Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Cross-Modal Alignment for Emotion-Consistent Text-to-Image Generation (ChinaMM 2026)

Qinglan Wei
Communication University of China
Ke Wang
Communication University of China
Jinglei Tan
Communication University of China
Yuan Zhang
Communication University of China
Long Ye
yelong@cuc.edu.cn

Communication University of China

Research Article

Keywords: Cross-Modal Alignment, Emotional Consistency, Text-to-Image Generation

Posted Date: September 29th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10467906/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Title Page

Article Title:

Cross-Modal Alignment for Emotion-Consistent Text-to-Image Generation (ChinaMM 2026)

Authors and Affiliations:

Qinglan Wei[1], Ke Wang[1], Jinglei Tan[3], Yuan Zhang[2], Long Ye[1,]*

[1]School of Data Science and Intelligent Media, Communication University of China, No.1 Dingfuzhuang East Street, Chaoyang District, Beijing 100024, China; qlwei@cuc.edu.cn (Q.W.); mercyken4536251@mails.cuc.edu.cn (K.W.); yelong@cuc.edu.cn (L.Y.)

[2]School of Information and Communication Engineering, Communication University of China, No.1 Dingfuzhuang East Street, Chaoyang District, Beijing 100024, China; yzhang@cuc.edu.cn (Y.Z.)

[3]Hainan International College, Communication University of China, Binhai Campus, Lingshui Li’an International Education Innovation Pilot Zone, Hainan, 572400, China ; tanj112@uni.coventry.ac.uk (J.T.)

[*]Correspondence: yelong@cuc.edu.cn (L.Y.)

Corresponding Author:

Long Ye

Email: yelong@cuc.edu.cn

Current/Permanent Address:

All authors’ current affiliations and addresses are as listed above. No author has changed affiliation since the completion of this work.

Acknowledgments:

This work was supported in part by the Hainan Provincial Joint Project of Li’an International Education Innovation Pilot Zone, in part by the National Natural Science Foundation of China under Grant 62301510, in part by the Fundamental Research Funds for the Central Universities, and in part by the Public Computing Cloud at Communication University of China.

Author Contributions:

Conceptualization, Qinglan Wei, Ke Wang and Long Ye; Methodology, Qinglan Wei, Ke Wang; Formal analysis and investigation, Ke Wang and Qinglan Wei; Writing-original draft preparation,

Qinglan Wei and Ke Wang; Writing-review and editing, Jinglei Tan, Long Ye and Yuan Zhang; Funding acquisition, Qinglan Wei; Supervision, Yuan Zhang and Long Ye.

Cross-Modal Alignment for Emotion-Consistent Text-to-Image Generation (ChinaMM 2026)

Abstract

Purpose: Cross-modal emotional alignment remains a major bottleneck in text-to-image generation, as semantic shifts between textual prompts and generated visuals often undermine output credibility. Existing methods typically rely on external emotion labels and lack a unified framework that jointly addresses emotional understanding and expression. To bridge this gap, we propose the Text-to-Image Emotion Consistent Generation (T2I-ECG) framework. Methods: The framework employs a two-stage strategy. In the first stage, it extracts emotional intent from natural-language texts and manifests it into emotion-enriched visual prompts, which guide the generation of preliminarily aligned images. In the second stage, a cross-modal enhancement network refines the images by injecting emotional semantics into the visual feature space, thereby improving emotional fidelity. Results: Extensive experiments on the EmoVisText6-3000 dataset demonstrate the effectiveness of our method, achieving notable improvements over recent approaches in related domains, with gains of 3.43% in binary emotion accuracy and 4.40% in six-class emotion accuracy. Conclusion: The proposed T2I-ECG framework effectively bridges the cross-modal emotional alignment gap, providing a unified solution that jointly handles emotional understanding and expression, and produces more credible emotion-consistent image generations.

Keywords: Cross-Modal Alignment; Emotional Consistency; Text-to-Image Generation

1. Introduction

Emotion, as a core dimension of human communication and cognitive experience, plays a key role in enhancing the efficiency of information transmission and the depth of emotional resonance [1, 2]. In the era of digital content, images serve as a powerful medium for conveying emotional and semantic information, capable of instantly evoking emotional responses from viewers [3, 4]. For example, scenes like “a birthday party” are typically associated with joyful visual elements, whereas “desolate ruins” tend to elicit feelings of sadness. With the increasing adoption of text-to-image (T2I) technologies, users are not only demanding semantic accuracy in generated content but also seeking emotional congruence.

Therefore, achieving precise cross-modal emotional alignment between textual intent and generated imagery remains a key challenge for improving the practicality, reliability, and robustness of T2I models. It is worth noting that the alignment pursued in this work refers to the transfer and synchronization of emotional semantics from text to visual emotion, rather than the construction of a shared multimodal representation space.

In recent years, diffusion-based models have driven significant progress in T2I generation, demonstrating strong capabilities in transforming textual descriptions into high-quality, semantically faithful images [5-7]. However, current models often lack sufficient sensitivity to emotional cues, frequently resulting in emotional semantic shifts in generated outputs. To address this, several methods have attempted preliminary solutions. For instance, methods like EmoGen [8] leverage the mapping between emotion and semantic spaces to guide the generation process toward predefined emotional

categories. Others like EmoEdit [9] apply emotion editing operations to existing images based on external emotional labels. Nonetheless, these approaches fail to fundamentally resolve the issue of emotional consistency. At their core, existing works do not establish a unified framework that integrates both emotional understanding and emotional expression.

To bridge this gap, we propose the Text-to-Image Emotion Consistent Generation (T2I-ECG) framework, which aims to align the emotional semantics of the input text with the visual affect of the generated image. Our framework adopts a two-stage optimization strategy. First, it parses emotional signals from the textual input and enhances the prompt through an emotion-mapping scheme. Then, an emotion-driven cross-modal optimization network refines emotional consistency within the visual feature space, further improving the emotional alignment of the output image. Extensive evaluations demonstrate that T2I-ECG achieves superior performance compared to recent methods in both binary emotion accuracy and six-class emotion accuracy, showing significant potential for emotional content creation.

Our contributions can be summarized as follows:

- We propose the T2I-ECG framework that pioneers an intent-driven paradigm for emotion-consistent generation, effectively shifting from traditional external label dependency to autonomous textual emotion parsing and visual alignment.
- We introduce a psychologically grounded emotion-mapping scheme that transforms abstract emotional states into concrete multi-dimensional visual cues, establishing an interpretable bridge that ensures semantic-affective synergy throughout the synthesis process.
- We contribute a dedicated dataset and evaluation benchmark for cross-modal emotional alignment in text-to-image generation, facilitating future research in this emerging direction.
- We demonstrate the flexibility of the framework by applying it to multiple T2I models, verifying its adaptability to different generation backbones.

2. Related Work

Recent years have witnessed remarkable progress in image generation and editing. In particular, diffusion-based generative models have demonstrated superior performance in generating high-quality images [5]. Building on this foundation, researchers have increasingly investigated fine-grained control over content, structure, and semantics in generated images, advancing the field of conditional image generation [10]. As an abstract and subjective condition, emotion has also become a growing area of interest in this field [4]. Accordingly, this section reviews two main threads of related work: general conditional image generation and emotion-driven image generation.

2.1. Conditional Image Generation

Conditional image generation aims to synthesize image content that aligns with specific input conditions. Early studies primarily focused on using single conditions, gradually evolving toward more flexible multimodal control mechanisms. Initial approaches based on Conditional Variational Autoencoders (CVAEs) [11] and Generative Adversarial Networks (GANs) [12] improved generation consistency by introducing class labels or semantic information. With the rapid development of diffusion models, the precision and fidelity of conditional generation have made a qualitative leap. GLIDE [13] first applied diffusion mechanisms to the text-to-image task, while models such as Stable Diffusion [7] and DALL • E [14] further expanded the capacity for large-scale semantic modeling.

To enhance the personalization of generated content, more advanced condition-guidance techniques have been proposed. For instance, Textual Inversion [15] and DreamBooth [16] enable concept-level image generation via embedding optimization and model fine-tuning, respectively. DreamArtist [17] and Custom Diffusion [18] achieve lightweight fine-tuning by modifying only the attention modules. ControlNet [19] and T2I-Adapter [20] introduce structural conditions such as edge maps to improve the model's structural sensitivity. IP-Adapter [21] incorporates reference image features via cross-attention, ensuring better attribute and style consistency in generated results.

While these methods have shown strong performance in semantically explicit generation tasks, they remain limited when handling more abstract conditions like emotion. On one hand, most methods lack explicit modeling mechanisms for emotional semantics; on the other hand, emotion, as a subjective and multidimensional construct, typically requires multi-level semantic coordination, placing higher demands on T2I systems. To address this, we propose a new task, Cross-Modal Alignment for Emotion-Consistent Text-to-Image Generation, aiming to help models better interpret emotional cues from text and generate emotionally aligned imagery with semantic clarity.

2.2. Emotion-Driven Image Generation

Emotion-driven image generation focuses on producing visual content that can evoke specific emotional responses. Earlier efforts primarily concentrated on Image Emotion Transfer (IET), which alters the emotional tone of an image while preserving its structural content. This was achieved by modifying either low-level visual features such as color [22-27] or higher-level stylistic elements [28, 29].

Yang et al. [22] were among the first to propose emotion editing by adjusting color and texture distributions. Wang et al. [23] and He et al. [24] established mappings between colors and emotions to perform emotion-specific image editing. Peng et al. [25] and Ali et al. [26] further introduced reference-based transformations, guiding color and texture adjustments with external examples. More recently, Liu et al. [27] improved earlier approaches by integrating deep learning techniques to enable more precise emotion control. Zhu et al. [28] employed GANs to disentangle different visual features and achieve emotion transformation while preserving structure. Weng et al. [29] adopted a Transformer-based architecture [30] to merge emotional and visual features, enabling hierarchical style-level emotion editing.

Recent works have further advanced emotion-driven generation with more structured and flexible modeling. EmoEdit [9] constructs an emotion factor tree to map abstract emotions to concrete visual elements for controllable editing. EmoGen [8] extends this direction to generative models by bridging emotion and semantic spaces for direct image synthesis. EmoKGEdit [31] enhances controllability via multimodal knowledge graphs and cue-driven reasoning. EmoSpace [32] introduces prototype-based representations to capture fine-grained emotional variations during generation.

Despite these advances, existing methods often rely on predefined emotion annotations or focus primarily on emotion modulation, lacking the ability to directly interpret emotional cues embedded in user prompts. To address this limitation, this work integrates emotional understanding and visual generation into a unified framework.

3. Method

To address the limitations of existing methods that rely heavily on emotion labels and lack the capability to directly interpret the emotional intent embedded in textual prompts, we propose Text-to-

Image Emotion Consistent Generation (T2I-ECG)—a unified framework that jointly models emotion understanding and emotion expression. T2I-ECG adopts a two-stage optimization strategy, namely Emotion-Intelligence Resonance Generation and Emotion-Guided Cross-Modal Enhancement, with the goal of aligning the emotional semantics of the generated image with those conveyed in the input text. The overall architecture is illustrated in Fig. 1.

3.1. Framework Overview

3.1.1. Emotion-Intelligence Resonance Generation

In this stage, the raw textual prompt is converted into an emotion-enriched visual description through two collaborating modules. The Emo-Detection Module first extracts a high-confidence emotion label, providing a core constraint for generation. The Emo-Manifestation Module then maps the detected emotion to concrete visual cues. Under the guidance of a mapping scheme, the Large Language Model (LLM) incorporates these cues into the original prompt to produce an Emo-Visual Description. This enhanced description enables the T2I model to generate an image with preliminary emotional alignment. For instance, as shown in Fig. 1, given the prompt "In the sunlight, a happy dog stands on the steps," the Emo-Detection Module identifies the emotion joy. Subsequently, the Emo-Manifestation Module enriches the prompt with joy-related cues such as "warm sunshine" and "colorful balloons".

3.1.2. Emotion-Guided Cross-Modal Enhancement

Building on the intermediate image generated in Stage 1, the Emo-Enhancement Network further refines emotional consistency at the visual feature level. Taking both the intermediate image and its associated emotion label as inputs, the network projects abstract emotional semantics into deeper visual representations, thereby enhancing emotional consistency while preserving the semantic structure of the original image. For example, when the target emotion is joy, the model reinforces the emotional tone by adding elements such as "smiling faces" or "colorful candy", without compromising the underlying visual content.

3.2. Emo-Detection Module

Accurately identifying emotional intent from a textual prompt is a crucial first step toward emotion-consistent image generation. In our framework, the Emo-Detection Module adopts RoBERTa [33] as the backbone due to its strong contextual understanding and robust transferability across downstream tasks. However, conventional RoBERTa-based emotion classifiers are predominantly trained on social media corpora and thus generalize poorly to visual description texts. Therefore, we train a new emotion detection module tailored specifically for text-to-image generation scenarios.

To align the training corpus with natural user prompts, we utilize ArtEmis [34], ArtEmis v2 [35], and Affection [36] datasets, which include human-written descriptions of emotional images annotated with eight emotion categories, mapping image semantics to human emotion. These sources cover diverse visual domains including artworks, daily scenes, human activities, and real-world objects. We systematically clean these datasets, retaining six emotion categories: amusement, awe, anger, disgust, fear, and sadness. Following Ekman's six basic emotions theory [37], we map the original dataset labels from eight emotions to six categories to better align with natural emotional expressions in language. Specifically, "amusement" is re-labeled as "joy" to reflect broader positive emotions, and "awe" is redefined as "surprise" to better capture reactions of shock and amazement.

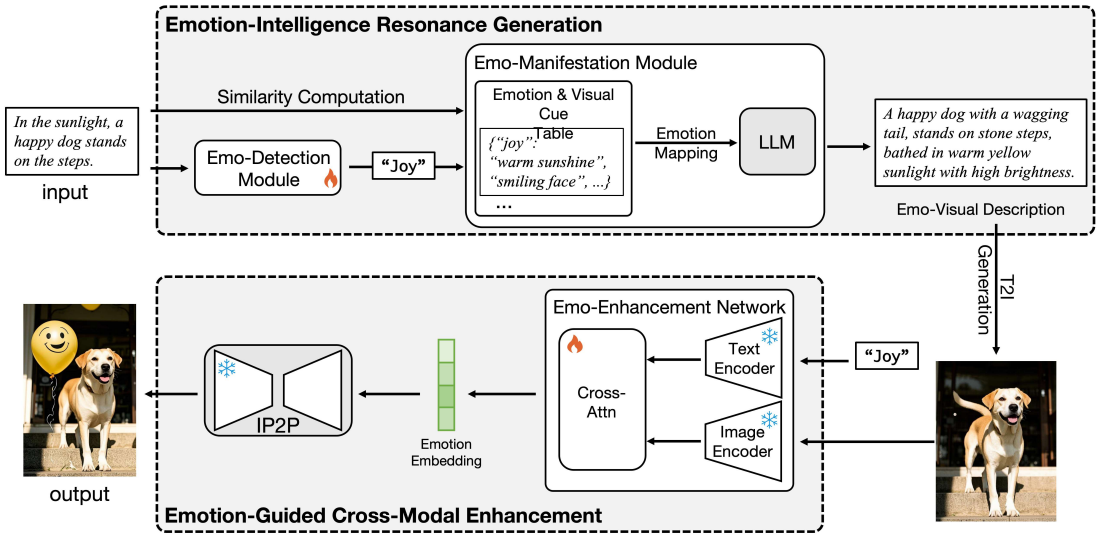


Fig. 1 The framework of the proposed T2I-ECG. It comprises two sequential stages: "Emotion-Intelligence Resonance Generation" and "Emotion-Guided Cross-Modal Enhancement

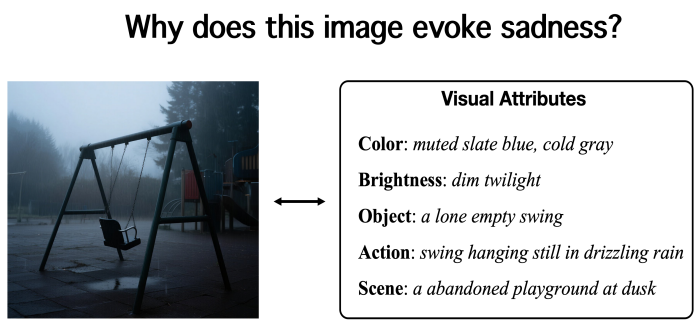


Fig. 2 An example of image-based emotion analysis utilizing the proposed emotion-mapping scheme

Ultimately, we construct a six-class emotion dataset for image generation prompts named EmoVisPrompt6, which contains 1,001,623 emotion-rich visually described texts. The six-class emotion label set can be denoted as: $E_6 = \{joy, surprise, anger, disgust, fear, sadness\}$. Given an input prompt t, the emotion detection module outputs the most probable emotion label $e \in E_6$. The prediction is given in (1):

$$e = \arg\max_{e_i \in E_6} f_{RoBERTa}(t; \theta_{FT}) \tag{1}$$

where $f_{RoBERTa}$ denotes the prediction function of RoBERTa and θ_{FT} represents the fine-tuned parameters.

```
Role: You are an expert visual prompts designer for text-to-image generation.

Task: You need to transform the {{input}} into a visual prompt.

Tip: The generated prompts should manifest {{emotion}} through visual elements.

Ref: Reference visual clues: {{visual_clues}}.

Output:
{
  "color": "...",
  "brightness": "...",
  "objects": [...],
  "actions": [...],
  "scene": "...",
  "image_prompt": "..."
}
```

Fig. 3 A sample prompt template for formalizing the emotion-mapping scheme

3.3. Emo-Manifestation Module

While the emotion label derived from Stage 1 captures the dominant affect conveyed by the input text, it does not specify how such affective states can be visually instantiated in generated imagery. To bridge the cross-modal semantic gap from textual emotion categories to explicit and tangible visual representations, we introduce the Emo-Manifestation Module. This module comprises two core components: an Emotion & Visual Cue Table and a structured emotion-mapping scheme, which collectively guide the LLM in generating emotion-enriched visual descriptions.

Inspired by the psychological work of Zhao et al. [38], which argues that emotions in visual content are conveyed through distinct attributes (such as color, brightness, object, action, and scene), we design a structured mapping scheme to formalize this insight. As illustrated in Fig. 2, combinations of cues such as “muted slate blue” and “a lone empty swing” collectively evoke sadness, exemplifying how specific visual elements shape affective perception. This principle underpins our mapping strategy, which enables the LLM to translate discrete emotion labels into concrete visual representations. As illustrated in Fig. 3, the mapping process is implemented through a structured prompt template.

However, LLM hallucinations may compromise the quality of generated prompts. The prevailing solution is Retrieval-Augmented Generation (RAG), which retrieves relevant knowledge from an external corpus to ground LLM outputs. Following this paradigm, we construct an Emotion & Visual Cue Table as an external knowledge base and design a retrieval-injection mechanism that feeds semantically aligned cues into the prompt template, thereby constraining the LLM's generation space.

The Emotion & Visual Cue Table is constructed via a data-driven extraction process based on EmoVisPrompt6. We first perform tokenization on all 1,001,623 prompts and extract adjective-noun pairs as candidate visual cues. These candidates are grouped by their corresponding emotion labels, and the top 500 most frequent entries per category are retained as preliminary candidates. We then manually curate this subset to preserve only cues with clear visual semantics and strong emotion specificity, yielding a final set of approximately 300 cues per emotion. This compact design prioritizes diversity and

interpretability over exhaustive coverage, forming a non-redundant, high-quality cue repository for subsequent generation.

To exemplify the table, we list some entries for each emotion category. For joy: “warm sunlight,” “smiling face,” “colorful balloons”; for surprise: “wide-open eyes,” “gasping mouth”; for anger: “red fire,” “clenched fist”; for disgust: “rotten food,” “messy space”; for fear: “dark shadow,” “creepy skull”; for sadness: “crying face,” “rainy sky”. These samples demonstrate clear visual semantics and distinct emotional boundaries across categories, laying a reliable foundation for the subsequent cue selection and prompt enrichment process.

During inference, only cues that are semantically aligned with the original prompt are selected, and the LLM (Qwen3-8B with default parameters) integrates them into the input prompt to produce an Emo-Visual Description. By leveraging retrieved visual cues as emotional grounding signals, the generated description provides the T2I model with richer and more reliable emotional context for the subsequent image-generation process.

Specifically, we compute the cosine similarity between the input vector v_t and candidate visual cue vectors v_c to measure their relevance. This similarity measure encourages the selection of visual cues that are both semantically relevant and emotionally consistent with the input prompt. The cosine similarity is defined as:

$$\mathrm{Sim} = \frac{v_t \cdot v_c}{\| v_t \| \| v_c \|} \tag{2}$$

Based on the computed similarities, the top k ($k = 5$) most relevant cues $\mathcal{R}$ from the candidate set V are selected as:

$$\mathcal{R} = \underset{v_c \in V}{\mathrm{Top}-k}\ \mathrm{Sim}(v_t, v_c) \tag{3}$$

3.4. Emo-Enhancement Network

Although the first stage introduces emotion-relevant cues into the text, emotional details may diminish during image synthesis. To address this issue, the Emo-Enhancement Network is introduced in Stage 2 to reinforce emotional fidelity directly in the visual feature space. This network takes both the intermediate image and its associated emotion label as inputs and injects emotional semantics into the deep visual representation while preserving the structural integrity of the original image content.

We build the network upon InstructPix2Pix (IP2P) [39], which is known for its high image fidelity and effective instruction following but lacks explicit modeling of emotional semantics. Inspired by recent work that introduces a lightweight Emotion Adapter [9] to improve IP2P’ s affective capacity, we design and train the Emo-Enhancement Network, which includes three key components: an Emotion Encoder, an Image Encoder, and a Cross-Attention Fusion Module. By jointly encoding the target emotion and image content, the network generates an emotion embedding that serves as a fine-grained affective constraint, enabling downstream editing to reinforce emotional cues without altering the semantic layout of the image.

During training, we use the EmoSet dataset [40], which contains triplets of (image, text annotation, emotion label), providing rich semantic supervision. Following standard diffusion training procedures [41], we freeze the original IP2P parameters and optimize only the Emo-Enhancement Network. This targeted integration allows the final output to exhibit clearer, more consistent emotional characteristics aligned with the intended affect. The loss function is shown in (4):

$$\mathcal{L}_{\mathrm{diff}} = \mathbb{E}_{x, e_t, \epsilon, t}[\| \epsilon_\theta(x_t, t, e_t) - \epsilon \|_2^2] \tag{4}$$

where x is the original image sample, x_t is the noised version at timestep t, e_t is the embedding of the target emotion, $\in \sim \mathcal{N}(0,1)$ is Gaussian noise, and $\epsilon_\theta(x_t, t, e_t)$ denotes the predicted noise. To prevent semantic collapse during emotion injection, the cross-attention mechanism enables spatially adaptive conditioning of emotional embeddings on image features, ensuring that emotional modulation does not override the underlying semantic structure.

4. Experiments and Results

4.1. Experimental Setup

Cross-modal Emotion Alignment (CEA) is a novel task that aims to synchronize the emotional intent embedded in natural language prompts with the affective atmosphere of the generated images, differing from prior tasks such as Visual Emotion Analysis (VEA), Image Emotion Transfer (IET), and Emotion Image Generation (EIG). Since this field remains in early exploration, unified evaluation benchmarks are currently lacking. To enable rigorous assessment, we construct a dedicated dataset and design task-specific metrics tailored to emotional consistency evaluation.

4.1.1. Dataset

We introduce EmoVisText6-3000, comprising 3,000 high-quality prompts reflecting diverse emotional tendencies. These are LLM-generated or web-sourced, with manual annotation and filtering to ensure balanced coverage of six basic emotions (joy, surprise, fear, disgust, sadness, anger) at 500 prompts per category.

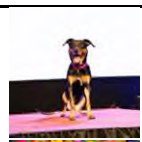
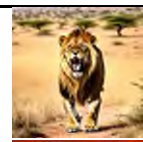
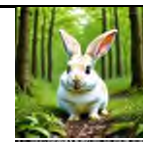
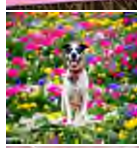
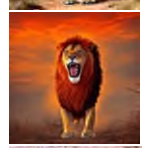
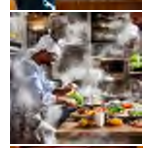
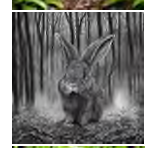
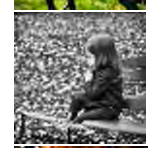
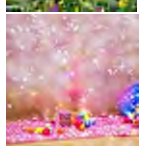
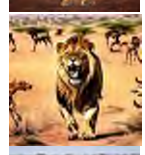
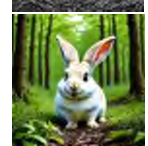
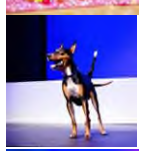
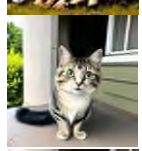
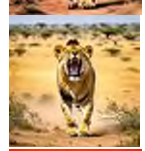
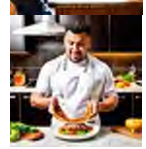
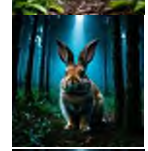
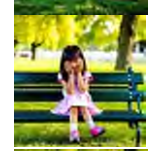
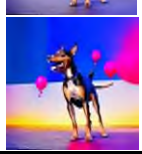
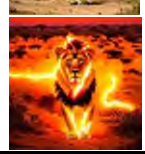
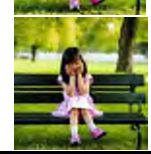
4.1.2. Metric

Drawing inspiration from evaluation protocols in VEA [42-45], IET [28, 29], and EIG [46], we formulate a task-specific version of the Emotion Accuracy (EmoAcc) metric. To guarantee fairness in the subsequent experiments, we use the image emotion classifier employed in EmoGen [8] and EmoEdit [9] for our evaluation. This classifier predicts the emotion category of each generated image, which is then compared against the ground-truth emotion of the input prompt. We report the performance using three levels of metrics. Specifically, Emo2 measures the binary emotion accuracy, grouping joy and surprise as positive, while categorizing anger, disgust, fear, and sadness as negative. Emo6 represents the six-class emotion accuracy, indicating the overall accuracy across all six categories. Furthermore, we report the per-class accuracy to provide a detailed evaluation for each individual emotion.

4.1.3. Baselines

We select representative approaches from related domains for subsequent experiments, including foundational text-to-image baselines Stable Diffusion 1.5 (SD1.5) [7] and Stable Diffusion 3 (SD3) [47], which represent the state-of-the-art in open-source image generation. We also include T2I-Adapter (T2I-A) [20] and IP-Adapter (IP-A) [21], representative methods for conditional image generation and fine-tuning. Furthermore, specialized emotional image generation frameworks are included for comparison, specifically EmoGen [8] for Emotion Image Generation (EIG) and EmoEdit [9] for Affective Image Manipulation (AIM).

Table 1 Qualitative Experiment

Method \ Prompt	*There is a happy dog on the stage*	*A shocked cat sat on the porch*	*A furious lion sat on the savanna*	*A man stood in the kitchen, feeling repulsed*	*A terrified rabbit in a forest*	*There is a little girl crying on a park bench*
SD3 [47]						
T2I-A [21]						
EmoEdit [9]						
+Stage 1						
+Stage 1 +Stage 2						

4.1.4. Implementation Details

All experiments are conducted on a single NVIDIA L40 GPU with 48GB memory. Standard training or fine-tuning procedures are applied across all modules. To mitigate GPU memory limitations, the majority of backbone parameters are frozen, restricting optimization to lightweight, task-specific layers. In the Emo-Manifestation Module, semantic relevance between the input and candidate cues is computed using the text embedding layer of the fine-tuned RoBERTa encoder.

4.2. Qualitative Comparison

We conduct qualitative comparisons with SD3, T2I-A, EmoEdit, and two variants of our framework (Stage 1 only and the full version) across six emotion scenarios, with prompts and generation results shown in Table 1. The results indicate that while general models like SD3 capture semantic entities, they exhibit limited capacity in expressing specific emotional intent. EmoEdit can accurately convey emotions in some scenarios, yet typically at the cost of compromising semantic fidelity to the original image. Overall, our method produces visually expressive results that effectively balance emotional intensity with semantic requirements across diverse scenarios.

These qualitative results highlight that the two-stage design of T2I-ECG, which combines interpretable emotion-to-visual mapping in Stage 1 with deep cross-modal feature enhancement in Stage

Table 2 Quantitative Experiment

Method	Emo Acc/%							
	Emo2↑	Emo6↑	Anger↑	Joy↑	Sadness↑	Fear↑	Disgust↑	Surprise↑
SD3	78.87	47.23	45.80	77.20	60.00	47.40	30.40	22.60
IP-A	81.27	50.37	50.60	82.80	69.00	49.20	31.40	19.20
T2I-A	82.63	52.10	48.00	86.00	67.80	53.00	38.20	19.60
EmoGen	85.37	61.47	72.00	90.60	73.60	60.40	58.00	14.20
EmoEdit	85.00	64.83	78.20	91.20	74.40	69.80	57.60	17.80
Ours	88.43	69.23	85.40	92.60	76.00	74.20	65.00	22.20

Table 3 Ablation Study

Method	Emo Acc/%		Runtime ↓
	Emo2 ↑	Emo6 ↑	
SD1.5 [7]	74.57	45.13	28.63s
+Stage 1	79.80	55.30	30.46s
+Stage 1 +Stage 2	83.63	63.07	33.15s

2, enables precise control over both emotional intensity and semantic fidelity. By jointly modeling emotion understanding and expression without relying on external labels, our method achieves superior emotional resonance while preserving the original prompt semantics, demonstrating clear advantages over existing label-driven or editing-based approaches.

4.3. Quantitative Comparison

To objectively evaluate emotional consistency, we conduct quantitative experiments on the EmoVisText6-3000 dataset. We compare our full two-stage framework against SD3, IP-A, T2I-A, EmoGen, and EmoEdit. We measure performance using EmoAcc at three levels: binary accuracy (Emo2), six-class accuracy (Emo6), and per-class accuracy.

As shown in Table 2, our method achieves superior performance on EmoAcc metrics. Specifically, Emo2 hits 88.43%, surpassing the strongest baseline, EmoEdit, by 3.43%. This demonstrates our framework's enhanced capability in capturing and expressing overall emotional polarity (positive versus negative). For Emo6, the accuracy reaches 69.23%, a 4.40% improvement over EmoEdit, further validating its effectiveness in fine-grained six-class emotion alignment.

Table 4. User Study

Method	Emotion Consistency ↑	Semantic Consistency ↑
SD1.5 [7]	21.29%	26.71%
T2I-A [20]	25.20%	25.09%
EmoEdit [9]	26.47%	22.24%
Ours	27.04%	25.96%

While our method achieves superior overall performance, a thorough per-class analysis reveals a consistent bottleneck across all evaluated approaches: the emotion category of surprise yields notably lower accuracy compared to the other five categories. We attribute this phenomenon to two primary factors. First, surprise is inherently complex and visually ambiguous, frequently sharing overlapping visual cues with fear and joy, which renders accurate expression and discrimination challenging. Second, the image emotion classifier, adopted from representative works in the field of emotion recognition, may exhibit limited discriminative power for surprise, potentially introducing evaluation bias. This limitation highlights the demand for developing more robust and standardized evaluation benchmarks, which we regard as a key direction for future work.

Importantly, to ensure experimental fairness, all comparisons in this work are conducted under settings that favor the baseline methods. Specifically, EmoGen and EmoEdit are provided with ground-truth emotion labels, while IP-A and T2I-A are equipped with reference images generated by the base model. In contrast, our framework operates solely on raw textual prompts, without any supplementary annotations or visual guidance. Despite these favorable settings for competing methods, our approach still achieves consistently superior performance, demonstrating the efficacy of jointly modeling emotion understanding and expression independent of external annotations.

This necessity stems from the fact that CEA is a novel task proposed in this work, for which no existing approach is natively designed. We thus adopt representative methods from adjacent domains, which inherently cannot directly infer emotional intent from raw text. Instead, they either depend on predefined emotion labels or require extra reference images for manipulation, obliging us to supply the necessary auxiliary inputs to ensure their proper functioning. Furthermore, since these baselines are primarily fine-tuned on Stable Diffusion backbones, all experiments are performed on various variants of this architecture to ensure a fair comparison. Preliminary small-scale tests also indicate that more advanced generative backbones have the potential to further enhance image quality, which points to a promising direction for future work.

4.4. Ablation Study

To verify the flexibility of T2I-ECG and the necessity of its components, we conduct ablation experiments with SD1.5 as the image generation backbone. Compared to SD3 employed in Table II, SD1.5 represents an earlier open-source standard with noticeable gaps in both generation quality and inference efficiency. Our intentional selection of this lower-capacity baseline serves to examine T2I-ECG's effectiveness under constrained conditions. This design also isolates efficiency variables

associated with backbone evolution, enabling accurate assessment of the framework's own computational overhead.

As shown in Table 3, Stage 1 alone already brings notable improvements, boosting Emo2 and Emo6 accuracy by 5.23% and 10.17%, respectively. This demonstrates that emotion-aware prompt enrichment is effective in capturing and injecting emotional intent into the generation process. Building upon this, the integration of Stage 2 further enhances performance, with Emo2 and Emo6 accuracy reaching 83.63% and 63.07%, respectively. This indicates that while Stage 1 provides a strong foundation for emotional alignment at the semantic level, Stage 2 plays a crucial role in refining emotional consistency within the visual feature space. The combined two-stage framework therefore achieves more robust and fine-grained emotional alignment than either stage alone.

Regarding inference efficiency, the full framework increases average runtime from 28.63s to 33.15s, incurring merely 4.52s additional overhead with a relative increase of approximately 15.8%. This cost primarily stems from emotional parsing and prompt enrichment in Stage 1, together with cross-modal feature editing in Stage 2. Notably, these components can directly benefit from optimizations in lightweight and advanced models, thereby demonstrating the flexibility of our framework. In preliminary tests, adopting a lightweight variant (Z-Image-Turbo [48]) as the backbone model compresses total generation time to under 10 seconds while maintaining high generation quality, showcasing the practical potential of T2I-ECG. Further systematic evaluation across diverse lightweight architectures remains for future work.

4.5. User Study

To validate human preference, we randomly sampled 300 images from the full set of 3,000 generated outputs for human evaluation, ensuring balanced coverage across all six emotion categories (joy, surprise, anger, disgust, fear, sadness) with 50 samples per class. We invited 15 participants to evaluate these images. In each session, participants were presented with the original prompt, the target emotion, and four anonymized outputs (SD1.5, T2I-A, EmoEdit, Ours), and asked: (1) Which image best evokes the targeted emotion? (2) Which image best aligns with the prompt’s semantics?

As shown in Table 4, T2I-ECG achieves a competitive preference rate of 27.04% for Emotion Consistency, showing a lead over specialized affective models like EmoEdit. This can be attributed to the nature of single-choice evaluation, where comparative methods may employ more aggressive emotional rendering strategies that appeal to voters on specific samples. Notably, T2I-ECG also maintains competitive performance in Semantic Consistency (25.96%), ranking second only to the base generative model SD1.5 (26.71%) while significantly outperforming T2I-A (25.09%) and EmoEdit (22.24%).

These results reflect an inherent tension between emotional intensity and semantic fidelity. Aggressive emotional rendering tends to distort the original scene composition, whereas strict semantic preservation often weakens affective expression. Our two-stage framework explicitly decouples these competing objectives. Stage 1 establishes the emotional direction via prompt-level enrichment, and Stage 2 refines affective details within the visual feature space while preserving the underlying semantic layout. This progressive design allows T2I-ECG to better balance the trade-off compared with single-stage editing or generation paradigms, yielding stronger emotional resonance with little compromise to semantic integrity.

5. Conclusion

5.1. Conclusion

In this paper, we introduce a new task, Cross-Modal Alignment for Emotion-Consistent Text-to-Image Generation, which emphasizes emotional consistency between textual prompts and generated images. We propose a novel framework that explicitly models and transfers emotional signals from text to image generation. Experimental results demonstrate that our method achieves superior emotional consistency compared to existing text-to-image diffusion models. We believe this work provides a meaningful step toward more affectively intelligent generative systems and opens new directions for research at the intersection of vision, language, and affective computing.

5.2. Limitations

While our method shows advantages in enhancing emotional consistency, several limitations remain. To ensure model stability, we adopt well-established architectures such as RoBERTa and IP2P as our backbone. However, more advanced multimodal models [46, 49-51] have recently emerged, which may further improve performance in emotion alignment. Additionally, the four-alternative forced-choice paradigm in our human preference study reflects only the most emotionally salient option, while unselected methods may equally possess valid emotional expressiveness. This constitutes a limitation in our experimental design, calling for future work to explore and construct more comprehensive, diverse, and systematic benchmarks.

Statements and Declarations

Funding:

This work was supported by the Hainan Provincial Joint Project of Li'an International Education Innovation pilot Zone, the National Natural Science Foundation of China (Grant No. 62301510), the Fundamental Research Funds for the Central Universities, and the Public Computing Cloud, CUC.

Competing Interests:

The authors declare that they are not members of the editorial board or advisory board of this journal, and have no competing interests to declare that are relevant to the content of this article.

Data Availability:

The EmoVisText6-3000 dataset constructed in this study is available from the corresponding author upon reasonable request. Publicly available datasets such as IEMOCAP, ArtEmis, ArtEmis v2, Affection, and EmoSet were used in this study and can be accessed through their respective official repositories.

Author Contributions:

Conceptualization, Qinglan Wei, Ke Wang and Long Ye; Methodology, Qinglan Wei, Ke Wang; Formal analysis and investigation, Ke Wang and Qinglan Wei; Writing-original draft preparation, Qinglan Wei and Ke Wang; Writing-review and editing, Jinglei Tan, Long Ye and Yuan Zhang; Funding acquisition, Qinglan Wei; Supervision, Yuan Zhang and Long Ye.

References

[1] M. de Vega, I. León, and J. M. Díaz, "The representation of changing emotions in reading comprehension," *Cognition and Emotion*, vol. 10, no. 3, pp. 303–321, 1996, doi: 10.1080/026999396380268.

[2] T. Brosch, G. Pourtois, and D. Sander, "The perception and categorisation of emotional stimuli: A review," in *Cognition and Emotion: Reviews of Current Research and Theories*, J. De Houwer and D. Hermans, Eds. Psychology Press, 2010, pp. 66–98.

[3] J. Li, C. Luo, Q. Zhang, et al., "Can emotional design really evoke emotion in multimedia learning?," *Int. J. Educ. Technol. High. Educ*., vol. 17, no. 24, 2020, doi: 10.1186/s41239-020-00198-y.

[4] S. Afzal, H. A. Khan, M. J. Piran, and J. W. Lee, "A comprehensive survey on affective computing: Challenges, trends, applications, and future directions," *IEEE Access*, vol. 12, pp. 96150–96168, 2024.

[5] C. Zhang, C. Zhang, M. Zhang, I. S. Kweon, and J. Kim, "Text-to-image diffusion models in generative AI: A survey," arXiv:2303.07909, 2024. [Online]. Available: https://arxiv.org/abs/2303.07909

[6] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Advances in Neural Information Processing Systems*, vol. 33, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, Eds. Curran Associates, Inc., 2020, pp. 6840–6851.

[7] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 10674–10685.

[8] J. Yang, J. Feng, and H. Huang, "EmoGen: Emotional image content generation with text-to-image diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2024, pp. 6358–6368, doi: 10.1109/CVPR52733.2024.00608.

[9] J. Yang, J. Feng, W. Luo, D. Lischinski, D. Cohen-Or, and H. Huang, "EmoEdit: Evoking emotions through image manipulation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2025, pp. 24690–24699.

[10] P. Cao, F. Zhou, Q. Song, and L. Yang, "Controllable generation with text-to-image diffusion models: A survey," arXiv:2403.04279, 2024. [Online]. Available: https://arxiv.org/abs/2403.04279

[11] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," arXiv:1312.6114, 2013. [Online]. Available: https://arxiv.org/abs/1312.6114.

[12] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 27, 2014, pp. 2672–2680.

[13] A. Q. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. McGrew, I. Sutskever, and M. Chen, "GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models," in *Proc. 39th Int. Conf. Mach. Learn. (ICML)*, vol. 162, PMLR, 2022, pp. 16784–16804.

[14] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with CLIP latents," arXiv:2204.06125, 2022. [Online]. Available: https://arxiv.org/abs/2204.06125

[15] R. Gal, Y. Alaluf, Y. Atzmon, O. Patashnik, A. H. Bermano, G. Chechik, and D. Cohen-Or, "An image is worth one word: Personalizing text-to-image generation using textual inversion," arXiv:2208.01618, 2022. [Online]. Available: https://arxiv.org/abs/2208.01618

[16] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman, "DreamBooth: Fine tuning text-to-image diffusion models for subject-driven generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023, pp. 22500–22510, doi: 10.1109/CVPR52729.2023.02155.

[17] Z. Dong, P. Wei, and L. Lin, "DreamArtist: Controllable one-shot text-to-image generation via positive-negative adapter," *Int. J. Comput. Vis.*, vol. 133, pp. 7037–7053, 2022.

[18] N. Kumari, B. Zhang, R. Zhang, E. Shechtman, and J.-Y. Zhu, "Multi-concept customization of text-to-image diffusion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023, pp. 1931–1941.

[19] L. Zhang, A. Rao, and M. Agrawala, “Adding conditional control to text-to-image diffusion models,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2023, pp. 3813 – 3824.
[20] C. Mou, X. Wang, L. Xie, J. Zhang, Z. Qi, Y. Shan, and X. Qie, “T2I-Adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models,” in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2024.
[21] H. Ye, J. Zhang, S. Liu, X. Han, and W. Yang, “IP-Adapter: Text compatible image prompt adapter for text-to-image diffusion models,” arXiv:2308.06721, 2023. [Online]. Available: https://arxiv.org/abs/2308.06721
[22] C.-K. Yang and L.-K. Peng, “Automatic mood transferring between color images,” *IEEE Comput. Graph. Appl.*, vol. 28, no. 2, pp. 52 – 61, Mar./Apr. 2008.
[23] X. Wang, J. Jia, and L. Cai, “Affective image adjustment with a single word,” *Vis. Comput.*, vol. 29, no. 11, pp. 1121 – 1133, Nov. 2013, doi: 10.1007/s00371-012-0755-3.
[24] L. He, H. Qi, and R. L. Zaretzki, “Image color transfer to evoke different emotions based on color combinations,” *Signal Image Video Process.*, vol. 9, pp. 1965 – 1973, 2015.
[25] K.-C. Peng, T. Chen, A. Sadovnik, and A. Gallagher, “A mixed bag of emotions: Model, predict, and transfer emotion distributions,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2015, pp. 860 – 868, doi: 10.1109/CVPR.2015.7298687.
[26] M. Ali and A. Rafaqat, “Automatic image transformation for inducing affect,” in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, 2017, doi: 10.5244/C.31.171.
[27] D. Liu, Y. Jiang, M. Pei, and S. Liu, “Emotional image color transfer via deep learning,” *Pattern Recognit.* Lett., vol. 110, pp. 16 – 22, 2018, doi: 10.1016/j.patrec.2018.03.015.
[28] S. Zhu, C. Qing, C. Chen, and X. Xu, “Emotional generative adversarial network for image emotion transfer,” *Expert Syst. Appl.*, vol. 216, Art. no. 119485, 2023.
[29] S. Weng, P. Zhang, Z. Chang, X. Wang, S. Li, and B. Shi, “Affective image filter: Reflecting emotions from text to images,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2023, pp. 10776 – 10785, doi: 10.1109/ICCV51070.2023.00992.
[30] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2017.
[31] J. Zhang and B. Fan, “EmoKGEdit: Training-free Affective Injection via Visual Cue Transformation,” arXiv:2601.12326, 2026. [Online]. Available: https://arxiv.org/abs/2601.12326
[32] B. Wang, X. Chen, Z. Qiu, L. Yuan, and Z. Wang, “EmoSpace: Fine-Grained Emotion Prototype Learning for Immersive Affective Content Generation,” arXiv:2602.11658, 2026. [Online]. Available: https://arxiv.org/abs/2602.11658
[33] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, “RoBERTa: A robustly optimized BERT pretraining approach,” arXiv:1907.11692, 2019. [Online]. Available: https://arxiv.org/abs/1907.11692
[34] P. Achlioptas, M. Ovsjanikov, K. Haydarov, M. Elhoseiny, and L. Guibas, “ArtEmis: Affective language for visual art,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021, pp. 11564 – 11574, doi: 10.1109/CVPR46437.2021.01140.
[35] Y. Mohamed, F. F. Khan, K. Haydarov, and M. Elhoseiny, “It is okay to not be okay: Overcoming emotional bias in affective image captioning by contrastive data collection,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 21231 – 21240, doi: 10.1109/CVPR52688.2022.02058.
[36] P. Achlioptas, M. Ovsjanikov, L. Guibas, and S. Tulyakov, “Affection: Learning affective explanations for real-world visual data,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023, pp. 6641 – 6651, doi: 10.1109/CVPR52729.2023.00642.
[37] P. Ekman, “Cross-cultural studies of facial expression,” in *Darwin and Facial Expression: A Century of Research in Review*, P. Ekman, Ed. Academic Press, 1973, pp. 169 – 222.
[38] S. Zhao et al., “Affective image content analysis: Two decades review and new perspectives,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 10, pp. 6729 – 6751, Oct. 2022, doi: 10.1109/TPAMI.2021.3094362.
[39] T. Brooks, A. Holynski, and A. A. Efros, “InstructPix2Pix: Learning to follow image editing instructions,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023, pp. 18392 – 18402, doi: 10.1109/CVPR52729.2023.01764.
[40] J. Yang, Q. Huang, T. Ding, D. Lischinski, D. Cohen-Or, and H. Huang, “EmoSet: A large-scale visual emotion dataset with rich attributes,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2023, pp. 20326 – 20337, doi: 10.1109/ICCV51070.2023.01864.
[41] D. Kumar, P. Väth, and M. Gregorová, “Loss functions in diffusion models: A comparative study,” in *Machine Learning and Knowledge Discovery in Databases. Research Track: Proc. Eur. Conf. Mach. Learn. Princ. Pract. Knowl. Discov. Databases (ECML PKDD)*, Part II, Springer, 2025, pp. 190 – 205, doi: 10.1007/978-3-032-05981-9_12.
[42] T. Rao, X. Li, and M. Xu, “Learning multi-level deep representations for image emotion classification,” *Neural Process. Lett.*, vol. 51, pp. 2043 – 2061, 2020.

[43] J. Yang, X. Gao, L. Li, X. Wang, and J. Ding, “SOLVER: Scene-object interrelated visual emotion reasoning network,” *IEEE Trans. Image Process.*, vol. 30, pp. 8686–8701, 2021, doi: 10.1109/TIP.2021.3118983.
[44] J. Yang, J. Li, X. Wang, Y. Ding, and X. Gao, “Stimuli-aware visual emotion analysis,” *IEEE Trans. Image Process.*, vol. 30, pp. 7432–7445, 2021, doi: 10.1109/TIP.2021.3106813.
[45] W. Zhang, X. He, and W. Lu, “Exploring discriminative representations for image emotion recognition with CNNs,” *IEEE Trans. Multimedia,* vol. 22, no. 2, pp. 515–523, 2020, doi: 10.1109/TMM.2019.2928998.
[46] S. Dang, Y. He, L. Ling, Z. Qian, N. Zhao, and N. Cao, “EmotiCrafter: Text-to-emotional-image generation based on valence-arousal model,” arXiv:2501.05710, 2025. [Online]. Available: https://arxiv.org/abs/2501.05710
[47] P. Esser et al., “Scaling rectified flow transformers for high-resolution image synthesis,” in *Proc. 41st Int. Conf. Mach. Learn. (ICML)*, PMLR, 2024, pp. 12606–12633.
[48] Z-Image Team, et al., “Z-Image: An efficient image generation foundation model with single-stream diffusion transformer,” arXiv:2511.22699, 2025. [Online]. Available: https://arxiv.org/abs/2511.22699
[49] Z. Yan et al., “EEdit: Rethinking the spatial and temporal redundancy for efficient image editing,” arXiv:2503.10270, 2025. [Online]. Available: https://arxiv.org/abs/2503.10270
[50] Y. Hu et al., “DCEdit: Dual-level controlled image editing via precisely localized semantics,” arXiv:2503.16795, 2025. [Online]. Available: https://arxiv.org/abs/2503.16795
[51] S. Liu et al., “Step1X-Edit: A practical framework for general image editing,” arXiv:2504.17761, 2025. [Online]. Available: https://arxiv.org/abs/2504.17761