

We compute SSI to rule out template-based "pseudo-reasoning" that would generate nearly identical explanations for the same emotion across different images. If reasoning is

truly image-dependent, then within the same emotion, reasoning texts should vary across images, leading to a lower average self-cosine similarity. Let $\mathcal{E}$ be the set of evaluated emotions. For each emotion $e \in \mathcal{E}$, consider n_e different images under that emotion, and let $R_{e,a}$ denote the reasoning text for the a-th image. We embed reasoning with $f(\cdot)$ using $\mathbf{v}(\cdot)$, and compute within-emotion average pairwise cosine similarity

$$SSI_e = \frac{2}{n_e(n_e-1)} \sum_{1 \le a < b \le n_e} \cos\big(\mathbf{v}(R_{e,a}), \mathbf{v}(R_{e,b})\big).$$

We then average across emotions

$$SSI_{\mathcal{E}} = \frac{1}{|\mathcal{E}|} \sum_{e \in \mathcal{E}} SSI_e.$$

where $\mathcal{E}$ is the emotion set, n_e is the number of valid images for emotion e, and $R_{e,a}$ is the reasoning text corresponding to image a under emotion e. Lower $SSI_{\mathcal{E}}$ indicates higher diversity and stronger image dependency (vs. a template baseline where the self-cosine is near 1).

Using the sentence embedding model `all-MiniLM-L6-v2`, we obtain $SSI_{\mathcal{E}} = 0.6554$. The template self-cosine control is control_self_cos_mean ≈ 1.000. We further perform a one-sample t-test with $H_0 : \mathbb{E}[SSI_{\mathcal{E}}] = 1$ (valid $n = 8$ emotions). The test yields $t = -12.682$ and $p = 2.19 \times 10^{-6}$, indicating that $SSI_{\mathcal{E}} < 1$ is statistically significant. This supports that the model's reasoning varies with image context rather than collapsing to identical templates.

4.6.4. Causal Alignment Under Counterfactual and Perturbation Tests

To systematically investigate the *causal* sensitivity of the ⟨reasoning⟩ → ⟨style_prompt⟩ pipeline to emotional dynamics, we establish a formal quantitative framework within the text embedding space. Specifically, we define a directional agreement metric based on textual perturbations to serve as the analytical foundation for subsequent counterfactual and ablation tests. Given a baseline pair $(R_{\text{orig}}, P_{\text{orig}})$ and its corresponding perturbed counterpart $(R_{\text{pert}}, P_{\text{pert}})$ generated under specific counterfactual or altered conditions, we map them into the representation space via a frozen text embedder $f(\cdot)$. We then formalize the semantic steering behaviors by constructing intermediate displacement vectors

$$\Delta\mathbf{v}_R = \mathbf{v}(R_{\text{pert}}) - \mathbf{v}(R_{\text{orig}}), \quad \Delta\mathbf{v}_P = \mathbf{v}(P_{\text{pert}}) - \mathbf{v}(P_{\text{orig}}).$$

The alignment and directional agreement between these two decoupled stages are evaluated via cosine similarity

$$\text{CosSim}(\Delta\mathbf{v}_R, \Delta\mathbf{v}_P) = \frac{\Delta\mathbf{v}_R \cdot \Delta\mathbf{v}_P}{\|\Delta\mathbf{v}_R\|\|\Delta\mathbf{v}_P\|}.$$

To capture the macro-level structural tendency across the evaluation set, we compute a consistency indicator $X = \mathbf{1}\big[\text{CosSim}(\Delta\mathbf{v}_R, \Delta\mathbf{v}_P) > 0\big]$ and define the overall Consistency Rate (CR) as

$$\text{CR} = \frac{1}{M} \sum_{i=1}^{M} X_i,$$

where M denotes the number of valid evaluation pairs, $\cdot$ represents the dot product, and $\|\cdot\|$ denotes the ℓ_2 vector norm. A higher CR signifies that intentional textual variance induces synchronized, unidirectional semantic updates across both rationales and prompts. Based on this generalized formulation, we execute three rigorous causal perturbation experiments.

Experiment 1: Counterfactual Style_Prompt Editing

We directly edit the intermediate <style_prompt> by swapping it with its opposite emotional polarity counterpart while keeping the downstream diffusion renderer fixed. This tests whether the final affective outcome is causally governed by the style realizations embedded in the prompt. Under this counterfactual editing, the fine-grained emotion fidelity drops significantly to Emo-A $= 0.2750$. This performance degradation indicates that polarity inversion in the middle representation severely degrades target emotional alignment, confirming the direct causal effect of <style_prompt> on the final image rendering.

Experiment 2: Reasoning-to-Prompt Shuffle Under Polarity Inversion

To verify whether the cross-stage steering direction depends on the exact semantic coupling between stages, we perform a shuffle baseline under full counterfactual emotion flips. Here, the perturbed state $(R_{\text{pert}}, P_{\text{pert}})$ is constructed by pairing the opposite-polarity reasoning R_{opp} with a mismatched opposite-polarity style prompt P_{opp} randomly shuffled within each evaluation batch. Utilizing the formulation defined above, the directional agreement completely collapses to $\text{CosSim}(\Delta\mathbf{v}_R, \Delta\mathbf{v}_P) = 0.0419$. This near-zero agreement demonstrates that the directional consistency is destroyed when the intermediate step-by-step coupling is broken, reinforcing the causal necessity of our paired inference.

Experiment 3: Removing Explicit Emotion Words Before Embedding Evaluation

Finally, we rule out evaluation leakage by stripping explicit emotion keywords from the text inputs before embedding extraction, while keeping the rendering pipeline unchanged. In this setup, the perturbed texts $(R_{\text{pert}}, P_{\text{pert}})$ represent the emotion-token-ablated variants of the original strings. After removing these explicit tokens, we recompute the causal alignment statistics and observe a minor drop in directional agreement ($\Delta\,\text{CosSim} = -0.0151$, $p = 3.17 \times 10^{-3}$). This result indicates that while the absolute alignment signal naturally weakens without explicit anchors, the underlying semantic steering direction remains highly consistent, proving that the causal alignment is not merely a trivial artifact of duplicate emotion words.

Overall, the three independent evaluations—counterfactual editing, coupling shuffle, and token ablation—yield highly consistent empirical trends: counterfactual editing degrades final fidelity, shuffling destroys cross-stage alignment, and removing explicit words preserves the core steering vector. Together, these findings provide compelling evidence that our decoupled pipeline possesses non-trivial causal sensitivity to emotional dynamics rather than exhibiting weak or spurious statistical correlations.

4.7. Robustness to Environmental Degradations

To verify real-world adaptability under input quality degradations (e.g., haziness and blur), we evaluate our framework's interoperability with advanced image restoration preprocessors, specifically IHDCP [38] and WBPGNDN [39]. Table 7 reports the quantitative performance gains (Δ) over the unenhanced degraded baseline.

Table 7. Quantitative performance gains (Δ) of stylization robustness via advanced restoration preprocessing over the degraded baseline. ↑ and ↓ indicate higher and lower values are preferred, respectively.

Condition	ΔCLIP ↑	ΔDINO ↑	ΔLPIPS ↓	ΔEmo-A ↑	ΔEmo-S ↑
`Clean` (pristine input)	+0.029	+0.205	−0.090	+0.025	+0.00010
+ `IHDCP`	+0.010	+0.090	−0.039	+0.037	+0.00011
+ `WBPGNDN`	+0.008	+0.043	−0.024	+0.038	+0.00017

The empirical data demonstrates that integrating either IHDCP [38] or WBPGNDN [39] substantially recovers semantic structures from environmental filtering ($\Delta\text{CLIP} \geq +0.008$, $\Delta\text{LPIPS} \leq -0.024$), effectively elevating corrupted inputs to facilitate stable downstream processing. Crucially, the contrast-enhancement and sharpening mechanisms intrinsic to these advanced restoration networks inherently amplify affective visual cues, leading to solid improvements in fine-grained emotional metrics (ΔEmo-A and ΔEmo-S).

5. Conclusions

5.1. Discussion

We propose LLaVA-Emo for Affective Image Stylization, where multimodal Chain-of-Thought is used as an explicit intermediate step rather than an implicit end-to-end mapping. Our framework decouples explanatory reasoning from executable $\langle$`style_prompt`$\rangle$, and renders the final stylization with a diffusion-based generator to ensure that the control signal is both interpretable and verifiable through observable output changes. During training, SFT mainly supervises the structured <reasoning> and output structure, while the true executability of <style_prompt> is ensured by our DPO stage using a render-and-reward objective. Experiments and analyses demonstrate that the reasoning-to-prompt pipeline enables more emotion-consistent artistic transformations while preserving semantic/structural fidelity.

5.2. Limitations

LLaVA-Emo still depends on the quality of intermediate reasoning and the ability of the generated style prompt to be faithful to the target emotion. Moreover, when the input intent is ambiguous or the visual description lacks key details, the reasoning step may still generate fluent directives, which can lead to unreliable or underspecified style prompts and propagate errors into the final stylization, causing the output style to drift.

In addition, although controllability across the evaluation emotions is strong, the current setting still uses a finite evaluation emotion set and may not fully guarantee robustness for rare or fine-grained affective concepts that are not covered during training. Finally, the multi-stage inference process (reasoning, prompt execution, and evaluation) introduces extra computational overhead compared with single-pass baselines, which may limit real-time deployment.

Author Contributions: Conceptualization, K.T.; methodology, K.T.; software, K.T.; validation, K.T.; formal analysis, K.T.; investigation, K.T.; resources, K.T.; data curation, K.T.; writing—original draft preparation, K.T.; writing—review and editing, Q.X.; visualization, K.T.; supervision, Q.X. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The model weights and demo code used in this study are openly available in Zenodo at https://zenodo.org/records/20268259 (accessed on 2 June 2026). The dataset used in this study is a subset of the publicly available Artemis dataset, which can be accessed at https://artemisdataset.org/ (accessed on 2 June 2026).

Acknowledgments: During the preparation of this work, the authors utilized several generative AI tools to assist in specific stages of the research pipeline. Specifically, a fine-tuned LLaVA-1.5-7B model was employed to generate reasoning and style prompts. Image rendering was performed using Stable Diffusion XL (sdxl-base-1.0) and FLUX.1 based on the generated text prompts. Additionally, Kimi-k1.5 and GPT-4o were utilized to participate in the quantitative evaluation of the generated indicators. Following the use of these tools, the authors carefully reviewed, verified, and edited the outputs, and assume full responsibility for the accuracy and integrity of the final manuscript.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A

Table A1. Detailed sampling and rendering configurations for the baseline framework.

Hyperparameter Component	Exact Value/Configuration
Base generator	Stable diffusion XL (SDXL) base 1.0
Control network	ControlNet (Canny)
Sampling steps	`steps` = 30
Classifier-free guidance	`guidance_scale` = 7.0
Random seed control	`seed` = 42
ControlNet conditioning scale	`controlnet_scale` = 0.5
Canny hysteresis thresholds	`canny_low` = 100, `canny_high` = 200

Appendix B

To provide a transparent justification for our multi-objective DPO reward hyperparameter selection, we execute a rigorous two-stage exploration protocol. We extract a representative subset of 600 training pairs as a proxy search domain. We then evaluate the parameter sensitivity under a single-variable incremental strategy, measuring performance on an independent validation set of 100 ArtEmis samples completely disjoint from any training splits. As compiled in Table A1, setting ($\omega_1 = 1.0$, $\omega_2 = 0.0$, $\omega_3 = 0.0$) benchmarks the raw multi-modal baseline. Introducing coarse-grained sentiment constraints ($\omega_2 = 0.5$) systematically enhances global emotional orientation but yields marginal structural improvements. Crucially, incorporating a moderate text-to-image executability reward ($\omega_3 = 0.2$) establishes a Pareto-optimal equilibrium across all five key metrics. Pushing the aesthetic weight further to $\omega_3 = 0.4$ inherently penalizes fine-grained emotional categorization, as the optimization shifts towards generic, non-artistic semantic structures to satisfy the rigid aesthetic prior. Therefore, the empirical configuration ($\omega_1 = 1.0, \omega_2 = 0.5, \omega_3 = 0.2$) is chosen as our core alignment objective.

Table A2. Hyperparameter sensitivity search on the proxy training domain. The selected optimal combination is highlighted in bold. ↑ and ↓ indicate higher and lower values are preferred, respectively. The best results are bolded, and the second best are underlined.

Reward Weights ($\omega_1, \omega_2, \omega_3$)	CLIP ↑	DINO ↑	LPIPS ↓	Emo-A ↑	Emo-S ↑
[1.0, 0.0, 0.0]	0.7231	0.5422	0.6541	0.2850	0.2210
[1.0, 0.5, 0.0]	0.7384	0.5815	0.6312	<u>0.3120</u>	<u>0.2740</u>
[**1.0**, **0.5**, **0.2**] (Ours)	<u>0.7594</u>	**0.6194**	**0.6173**	**0.3473**	**0.3270**
[1.0, 0.5, 0.4]	**0.7610**	<u>0.5985</u>	<u>0.6215</u>	0.2912	0.2450

Appendix C

To eliminate the impact of structural randomness and ensure strict statistical rigor, all evaluations for our final LLaVA-Emo model are macro-averaged across three distinct random seeds (42, 100, and 262). As summarized in Table A3, our framework demonstrates exceptional stability across all metrics, with fluctuations strictly bounded within $\pm 0.5\%$. Specifically, the core affective alignment metrics yield a consistent output of 0.3475 ± 0.0036 for Emo-A and 0.3269 ± 0.0034 for Emo-S, firmly proving that the proposed emotion-aware steering and Direct Preference Optimization (DPO) mechanisms achieve deterministic and reproducible convergence rather than relying on fortuitous weight initializations.

Table A3. Quantitative results of LLaVA-Emo evaluated across multiple random seeds on the EmoEdit dataset. ↑ and ↓ indicate higher and lower values are preferred, respectively.

Seed Configuration	CLIP ↑	DINO ↑	LPIPS ↓	Emo-A ↑	Emo-S ↑
SEED = 42	0.7594	0.6194	0.6173	0.3473	0.3270
SEED = 100	0.7562	0.6145	0.6198	0.3441	0.3235
SEED = 262	0.7611	0.6213	0.6132	0.3512	0.3302
Mean ± SD	0.7589 ± 0.0025	0.6184 ± 0.0035	0.6168 ± 0.0033	0.3475 ± 0.0036	0.3269 ± 0.0034

Appendix D

To ensure data quality and verify the behavior of our automated filtering framework, we conduct a manual inspection on a targeted subset. This validation pool is constructed by sorting the initially retained samples in ascending order based on their cosine similarity scores and selecting the bottom 10%.

The manual inspection is performed by a single annotator who reviews each instance using only the artwork image and its paired human comment to make a binary decision. An instance is permanently discarded if the comment is: (i) extremely short or ambiguous; (ii) missing key visual elements required to support the described content; (iii) inconsistent with the target emotional intent; or (iv) clearly not supported by the visual evidence in the artwork. Otherwise, the sample is retained. Since this quality check is strictly limited to binary classification rather than additional data annotation, a single annotator is sufficient for the purpose of this workflow.

Table A4 provides a comprehensive qualitative exhibition of four retained and four discarded samples along with their titles, emotion labels, human comments, and the underlying manual inspection rationales.

Table A4. Exemplar case studies of retained and discarded samples during the manual inspection phase.

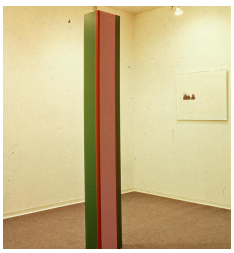
Artwork	Visual	Emotion	Human Comment	Manual Inspection Rationale
Group 1: Retained Samples (Salvaged due to Deep Affective/Implicit Alignment)				
Fishermen at Sea by William Turner		Fear	There is a spooky feel to this image, with the waters churning and the green sky overhead.	**RETAINED**: Reasonable comments, consistent with the elements of the image.
The Church in Nuenen in Winter, 1883, by Vincent van Gogh		Fear	The look of the jagged barren trees and the house in solitude makes it feel haunted and abandoned.	**RETAINED**: Reasonable comments, consistent with the elements of the image.
Group 2: Discarded Samples (Eliminated due to Semantic/Contextual Failure)				
Second Requiem, 1977, by Anne Truitt		Anger	Not many colors.	**DISCARDED**: The comment is too short and lacks visual support.

Table A4. *Cont.*

Artwork	Visual	Emotion	Human Comment	Manual Inspection Rationale
Untitled #235, 2007, by Claude Viallat		Excitement	Nice design.	**DISCARDED**: The meaning is ambiguous and lacks visually supporting elements.

Appendix E

To offer a more detailed view of emotion controllability and address potential evaluation bias, we provide a normalized confusion matrix on the EmoEdit test set using our evaluation classifier. As illustrated in Figure A1, each cell tracks the predicted emotions (γ_{pred}) against the target emotions (γ_{true}). Given the inherent difficulty of affective stylization, visually related emotion pairs can occasionally lead to off-diagonal predictions. However, the resulting patterns remain structured rather than scattered, indicating that the framework maintains meaningful, target-dependent control across different emotional categories.

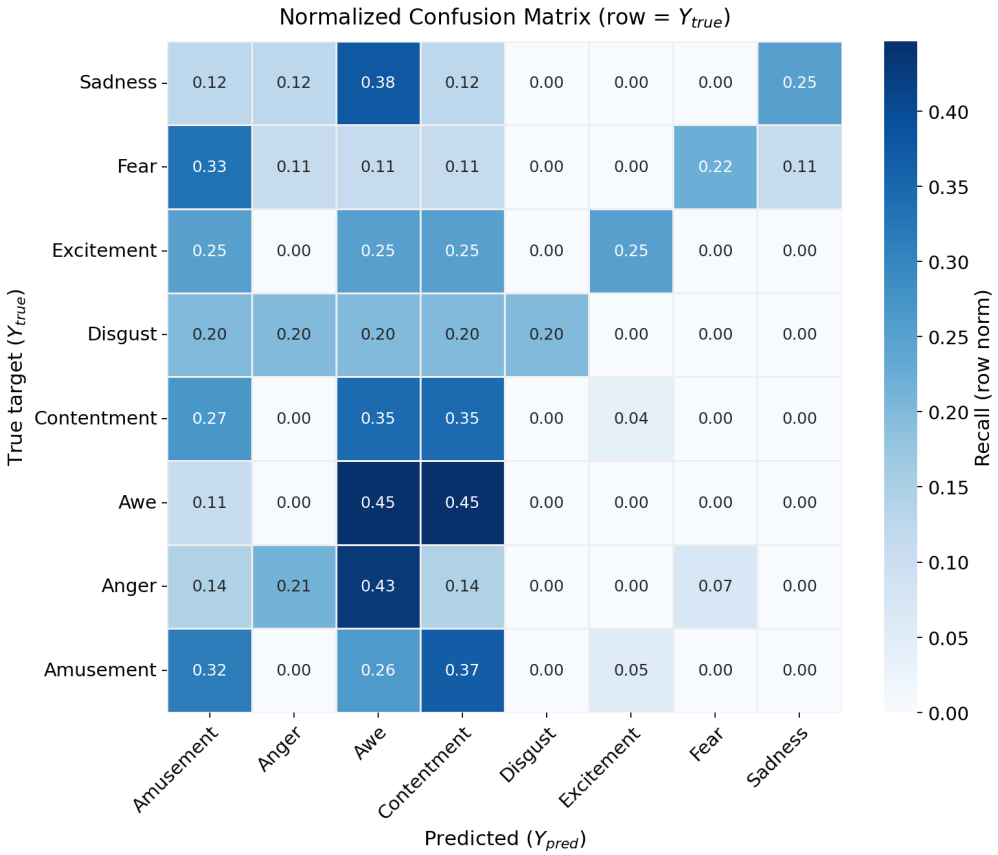


Figure A1. Normalized confusion matrix on the EmoEdit test.

References

1. Barwell, I. How Does Art Express Emotion? *J. Aesthet. Art Crit.* **1986**, *45*, 175–181. [CrossRef]
2. Joshi, D.; Datta, R.; Fedorovskaya, E.; Luong, Q.T.; Wang, J.Z.; Li, J.; Luo, J. Aesthetics and Emotions in Images. *IEEE Signal Process. Mag.* **2011**, *28*, 94–115. [CrossRef]

3. Xiao, Y.; Song, L.; Chen, Y.; Luo, Y.; Chen, Y.; Gan, Y.; Huang, W.; Li, X.; Qi, X.; Shan, Y. Mindomni: Unleashing reasoning generation in vision language models with rgpo. *Adv. Neural Inf. Process. Syst.* **2026**, *38*, 88786–88810.
4. Fang, R.; Duan, C.; Wang, K.; Huang, L.; Li, H.; Tian, H.; Yan, S.; Yu, W.; Zeng, X.; Dai, J.; et al. GoT: Unleashing Reasoning Capability of MLLM for Visual Generation and Editing. *Adv. Neural Inf. Process. Syst.* **2026**, *38*, 67680–67708.
5. Achlioptas, P.; Ovsjanikov, M.; Haydarov, K.; Elhoseiny, M.; Guibas, L. ArtEmis: Affective Language for Visual Art. In *Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021*; IEEE: New York, NY, USA, 2021; pp. 11564–11574. [CrossRef]
6. Mohamed, Y.; Khan, F.F.; Haydarov, K.; Elhoseiny, M. It is Okay to Not Be Okay: Overcoming Emotional Bias in Affective Image Captioning by Contrastive Data Collection. In *Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022*; IEEE: New York, NY, USA, 2022; pp. 21231–21240. [CrossRef]
7. Zhang, C.; Xie, H.; Wen, B.; Zuo, S.; Zhang, R.; Cheng, W.H. Emoart: A multidimensional dataset for emotion-aware artistic generation. In *Proceedings of the 33rd ACM International Conference on Multimedia, Dublin, Ireland, 27–31 October 2025*; ACM: New York, NY, USA, 2025; pp. 12644–12650.
8. Aslan, S.; Castellano, G.; Digeno, V.; Migailo, G.; Scaringi, R.; Vessio, G. Recognizing the Emotions Evoked by Artworks Through Visual Features and Knowledge Graph-Embeddings. In *Proceedings of the Image Analysis and Processing. ICIAP 2022 Workshops, Lecce, Italy, 23–27 May 2022*; Mazzeo, P.L., Frontoni, E., Sclaroff, S., Distante, C., Eds.; Springer International Publishing: Cham, Switzerland, 2022; pp. 129–140. [CrossRef]
9. Chen, T.; Garcia, N.; Li, L.; Nakashima, Y. Exploring Emotional Stimuli Detection in Artworks: A Benchmark Dataset and Baselines Evaluation. *J. Imaging* **2024**, *10*, 136. [CrossRef]
10. Ishikawa, S.; Sugiura, K. Affective Image Captioning for Visual Artworks Using Emotion-Based Cross-Attention Mechanisms. *IEEE Access* **2023**, *11*, 24527–24534. [CrossRef]
11. Fu, T.J.; Wang, X.E.; Wang, W.Y. Language-Driven Artistic Style Transfer. In *Proceedings of the Computer Vision–ECCV 2022, Tel Aviv, Israel, 23–27 October 2022*; Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T., Eds.; Springer Nature: Cham, Switzerland, 2022; pp. 717–734. [CrossRef]
12. Huang, N.; Zhang, Y.; Tang, F.; Ma, C.; Huang, H.; Dong, W.; Xu, C. DiffStyler: Controllable Dual Diffusion for Text-Driven Image Stylization. *IEEE Trans. Neural Netw. Learn. Syst.* **2025**, *36*, 3370–3383. [CrossRef] [PubMed]
13. Kwon, G.; Ye, J.C. In *Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022*; IEEE: New York, NY, USA, 2022; pp. 18041–18050. [CrossRef]
14. Zhang, Y.; Huang, N.; Tang, F.; Huang, H.; Ma, C.; Dong, W.; Xu, C. Inversion-based Style Transfer with Diffusion Models. In *Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023*; IEEE: New York, NY, USA, 2023; pp. 10146–10156. [CrossRef]
15. Gatys, L.A.; Ecker, A.S.; Bethge, M. Image Style Transfer Using Convolutional Neural Networks. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016*; IEEE: New York, NY, USA, 2016; pp. 2414–2423. [CrossRef]
16. Hertz, A.; Voynov, A.; Fruchter, S.; Cohen-Or, D. Style Aligned Image Generation via Shared Attention. In *Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 16–22 June 2024*; IEEE: New York, NY, USA, 2024; pp. 4775–4785. [CrossRef]
17. Xing, P.; Wang, H.; Sun, Y.; Wang, Q.; Bai, X.; Ai, H.; Huang, R.; Li, Z. CSGO: Content-Style Composition in Text-to-Image Generation. In Proceedings of the Thirty-Ninth Annual Conference on Neural Information Processing Systems, San Diego, CA, USA, 2–7 December 2025.
18. Picard, R.W. *Affective Computing*; The MIT Press: Cambridge, MA, USA, 1997. [CrossRef]
19. Liu, Z.S.; Wang, L.W.; Siu, W.C.; Kalogeiton, V. Name your style: Text-guided artistic style transfer. In *Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Vancouver, BC, Canada, 17–24 June 2023*; IEEE: New York, NY, USA, 2023; pp. 3530–3534. [CrossRef]
20. Liu, D.; Jiang, Y.; Pei, M.; Liu, S. Emotional image color transfer via deep learning. *Pattern Recognit. Lett.* **2018**, *110*, 16–22. [CrossRef]
21. Yang, C.K.; Peng, L.K. Automatic Mood-Transferring between Color Images. *IEEE Comput. Graph. Appl.* **2008**, *28*, 52–61. [CrossRef] [PubMed]
22. Lin, Q.; Zhang, J.; Ong, Y.S.; Zhang, M. Make Me Happier: Evoking Emotions Through Image Diffusion Models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Honolulu, HI, USA, 19–23 October 2025*; IEEE: New York, NY, USA, 2025; pp. 16367–16376.
23. Sun, S.; Jia, J.; Wu, H.; Ye, Z.; Xing, J. MSNet: A Deep Architecture Using Multi-Sentiment Semantics for Sentiment-Aware Image Style Transfer. In *Proceedings of the ICASSP 2023—2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 4–10 June 2023*; IEEE: New York, NY, USA, 2023; pp. 1–5. [CrossRef]

24. Yang, J.; Feng, J.; Huang, H. EmoGen: Emotional Image Content Generation with Text-to-Image Diffusion Models. In *Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 16–22 June 2024*; IEEE: New York, NY, USA, 2024; pp. 6358–6368. [CrossRef]
25. Yang, J.; Feng, J.; Luo, W.; Lischinski, D.; Cohen-Or, D.; Huang, H. EmoEdit: Evoking Emotions through Image Manipulation. In *Proceedings of the 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 11–15 June 2025*; IEEE: New York, NY, USA, 2025; pp. 24690–24699. [CrossRef]
26. Yang, J.; Bai, Z.; Huang, H. EmoStyle: Emotion-Driven Image Stylization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Denver, CO, USA, 3–7 June 2026*; IEEE: New York, NY, USA, 2026.
27. Li, J.; Li, D.; Savarese, S.; Hoi, S. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv* **2023**, arXiv:2301.12597.
28. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning Transferable Visual Models from Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning, PMLR, Virtual, 18–24 July 2021*; Meila, M., Zhang, T., Eds.; Proceedings of Machine Learning Research; International Conference on Machine Learning (ICML): San Diego, CA, USA, 2021; Volume 139, pp. 8748–8763.
29. Zhang, L.; Rao, A.; Agrawala, M. Adding Conditional Control to Text-to-Image Diffusion Models. In *Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 1–6 October 2023*; IEEE: New York, NY, USA, 2023; pp. 3813–3824. [CrossRef]
30. Mikels, J.A.; Fredrickson, B.L.; Larkin, G.R.; Lindberg, C.M.; Maglio, S.J.; Reuter-Lorenz, P.A. Emotional category data on images from the International Affective Picture System. *Behav. Res. Methods* **2005**, *37*, 626–630. [CrossRef]
31. Yang, J.; Huang, Q.; Ding, T.; Lischinski, D.; Cohen-Or, D.; Huang, H. EmoSet: A Large-scale Visual Emotion Dataset with Rich Attributes. In *Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 1–6 October 2023*; IEEE: New York, NY, USA, 2023; pp. 20326–20337. [CrossRef]
32. Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H.V.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; et al. DINOv2: Learning Robust Visual Features Without Supervision. *arXiv* **2023**, arXiv:2304.07193.
33. Zhang, R.; Isola, P.; Efros, A.A.; Shechtman, E.; Wang, O. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–26 June 2018*; IEEE: New York, NY, USA, 2018; pp. 586–595.
34. Zhang, Z.; Zhang, Q.; Lin, H.; Xing, W.; Mo, J.; Huang, S.; Xie, J.; Li, G.; Luan, J.; Zhao, L.; et al. Towards Highly Realistic Artistic Style Transfer via Stable Diffusion with Step-aware and Layer-aware Prompt. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24, Jeju, Republic of Korea, 3–9 August 2024*; Larson, K., Ed.; International Joint Conferences on Artificial Intelligence Organization: Darmstadt, Germany, 2024; Volume 8, pp. 7814–7822. [CrossRef]
35. Brooks, T.; Holynski, A.; Efros, A.A. InstructPix2Pix: Learning to Follow Image Editing Instructions. In *Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023*; IEEE: New York, NY, USA, 2023; pp. 18392–18402.
36. Brack, M.; Friedrich, F.; Kornmeier, K.; Tsaban, L.; Schramowski, P.; Kersting, K.; Passos, A. LEDITS++: Limitless Image Editing Using Text-to-Image Models. In *Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 16–22 June 2024*; IEEE: New York, NY, USA, 2024; pp. 8861–8870. [CrossRef]
37. Weng, S.; Zhang, P.; Chang, Z.; Wang, X.; Li, S.; Shi, B. Affective Image Filter: Reflecting Emotions from Text to Images. In *Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 1–6 October 2023*; IEEE: New York, NY, USA, 2023; pp. 10776–10785. [CrossRef]
38. Zhang, S.; Zhang, X.; Shen, L.; Wan, S.; Ren, W. Wavelet-based physically guided normalization network for real-time traffic dehazing. *Pattern Recognit.* **2025**, *172*, 112451. [CrossRef]
39. Liu, Y.; Li, T.; Tan, C.; Ren, W.; Ancuti, C.; Lin, W. IHDCP: Single Image Dehazing Using Inverted Haze Density Correction Prior. *IEEE Trans. Image Process.* **2026**, *35*, 1448–1461. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.