EmotiCrafter: Text-to-Emotional-Image Generation based on Valence-Arousal Model

Shengqi Dang [1,2†], Yi He [1†], Long Ling [1], Ziqing Qian [1], Nanxuan Zhao [3], Nan Cao [1,2*]

[1]Tongji University [2]Shanghai Innovation Institute [3]Adobe Research

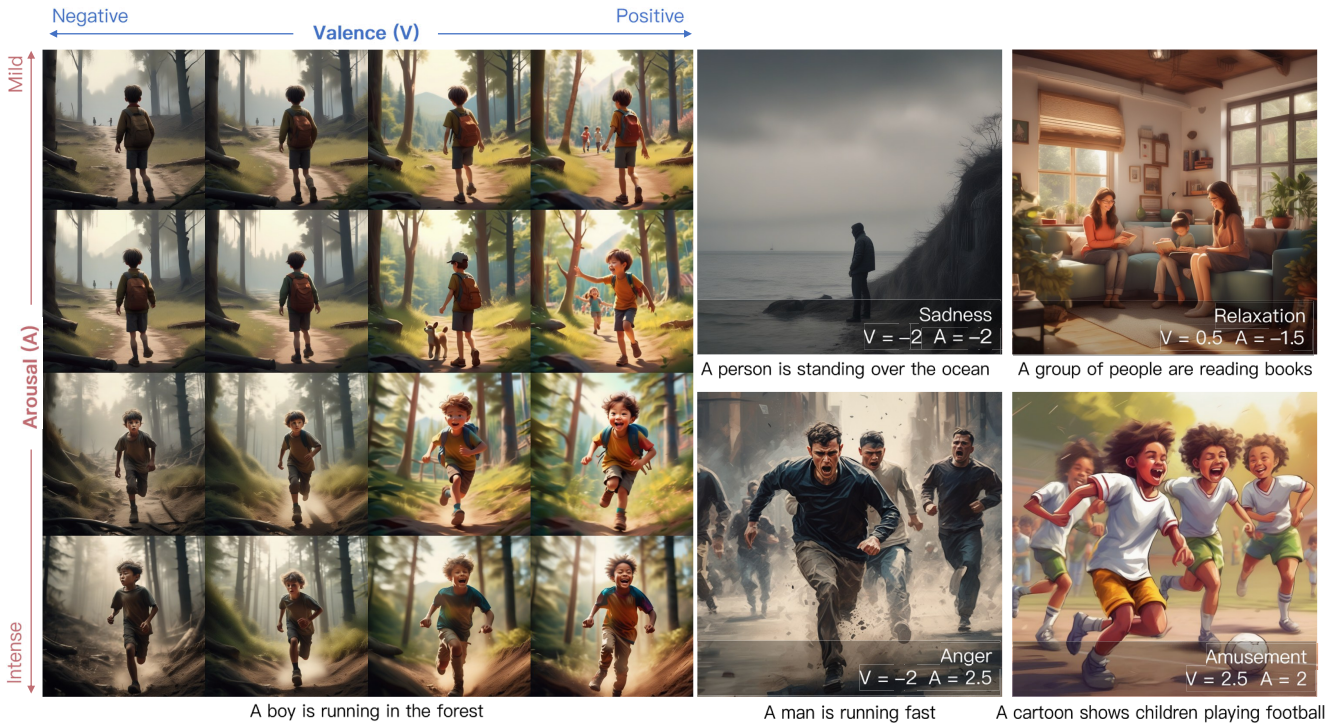


Figure 1. Generated images of the *EmotiCrafter*. Given a prompt and Valence-Arousal (V-A) values, our method can generate an emotional image that reflects the input prompt and aligns with the specified emotion values. By adjusting the V-A values, our method can also generate images that evoke discrete emotions, such as sadness, relaxation, anger, and amusement.

Abstract

Recent research shows that emotions can enhance users' cognition and influence information communication. While research on visual emotion analysis is extensive, limited work has been done on helping users generate emotionally rich image content. Existing work on emotional image generation relies on discrete emotion categories, making it challenging to capture complex and subtle emotional nuances accurately. Additionally, these methods struggle to control the specific content of generated images based on text prompts. In this paper, we introduce the task of continuous emotional image content generation (C-EICG) and present EmotiCrafter, a general emotional image generation model that generates images based on free text prompts and Valence-Arousal (V-A) values. It leverages a novel emotion-embedding mapping network to fuse V-A values

†Shengqi Dang and Yi He contributed equally to this work.

*Nan Cao is the corresponding author.

*Shengqi Dang, Yi He, Long Ling, Ziqing Qian, and Nan Cao are with the Intelligent Big Data Visualization Lab, Tongji University. Shengqi Dang and Nan Cao are also with the Shanghai Innovation Institute. Email: {dangsq123, heyi_11}@tongji.edu.cn, lucyling0224@gmail.com, 2411920@tongji.edu.cn, nan.cao@gmail.com.

*Nanxuan Zhao is with Adobe Research. Email: nanxuanzhao@gmail.com.

*Code: https://github.com/idvxlab/EmotiCrafter

into textual features, enabling the capture of emotions in alignment with intended input prompts. A novel loss function is also proposed to enhance emotion expression. The experimental results show that our method effectively generates images representing specific emotions with the desired content and outperforms existing techniques.

1. Introduction

Emotions are fundamental to human experiences and play a critical role in shaping how people perceive and interact with the world. Research has shown that emotions affect memory [15, 19, 25, 45] and comprehension [16, 38, 40], which are crucial for effective communication. As a result, content creators increasingly recognize the importance of incorporating emotions to enhance audience engagement.

While research on visual emotion analysis is extensive [13, 49, 58], there is limited work on generating emotionally rich image content. Some early studies explored emotional content generation techniques within specific domains such as facial expressions [2, 46] or landscapes [24]. EmoGen [50] generates images based on a given emotion tag (e.g., happy or sad), which pioneered the domain-free emotional image content generation (EICG) task. However, it has two critical limitations: (1) the image is generated from an emotion tag instead of a text prompt, making the generated content difficult to control; (2) while the discrete emotion tags used in EmoGen are easy to understand, psychologists have not achieved a consensus on the emotion categories [60]. The limited scope of discrete emotion tags falls short of capturing nuanced emotions.

To address the above issues, we propose the continuous emotional image content generation (C-EICG) task and present *EmotiCrafter*, the first C-EIGC model that generates emotional images using free-text prompts and continuous Valence-Arousal (V-A) values defined in the V-A model (a well-known psychological continuous emotion model) [33]. V-A model represents emotions in a two-dimensional Cartesian space (Figure 2), where Valence quantifies pleasantness (negative to positive) and Arousal measures intensity (calm to excited). We utilize a [-3, 3] range for Valence and Arousal [10, 21]. The continuous V-A space enables smooth transitions and nuanced emotional shifts beyond the capability of discrete labels (e.g., shifts from "bored" to "tired" of a character's state in a video). Specific emotional values in this space have been investigated in prior work [34], and such fine-grained modeling is particularly beneficial for human–computer interaction [2]. Leveraging this model, *EmotiCrafter* captures subtle affective variations via precise (V, A) positioning. The contributions of this paper are as follows:

- We propose a novel task, continuous emotional image content generation (C-EICG), and develop the first dedicated model for this purpose. Our model introduces an emotion-embedding network that integrates continuous Valence-Arousal (V-A) values with text prompts. These fused features are then injected into Stable Diffusion XL [28] using cross-attention mechanisms, enabling precise control over both content and emotional expressions in the generated images.
- We propose a novel loss function that enhances the emotional resonance of generated images. By amplifying the difference between neutral and emotional text features, our approach enables the model to capture more distinct emotional variations. Additionally, the loss function incorporates the V-A distribution to address data imbalance, further refining the model's ability to generate images with rich, accurate emotional expressions.
- We constructed an emotional prompts dataset to train the emotion-embedding network, where each sample consists of a neutral prompt and an emotional prompt that share the same core meaning but express a specific emotion corresponding to a given pair of V-A values.

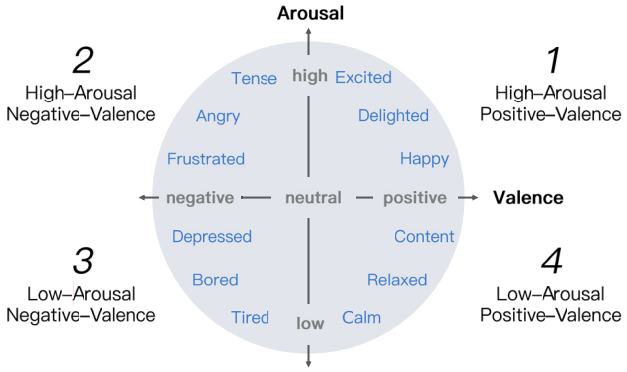


Figure 2. Valence-Arousal model.

2. Related work

In this section, we present a review of the related work, specifically focusing on visual emotion analysis, image emotion transfer, and conditional image generation.

2.1. Visual Emotion Analysis

Visual emotion analysis refers to the computational recognition and interpretation of human emotion [57] in visual media, such as images or videos. It has been a prominent research area, with most efforts focusing on the classification of discrete emotions [4, 42, 47, 48, 52]. However, discrete emotion categories limit the ability to capture the nuanced emotions [7], leading to an increased focus on continuous emotion analysis in images [59].

Much of the continuous emotion analysis remains centered on facial expression analysis [9, 11, 22, 37, 54]. While effective, facial analysis can overlook crucial con-

textual information that influences emotional interpretation. This drives studies to focus on emotions in objects or individuals within their environment, rather than just faces[12, 13, 20, 21]. For example, Kosti *et al.* [12] combined person-specific characteristics with the context of the scene to predict continuous emotional dimensions such as arousal, valence, and dominance. Kragel *et al.* [13] proposed EmoNet to extract visual features such as facial expressions, body posture, and scene elements from images to predict valence and arousal values. Recently, Mertens *et al.*[21] tested multiple backbones (such as ResNet, CLIP, DINO, etc.) for prediction of valence and arousal, demonstrating strong performance across various architectures.

Previous studies have achieved high accuracy in predicting continuous emotions in images, demonstrating a correlation between continuous emotions and visual elements. Building on this insight, our work moves beyond prediction to explore how continuous emotions can be actively embedded within generated content.

2.2. Image Emotion Transfer

Image Emotion Transfer (IET) focuses on editing the content of the images to evoke different emotions [17, 27, 43, 61]. For instance, Peng *et al.* [27] achieved emotion transfer by adjusting color tones and texture-related features. Zhu *et al.* [61] introduced a method that separates high-level emotion-relevant features (e.g., object shapes and scene layout) from low-level emotion-relevant features (e.g., brightness). By applying GANs, they transferred emotions between images while preserving their original structure. Building on these methods, Weng *et al.* [43] proposed the Affective Image Filter, which uses a multi-modal transformer to process both text and image inputs. With the emergence of text-to-image models, IET has expanded into new applications based on instructive commands. For example, EmoEdit [51] used GPT-4V to build emotion factor trees that map abstract emotions to specific visual elements and employed the InstructPix2Pix model to apply emotion-driven content and color adjustments to images.

Current methods primarily extract specific features, focusing on certain visual elements or emotional cues, which can limit the depth of emotional expression. In contrast, our method broadly learns a variety of emotion-influencing features, and it could accept natural language prompts as input.

2.3. Conditional Image Generation

Conditional image generation aims to create images that align with specific input conditions, such as text [3, 28, 30, 31], reference images [53], subjects [6, 32], and depth maps [23, 55]. To enhance quality, researchers have developed specialized approaches, such as Diffusion Transformers (DiT)[26], which use transformer-based diffusion for denoising, and Visual Autoregressive Modeling (VAR)[36], which encodes images into discrete tokens for autoregressive prediction across scales. However, these state-of-the-art methods rely on discrete labels, limiting flexibility and control. In contrast, our approach offers greater generalization by leveraging Stable Diffusion XL [28] to generate emotional images from free-text prompts.

Despite these progresses, incorporating emotion as a condition for image generation remains underexplored. EmoGen [50] pioneered the generation of emotional image content (EIGC) by mapping emotional features to semantic features to generate emotional images. However, EmoGen struggles to understand natural language, limiting its capacity to effectively control specific content. Its reliance on discrete emotions also limits its practical applicability.

Our method bridges this gap by embedding continuous emotion into textual features, enabling image generation models to use continuous emotion for emotion control. Unlike label-based methods, our method supports free-text prompts for flexible content control.

3. Method

In this section, we introduce the technique details of the proposed *EmotiCrafter*.

3.1. Overview

Our method generates emotional images I_{emo} from two inputs (Figure 3): a free-text prompt describing the desired content, and a pair of V-A values (v, a) specifying the emotion. First the prompt encoder $\mathcal{E}$ converts the text prompt into feature f_n, which is then processed by an emotion-embedding network $\mathcal{M}$ to produce emotional prompt feature $\hat{f}_e$ that integrate the V-A values:

$$\hat{f}_e = \mathcal{M}(f_n|(v, a)) \tag{1}$$

Next, this feature is injected into Stable Diffusion XL $\mathcal{G}$ via its cross-attention mechanism to generate the emotional images: $I_{emo} = \mathcal{G}(\hat{f}_e)$. To enhance emotional expressiveness, we introduce a loss function (Fig.3(b.2)) that leverages the V-A distribution and emphasizes the differences between neutral and emotional prompt features. This ensures that the generated images accurately convey both the intended emotions and content. Additionally, we construct a dataset (Fig.3(a)) that pairs neutral and emotional prompts with the corresponding V-A values.

3.2. Emotion-Embedding Network

The emotion-embedding network $\mathcal{M}$ generates the emotional prompt feature by integrating a pair of V-A values with a neutral prompt feature (Fig.3(b.1)). First, a V-A encoder converts the V-A values into feature vectors. Then, an emotion injection transformer—modified from GPT-2[29]—fuses these vectors with the neutral prompt

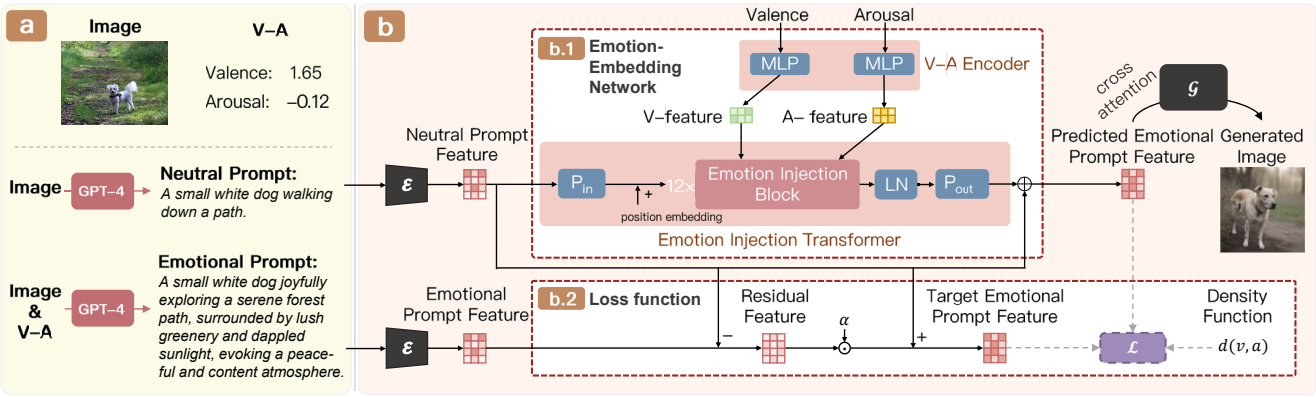


Figure 3. Overview of our method. Specifically, we take the following steps. (a) We collect an image dataset annotated with V-A values, neutral prompts, and emotional prompts. These prompts are then encoded into features by prompt encoder $\mathcal{E}$. (b) Next, we design (b.1) an emotion-embedding network $\mathcal{M}$ to embed V/A values into textual features based on the transformer architecture, and (b.2) a specialized loss function to enhance the emotional resonance of generated images. The output of the mapping network serves as the condition for the image generation model $\mathcal{G}$ to generate emotional images.

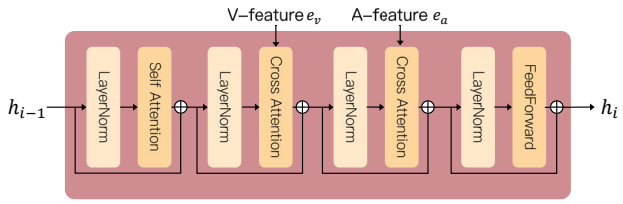


Figure 4. Structure of Emotion Injection Block. It accepts hidden state h_{i-1} as input and produces h_i as output. The V-feature e_v and A-feature e_a represent the emotion features, which are injected through the cross-attention module.

feature, preserving the original textual context while infusing emotional content.

V-A Encoder. The V-A Encoder converts a pair of V-A values into feature vectors using two separate multilayer perceptrons (MLPs). One MLP processes the Valence value to produce V-feature e_v, and the other processes the Arousal value to generate A-feature e_a. These features are then fed into the emotion injection transformer network for further emotion infusion.

Emotion Injection Transformer (EIT). The Emotion Injection Transformer (EIT) leverages a modified GPT-2 architecture to seamlessly integrate V/A-features into textual features. Its process consists of three stages: input projection, emotion injection, and output projection.

First, we project the input neutral prompt feature f_n into the transformer's feature space:

$$h_0 = \mathrm{P_{in}}(f_n) + \mathrm{PE}, \tag{2}$$

where h_0 represents the initial hidden state; $\mathrm{P_{in}}(\cdot)$ is a linear projection layer, and PE denotes positional embedding [39].

In the next, we inject emotion into f_n via 12 sequential **Emotion Injection Block**s (**EIB**s) corresponding to 12 transformer blocks. Each block outputs a hidden state:

$$h_i = \mathrm{EIB}(h_{i-1}, e_v, e_a), \quad i \in \{1, \ldots, 12\} \tag{3}$$

where h_i is the output of the i-th $\mathrm{EIB}(\cdot)$. As shown in Figure 4, each EIB enhances the transformer block through a cross-attention mechanism:

$$h'_i = \text{self-attn}(\mathrm{LN}(h_{i-1})) + h_{i-1} \tag{4}$$

$$h_i^{(v)} = \text{cross-attn}(\mathrm{LN}(h'_i), e_v) + h'_i \tag{5}$$

$$h_i^{(v,a)} = \text{cross-attn}(\mathrm{LN}(h_i^{(v)}), e_a) + h_i^{(v)} \tag{6}$$

$$h_i = \mathrm{fnn}(\mathrm{LN}(h_i^{(v,a)})) + h_i^{(v,a)} \tag{7}$$

where $h', h^{(v)}, h^{(v,a)}$ are the intermediate hidden variables; $\mathrm{LN}(\cdot)$ is the LayerNorm; self-attn$(\cdot)$ denotes self-attention, employed to capture context dependencies; cross-attn$(\cdot)$ is cross-attention for injecting e_v and e_a; fnn$(\cdot)$ is a feed-forward network that adapts the complexity of the emotional embedding process. We also remove the causal mask typically used for next-token prediction from the original transformer model to fit our task.

Finally, the output of the last (i.e., the 12th) injection block h_{12} is projected back to SDXL's prompt feature space via $\mathrm{P_{out}}$ (a linear layer and a LayerNorm layer) to predict the residual between emotional and neutral prompt features, which represents a semantic shift between emotional and neutral prompts:

$$\hat{f}_r = \mathrm{P_{out}}(\mathrm{LN}(h_{12})) \tag{8}$$

The final emotional prompt feature is obtained by adding this residual to the original neutral prompt feature:

$$\hat{f}_e = \hat{f}_r + f_n \tag{9}$$

The above emotion embedding network is trained by minimizing the averaged expectation of the difference between the predicted emotional prompt feature $\hat{f}_e = \mathcal{M}(f_n|(v,a))$ and the scaled target emotional prompt feature f_e^t, using the loss function described in Equation 10.

$$\mathcal{L} = \frac{1}{n}\mathbb{E}\left(\frac{1}{d(v,a)}\|\hat{f}_e - f_e^t\|^2\right) \tag{10}$$

where n is the number of feature elements; $\mathbb{E}(\cdot)$ denotes the expectation; $d(v,a)$ is a density function that describes the distribution of V-A values in the training sample.

To effectively embed emotions and address the challenges posed by the uneven distribution of V-A values in the dataset, this loss function incorporates two key strategies to improve the model's performance:

Scaled Residual Learning. To better capture pronounced emotional changes in generated images, we enlarge the target residuals:

$$f_e^t = f_n + \alpha \underbrace{(f_e - f_n)}_{\text{residual feature}}, \tag{11}$$

where f_e is the emotional prompt feature, f_n is the neutral prompt feature, and α is a scale factor, we set its value to 1.5 based on the ablation study.

V-A Density Weighting. To mitigate the effects of the imbalanced distribution of the training samples, we weigh the loss inversely proportional to the density of training samples in the V-A space. The density is estimated using Kernel Density Estimation (KDE) [5] with a Gaussian kernel, denoted as $d(v,a)$:

$$d(v,a) = \frac{1}{n}\sum_{i=1}^{n} K_H((v,a) - (v_i,a_i)), \tag{12}$$

where K_H is a 2D Gaussian kernel with bandwidth H; n is the number of training samples; (v_i,a_i) are the V-A values of the i-th training sample. The bandwidth H is selected using Silverman's rule of thumb to provide optimal smoothing of the density estimation.

3.3. Dataset and Training

To train the emotion-embedding network, we constructed a dataset of paired neutral and emotional prompts with corresponding Valence-Arousal (V-A) values (Figure 3(a)). These pairs were automatically generated using GPT-4 based on 39,843 images with human-annotated V-A values from publicly available datasets, including OASIS [14], EMOTIC [12], and FindingEmo [21]. Specifically, GPT-4 was used to generate neutral prompts with objective image descriptions and emotional prompts that emphasize affective attributes such as color, lighting, and texture, which influence emotional perception. To ensure data reliability, all LLM-generated prompts underwent crowd-sourced verification, with disagreements resolved through a voting mechanism among annotators.

The proposed emotion embedding network is trained on two NVIDIA A800 GPUs using the aforementioned dataset. We employ the AdamW [18] optimizer with a weight decay of 1e-5 and a learning rate of 1e-3. The training process spans 200 epochs with a batch size of 768, completing in approximately 7 to 8 hours.

4. Evaluation

4.1. Generation Results

Figure 1 shows the proposed technique's ability to achieve continuous and effective control over both emotion and content during image generation. Meanwhile, Figure 5 highlights four key capabilities: (a) emotion-content decoupling, where V-A values override emotional cues in the prompt, allowing typically positive concepts to be rendered with negative emotions; (b) compatibility with discrete emotion categories; (c) content-independent generation, where images generated from empty prompts and specified V-A values maintain emotional consistency without semantic constraints; and (d) fine-grained emotional control, demonstrated through V-A increments of 0.2, showcasing the model's sensitivity to subtle emotional variations.

4.2. Comparisons

To estimate the effectiveness of the proposed method, we built four baselines based on existing techniques.

Baselines. We established baselines for comparison using two strategies: (1) directly injecting emotion features into the image generation modules of SDXL, such as UNet, and (2) modifying the input text prompt using emotion features to influence the generated image's content, similar to the proposed method.

As a result, four different baselines were built: (1) *Cross Attention*: inject the emotion features (e_v, e_a) into the UNet in SDXL via its cross-attention mechanism based on IP-Adapter [53]; (2) *Time Embedding*: directly add emotion features (e_v, e_a) to the time embedding of the UNet in SDXL. (3) *Textual Inversion*: use the text inversion technique [6] to embed emotion features (e_v, e_a) into prompt templates with predefined emotion placeholders. (4) *GPT-4+SDXL (GPT-SD)*: use GPT-4 [1] to rewrite the input text according to (v,a) values to generate an emotional SDXL prompt for image generation.

Qualitative Comparison. We evaluate the generated images based on three criteria: (1) the effectiveness of emotion embedding, (2) image-prompt similarity, and (3) the continuity of emotional variations as V-A values change. Figure 6 demonstrates that the baseline methods—Cross Attention, Time Embedding, and Textual Inversion—tend to produce nearly identical outputs regardless of emotional variation. This is because the loss terms of these baselines primarily align low-level image features (e.g., SDXL's latent space), which are highly correlated with prompt content but struggle to capture subtle emotional cues when both prompts and V-A values are provided. Notably, Tex-

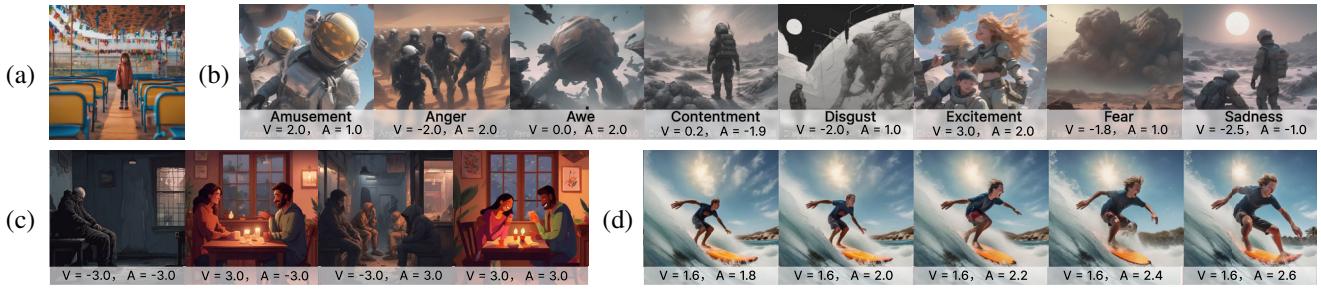


Figure 5. Results under multiple inputs. (a) Overriding semantic content ('a child in the amusement park') with sad V-A (-2,-2); (b) Discrete emotion mapping in V-A space as emotion input; (c) Empty-prompt generation with pure emotion condition; (d) Fine-grained control of V-A variations with a granularity of 0.2.

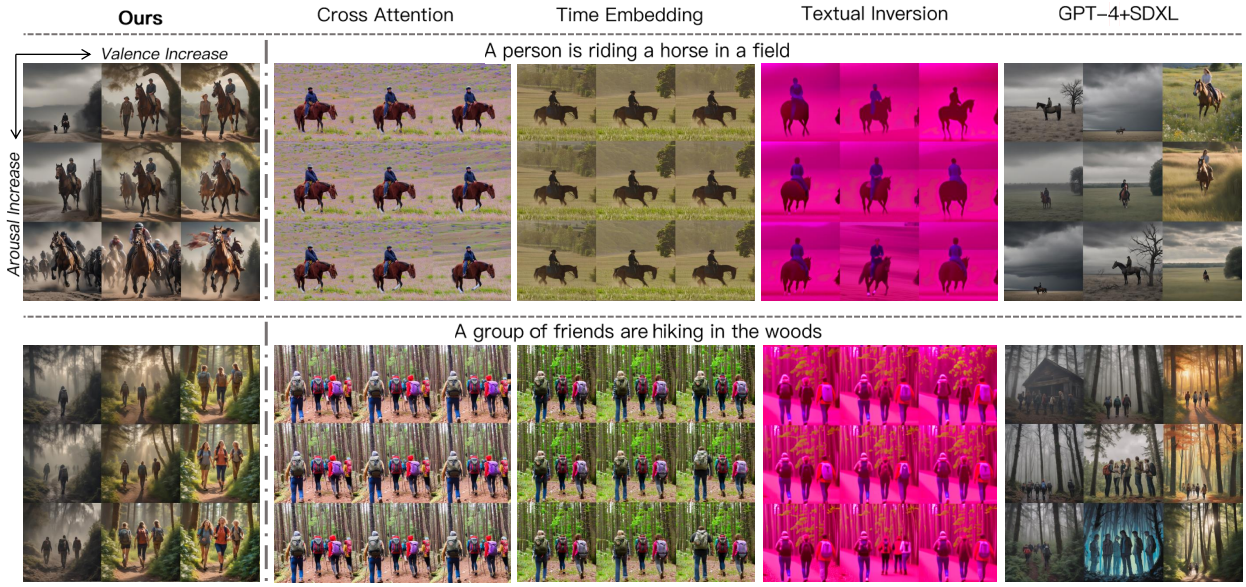


Figure 6. Qualitative comparisons with baselines. These images are generated at varying V-A values, specifically -1.5, 0, and 1.5. Only our approach and the GPT-4+SDXL successfully generate images that clearly reflect emotional variations. Notably, our results show enhanced continuity, indicating superior controllability over continuous V-A values compared to the GPT-4+SDXL.

tual Inversion often produces images with a persistent purple tint. Furthermore, all methods faithfully generate content aligned with the given prompts. However, for image continuity, we focus on comparing our method with GPT-4+SDXL, as they are the only two capable of generating distinct emotional variations. Figure 7 provides a more detailed comparison, showing that our method maintains smooth emotional transitions even under extreme V-A conditions, whereas GPT-4+SDXL introduces noticeable discontinuities (e.g., in the V=3 column).

Quantitative Comparison. We compare our method against several baselines using the following metrics: (1) *V/A-Error* evaluates the absolute error between the predicted V/A of the generated images and the input V/A. (2) *CLIPScore* [8] assesses the similarity between the input text and the generated images. (3) *CLIP-IQA* [41] leverages a pre-trained CLIP model to evaluate image quality without requiring reference images. (4) *LPIPS-Continuous* utilizes the Learned Perceptual Image Patch Similarity [56] to measure the continuity of the change in image as V/A changes.

	A-Error ↓	V-Error ↓	CLIPScore ↑	CLIP-IQA ↑
Cross Attention	1.923±1.153	2.080±1.438	26.266±2.381	**0.949±0.046**
Time Embedding	1.941±1.168	2.031±1.348	**26.566±2.125**	0.786±0.164
Textual Inversion	1.958±1.188	1.923±1.170	22.346±3.594	0.370±0.111
GPT-4+SDXL	1.860±1.090	1.517±1.060	25.907±1.949	0.906±0.066
Ours	**1.828±1.085**	**1.510±1.074**	23.067±2.655	0.881±0.099

Table 1. Comparison on emotion accuracy, prompt fidelity, and image quality across different baselines, evaluated on 3,300 images per method (132 prompts × 5 V values × 5 A values). Our method achieves the highest performance in emotion accuracy while maintaining comparable results in prompt fidelity and image quality. The slight decrease in prompt fidelity is expected, as modifying emotional content affects semantic alignment.

As shown in Table 1, our method achieves the lowest (best) *V/A-Error* on average. While Cross Attention and Time Embedding achieve the highest *CLIP-IQA* and *CLIPScore*, respectively, these methods fail to generate

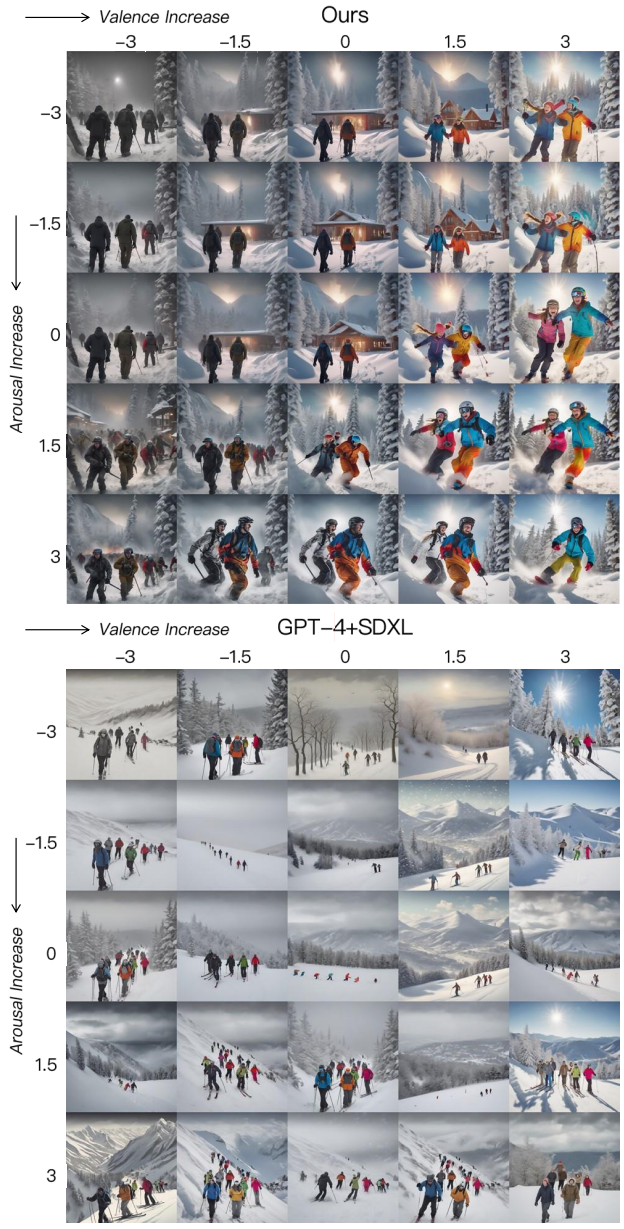


Figure 7. A more comprehensive comparison with the GPT-4+SDXL for the prompt "A group of people is skiing in a snow hill." Our approach maintains continuity even under extreme V/A conditions. Conversely, GPT-4+SDXL displays noticeable discontinuities (e.g., in the V=3 column).

emotionally expressive images (Figure 6).

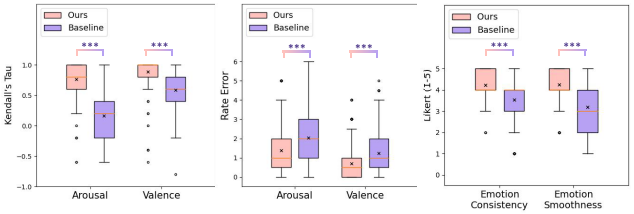
Our method exhibits a slight decrease in *CLIPScore* compared to the baselines, which we attribute to the inherent trade-off between emotional modulation and strict semantic alignment. However, the high *CLIP-IQA* score indicates that our method produces high-quality images. Additionally, as shown in Table 2, our approach demonstrates superior continuity compared to GPT-4 + SDXL.

	Ours	GPT-4+SDXL
LPIPS-Continuous↓	**0.220±0.064**	0.361±0.059

Table 2. Continuity comparison between our method and baseline.

		Ours	**GPT-4+SDXL**
Study I	A-Ranking Consistency ↑	**0.759±0.273**	0.165±0.379
	V-Ranking Consistency ↑	**0.887±0.245**	0.584±0.259
	A-Error ↓	**1.327±1.120**	2.029±1.446
	V-Error ↓	**0.692±0.682**	1.229±1.026
Study II	Emotion Consistency ↑	**4.215±0.715**	3.525±1.065
	Emotion Smoothness ↑	**4.240±0.828**	3.195±1.163

Table 3. Our method outperformed the baseline.



(a) V/A Ranking (Study I) (b) V/A Error (Study I) (c) Emotion Consistency and Smoothness (Study II)

Figure 8. User Study Results (*p<0.05, **p<0.01, ***p<0.001)

4.3. User Study

We conducted two user studies with 20 college students to evaluate the effectiveness of our method by comparing it to *GPT-4+SDXL* (the baseline).

Study I. In the first experiment, we evaluated whether the generated images' emotions align with human perception. Two image collections were prepared—one for Arousal (A) and one for Valence (V)—each comprising 20 sets (10 from our method and 10 from the baseline) of 5 randomized images with varying A or V values. Participants reordered the images by perceived intensity and estimated each image's V and A values. We then computed Kendall's τ_b (with $\tau_b = 1$ indicating perfect alignment) to assess ordering accuracy and calculated the absolute error between the estimated and ground truth values.

Study II. In the second experiment, we assessed whether the generated images could effectively reflect continuous emotional changes (i.e., V-A values). We generated 20 image sets (10 per method), each containing 25 images with V-A values varying gradually from -3 to +3 (Figure 7). Participants rated each set on a 5-point Likert scale regarding (1) the alignment between V-A changes and image content, and (2) the smoothness of the content transition.

Analysis & Results. We conducted a Shapiro-Wilk test [35] to assess normality and applied the Wilcoxon Signed Rank test [44] to evaluate statistical significance. Our results indicate that our method outperformed the baseline across all metrics (Table 3), with statistically significant improvements in V/A Ranking, V/A Error, Consistency, and Smoothness (Figure 8).

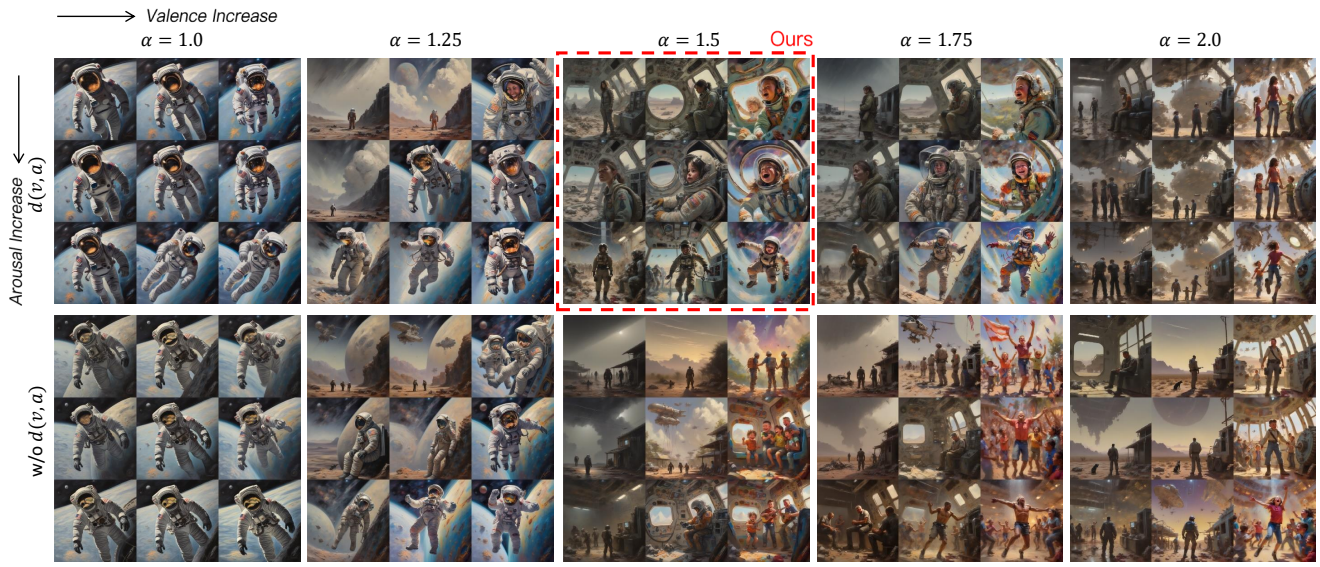


Figure 9. Abalation Study. Images are generated from the prompt "An oil painting shows an astronaut." As α increases, image-prompt similarity decreases, while emotional variations increase. The usage of $d(a, v)$ enhances the accuracy of emotional changes.

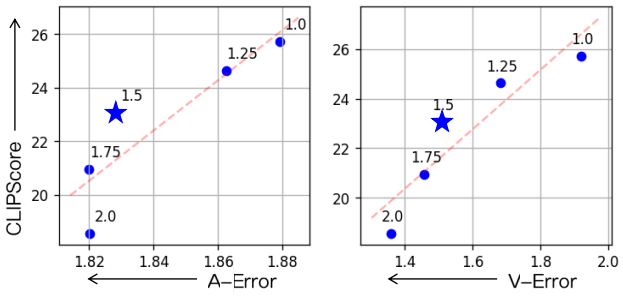


Figure 10. The effectiveness of the scaling factor α. The method with $\alpha = 1.5$ (★) surpasses the performance indicated by the regression line (- - -).

	A-Error ↓	V-Error ↓	CLIPScore ↑
Ours	**1.828±1.085**	**1.510±1.074**	**23.067±2.655**
w/o $d(v, a)$	1.829±1.083	1.546±1.082	21.977±0.066

Table 4. The effectiveness of $d(v, a)$.

4.4. Ablation Study

We performed ablation experiments to assess the contribution and effectiveness of the proposed loss function.

Effectiveness of the Scaling Factor α. We evaluated how varying α affects CLIPScore and V/A-Error (Figure 10). As α increases, CLIPScore decreases (indicating reduced semantic alignment), while V/A-Error also decreases (indicating improved emotional accuracy). This trend is further illustrated by the examples in Figure 9. Based on these findings, we set $\alpha = 1.5$ as an optimal trade-off, though users can adjust α to suit their specific needs.

Effectiveness of the Density Weighting $d(v, a)$. We compare our full method with a variant that omits d(v,a) from the loss function. As shown in Table 4, including d(v,a) leads to improvements in both CLIPScore and V/A-Error. This positive effect is further illustrated in Figure 9.

5. Conclusion and Limitations

In this paper, we introduce continuous emotional image content generation (C-EICG) and present *EmotiCrafter*, a novel method that generates emotionally expressive images using continuous Valence-Arousal (V-A) values. Our emotion-embedding network integrates V-A values into textual features, and extensive experiments show that our approach reliably aligns images with both user prompts and specified emotions. We believe this work will advance affective computing and image generation, and we will release our code and data to foster further research.

Although our method achieved promising results, it still has some limitations need to be addressed in the future. First, controlling image generation based on arousal remains more challenging than controlling based on valence. This is consistent with prior research in visual emotion analysis, which has found that arousal is harder to predict due to lower inter-annotator agreement [21]. Second, our approach frequently generates images featuring human activities even when such activities are not mentioned in the prompts. This likely stems from the limited representation of non-human scenes in our training data and could be mitigated by incorporating a more diverse range of non-human scenarios. Third, our method occasionally modifies users' input prompts to better align with the specified emotional prompts, resulting in a slight semantic shift that affects the generated images. We believe this issue could be addressed by adding a semantic preservation term to the loss function.

6. Acknowledgments

Nan Cao is the corresponding author. This work was supported by the National Key Research and Development Program of China (2023YFB3107100).

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 5

[2] Bita Azari and Angelica Lim. Emostyle: One-shot facial expression editing using continuous emotion parameters. In *Proceedings of WACV*, pages 6385–6394, 2024. 2

[3] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, et al. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8, 2023. 3

[4] Damian Borth, Rongrong Ji, Tao Chen, Thomas Breuel, and Shih-Fu Chang. Large-scale visual sentiment ontology and detectors using adjective noun pairs. In *Proceedings of ACM MM*, pages 223–232, 2013. 2

[5] Yen-Chi Chen. A tutorial on kernel density estimation and recent advances. *Biostatistics & Epidemiology*, 1(1):161–187, 2017. 5

[6] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, et al. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *Proceedings of ICLR*, 2023. 3, 5

[7] Hatice Gunes, Björn Schuller, Maja Pantic, and Roddy Cowie. Emotion representation, analysis and synthesis in continuous space: A survey. In *Proceedings of IEEE international conference on automatic face & gesture recognition*, pages 827–834, 2011. 2

[8] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528, 2021. 6

[9] Stephen Khor Wen Hwooi, Alice Othmani, and Aznul Qalid Md. Sabri. Deep learning-based approach for continuous affect prediction from facial expression images in valence-arousal space. *IEEE Access*, 10:96053–96065, 2022. 2

[10] Ahmed Khota, Eric Cooper, Yu Yan, and Mate Kovacs. Modelling emotional valence and arousal of non-linguistic utterances for sound design support. In *Proceedings of the International Conference on Kansei Engineering and Emotion Research*, pages 507–516, 2022. 2

[11] Dimitrios Kollias. Abaw: Valence-arousal estimation, expression recognition, action unit detection & multi-task learning challenges. In *Proceedings of CVPR*, pages 2328–2336, 2022. 2

[12] Ronak Kosti, Jose M. Alvarez, Adria Recasens, and Agata Lapedriza. Emotion recognition in context. In *Proceedings of CVPR*, pages 1667–1675, 2017. 3, 5

[13] Philip A Kragel, Marianne C Reddan, Kevin S LaBar, and Tor D Wager. Emotion schemas are embedded in the human visual system. *Science advances*, 5(7):eaaw4358, 2019. 2, 3

[14] Benedek Kurdi, Shayn Lozano, and Mahzarin R Banaji. Introducing the open affective standardized image set (oasis). *Behavior research methods*, 49:457–470, 2017. 5

[15] Angela Y Lee and Brian Sternthal. The effects of positive mood on memory. *Journal of consumer research*, 26(2):115–127, 1999. 2

[16] Jiansheng Li, Chuanlan Luo, Qi Zhang, and Rustam Shadiev. Can emotional design really evoke emotion in multimedia learning? *International Journal of Educational Technology in Higher Education*, 17:1–18, 2020. 2

[17] Da Liu, Yaxi Jiang, Min Pei, and Shiguang Liu. Emotional image color transfer via deep learning. *Pattern Recognition Letters*, 110:16–22, 2018. 3

[18] I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5

[19] Olga Megalakaki, Ugo Ballenghein, and Thierry Baccino. Effects of valence and emotional intensity on the comprehension and memorization of texts. *Frontiers in Psychology*, 10:179, 2019. 2

[20] Liyu Meng, Yuchen Liu, Xiaolong Liu, Zhaopei Huang, Yuan Cheng, Meng Wang, Chuanhe Liu, and Qin Jin. Multimodal emotion estimation for in-the-wild videos. *arXiv preprint arXiv:2203.13032*, 2022. 3

[21] Laurent Mertens, Elahe Yargholi, Hans Op de Beeck, Jan Van den Stock, and Joost Vennekens. Findingemo: An image dataset for emotion recognition in the wild. *Advances in Neural Information Processing Systems*, 37:4956–4996, 2024. 2, 3, 5, 8

[22] Ali Mollahosseini, Behzad Hasani, and Mohammad H. Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1):18–31, 2019. 2

[23] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, et al. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453*, 2023. 3

[24] Chanjong Park and In-Kwon Lee. Emotional landscape image generation using generative adversarial networks. In *Proceedings of ACCV*, pages 573–590, 2020. 2

[25] Monika Pawłowska and Ewa Magier-Łakomy. The influence of emotional and non-emotional concepts activation on information processing and unintentional memorizing. *Polish Psychological Bulletin*, pages 150–159, 2011. 2

[26] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of ICCV*, pages 4195–4205, 2023. 3

[27] Kuan-Chuan Peng, Tsuhan Chen, Amir Sadovnik, and Andrew C Gallagher. A mixed bag of emotions: Model, predict, and transfer emotion distributions. In *Proceedings of CVPR*, pages 860–868, 2015. 3

[28] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, et al. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *Proceedings of ICLR*, 2024. 2, 3

[29] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 3

[30] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1 (2):3, 2022. 3

[31] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of CVPR*, pages 10684–10695, 2022. 3

[32] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of CVPR*, pages 22500–22510, 2023. 3

[33] James A Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161, 1980. 2

[34] James A Russell and Albert Mehrabian. Evidence for a three-factor theory of emotions. *Journal of research in Personality*, 11(3):273–294, 1977. 2

[35] S Shaphiro and MBJB Wilk. An analysis of variance test for normality. *Biometrika*, 52(3):591–611, 1965. 7

[36] Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *Advances in neural information processing systems*, 37:84839–84865, 2025. 3

[37] Antoine Toisoul, Jean Kossaifi, Adrian Bulat, Georgios Tzimiropoulos, and Maja Pantic. Estimation of continuous valence and arousal levels from faces in naturalistic conditions. *Nature Machine Intelligence*, 3(1):42–50, 2021. 2

[38] Chai M Tyng, Hafeez U Amin, Mohamad NM Saad, and Aamir S Malik. The influences of emotion on learning and memory. *Frontiers in psychology*, 8:235933, 2017. 2

[39] Ashish Vaswani. Attention is all you need. In *Proceedings of NeurIPS*, page 6000–6010, 2017. 4

[40] Manuelde Vega. The representation of changing emotions in reading comprehension. *Cognition & emotion*, 10(3):303–322, 1996. 2

[41] Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *Proceedings of AAAI*, pages 2555–2563, 2023. 6

[42] Yan Wang, Wei Song, Wei Tao, Antonio Liotta, Dawei Yang, Xinlei Li, et al. A systematic review on affective computing: Emotion models, databases, and recent advances. *Information Fusion*, 83:19–52, 2022. 2

[43] Shuchen Weng, Peixuan Zhang, Zheng Chang, Xinlong Wang, Si Li, and Boxin Shi. Affective image filter: Reflecting emotions from text to images. In *Proceedings of ICCV*, pages 10810–10819, 2023. 3

[44] Frank Wilcoxon. Individual comparisons by ranking methods. In *Breakthroughs in statistics: Methodology and distribution*, pages 196–202. 1992. 7

[45] Weizhen Xie and Weiwei Zhang. Negative emotion enhances mnemonic precision and subjective feelings of remembering in visual long-term memory. *Cognition*, 166:73–83, 2017. 2

[46] Chao Xu, Junwei Zhu, Jiangning Zhang, Yue Han, Wenqing Chu, Ying Tai, et al. High-fidelity generalized emotional talking face generation with multi-modal emotion space learning. In *Proceedings of CVPR*, pages 6609–6619, 2023. 2

[47] Liwen Xu, Zhengtao Wang, Bin Wu, and Simon Lui. Mdan: Multi-level dependent attention network for visual emotion analysis. In *Proeedings of CVPR*, pages 9479–9488, 2022. 2

[48] Dingkang Yang, Shuai Huang, Shunli Wang, Yang Liu, Peng Zhai, Liuzhen Su, et al. Emotion recognition for multiple context awareness. In *Proceedings of ECCV*, pages 144–162, 2022. 2

[49] Jufeng Yang, Dongyu She, and Ming Sun. Joint image emotion classification and distribution learning via deep convolutional neural network. In *Proceedings of IJCAI*, pages 3266–3272, 2017. 2

[50] Jingyuan Yang, Jiawei Feng, and Hui Huang. Emogen: Emotional image content generation with text-to-image diffusion models. In *Proceedings of CVPR*, pages 6358–6368, 2024. 2, 3

[51] Jingyuan Yang, Jiawei Feng, Weibin Luo, Dani Lischinski, Daniel Cohen-Or, and Hui Huang. Emoedit: Evoking emotions through image manipulation. *arXiv preprint arXiv:2405.12661*, 2024. 3

[52] Xiaocui Yang, Shi Feng, Daling Wang, and Yifei Zhang. Image-text multimodal emotion classification via multi-view attentional network. *IEEE Transactions on Multimedia*, 23: 4014–4026, 2021. 2

[53] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023. 3, 5

[54] Ligang Zhang, Dian Tjondronegoro, and Vinod Chandran. Representation of facial expression categories in continuous arousal–valence space: feature and correlation. *Image and Vision Computing*, 32(12):1067–1079, 2014. 2

[55] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of ICCV*, pages 3836–3847, 2023. 3

[56] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of CVPR*, pages 586–595, 2018. 6

[57] Zixing Zhang, Liyizhe Peng, Tao Pang, Jing Han, Huan Zhao, and Björn W Schuller. Refashioning emotion recognition modelling: The advent of generalised large models. *IEEE Transactions on Computational Social Systems*, pages 6690–6704, 2024. 2

[58] Sicheng Zhao, Yue Gao, Xiaolei Jiang, Hongxun Yao, Tat-Seng Chua, and Xiaoshuai Sun. Exploring principles-of-art features for image emotion recognition. In *Proceedings of ACM MM*, pages 47–56, 2014. 2

[59] Sicheng Zhao, Hongxun Yao, and Xiaolei Jiang. Predicting continuous probability distribution of image emotions in valence-arousal space. In *Proceedings of ACM MM*, page 879–882, 2015. 2

[60] Sicheng Zhao, Xingxu Yao, Jufeng Yang, et al. Affective image content analysis: Two decades review and new perspectives. *IEEE TPAMI*, 44(10):6729–6751, 2022. 2

[61] Siqi Zhu, Chunmei Qing, Canqiang Chen, and Xiangmin Xu. Emotional generative adversarial network for image emotion transfer. *Expert Systems with Applications*, 216:119485, 2023. 3