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

Supplementary Material

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

1. Dataset Detail

We present more details of our dataset, including the dataset preparation and the distribution of the emotion values.

Detail of Dataset Preparation. We generate the training set using GPT-4 based on a collection of images with human-annotated V-A values to ensure the content alignment of the generated neutral and emotional prompts.

The GPT-4 prompt for generating neutral prompts based on a given image I is as follows:

> Read the input image carefully. Generate a neutral, concise prompt using a simple subject-verb-object structure with minimal necessary adjectives. Focus on the primary subject and its immediate action or setting, without extra details. Keep it straightforward and under 15 words.

The GPT-4 prompt for generating emotional prompts of I is as follows:

> Read the input image carefully. The input image's arousal value is {arousal} and its valence value is {valence}. Both the values are in the range of [-3, 3]. Based on the image, please generate a prompt for text-to-image generation models. The prompt should be concisely, precisely, and emotionally describing the image content. In particular, the prompt should precisely describe the affective aspects of the image, i.e., color, lighting and shadows, composition, facial expressions and body language (if people are in the image), background and setting, texture and surface, contrast and tonality, symbols and icons, motion and blur, content and theme. The prompt should be no more than 75 words.

Distribution of Emotion Values. As shown in Figure 1, we plot the distribution of Valence and Arousal of our collected dataset. The extreme valence or arousal values are rare in the training dataset.

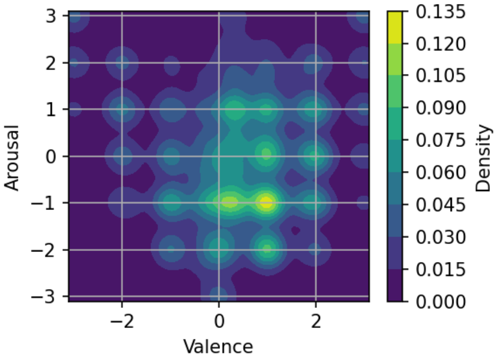


Figure 1. Distribution of V-A.

2. Network Details

We provide additional details on the network implementation. The Emotion Injection Transformer (EIT) has a feature dimension of 768, with V/A features represented by 16 vectors of the same dimension. These features are incorporated into the text features via a cross-attention mechanism. The V-A encoder employs a two-layer MLP with a 256-dimensional latent space, mapping V/A features to a 768-dimensional representation. The Emotion Injection Transformer (EIT) is implemented based on the GPT-2 architecture, with the causal mask removed and transformer blocks adapted into our Emotion Injection Block (EIB). For P_{out}, we apply a linear transformation from 768 to 2048, followed by Layer Normalization with an affine transformation to scale residual features effectively.

3. Experiment Details

We provide further details on the experimental evaluation, covering baseline implementation, test set construction, and the definition of evaluation metrics.

3.1. Comparative Baselines

We compared our method with four baselines and provided details on their implementation in this section.

Cross Attention. Following IP-Adapter [7], we concatenate the V-feature e_v, the A-feature e_a and the prompt feature f_n to form a conditional feature. This feature is fed into the cross-attention mechanism of the SDXL UNet, enabling emotion injection during image generation.

Time Embedding. In this baseline, emotion features (e_v, e_a) are directly embedded into the time embedding component of the SDXL UNet. Specifically, the features are first averaged and then added to the time embedding to incorporate emotional information into the generation process.

Textual Inversion. Using the Textual Inversion method by Gal et al. [1], we embed emotions through learnable token embeddings within structured prompts containing an emotion placeholder token $\{E^*\}$. For example, the template "a painting in the emotion of $\{E^*\}$" utilizes the embedding for $\{E^*\}$, constructed by summing emotion features e_v and e_a, to infuse the intended emotion.

GPT-4+SDXL. We leverage GPT-4 to modify input prompts based on specified V-A values before passing them to SDXL for image generation. The prompt for GPT-4 is:

> Given a prompt {input prompt}, please generate the image corresponding to specific combinations of arousal {arousal} and valence {valence} values based on a standardized scale where both arousal and valence range from -3 to 3. On this scale, a valence value of -3 represents extreme negativity (very unpleasant), while a value of 3 indicates extreme positivity (very pleasant). Likewise, an arousal value of -3 denotes extremely low arousal (very calm or sleepy), whereas a value of 3 signifies extremely high arousal (very excited or tense). Provide the image description alone, without any additional text or explanation.

3.2. Test Image Generation Procedure

We generated a total of 3,300 test images using 132 diverse prompts, encompassing various human scenes, as well as object scenes. For each prompt, we varied both V and A across five values: -3, -1.5, 0, 1.5, and 3. This combination of V and A values resulted in a total of 25 distinct images per prompt. These images served as the basis for our comparative analysis across different metrics.

3.3. Evaluation Metrics

We employed the following metrics to assess the performance of our method:

V/A-Error. This metric measures the absolute difference between the predicted V/A values of generated images and the input V/A values. Utilizing the pre-trained V/A predictor from FindingEMO [4], the errors are calculated as:

$$\text{V-Error} = \frac{1}{n}\sum_{i=1}^{n} |v_i - \hat{v}_i| \tag{1}$$

$$\text{A-Error} = \frac{1}{n}\sum_{i=1}^{n} |a_i - \hat{a}_i| \tag{2}$$

where n is the number of images, v_i/a_i are the input V/A values, and $\hat{v}_i/\hat{a}_i$ are the predicted values. Lower errors indicate closer alignment with the intended emotions.

CLIPScore. To evaluate the semantic alignment between input prompts and generated images, we employed CLIPScore [2], defined as:

$$\text{CLIPScore} = \frac{1}{n}\sum_{i=1}^{n} \max(0, 100\cos(T_i, I_i)) \tag{3}$$

where T_i and I_i represents the CLIP features of input text prompt and the corresponding generated image. Higher scores denote better correspondence between the textual prompts and visual content.

CLIP-IQA. For assessing image quality without image distribution, we utilized CLIP-IQA [5]. This metric leverages a pre-trained CLIP model to evaluate the vision quality of the images. We calculated the average CLIP-IQA score over the generated images, with higher scores reflecting superior image quality.

LPIPS-Continuous. We measured the continuity of image transitions relative to changes in V/A values using the Learned Perceptual Image Patch Similarity (LPIPS) metric [9]. This was calculated as:

$$\text{LPIPS-Continuous} = \frac{1}{2}(L_V + L_A) \tag{4}$$

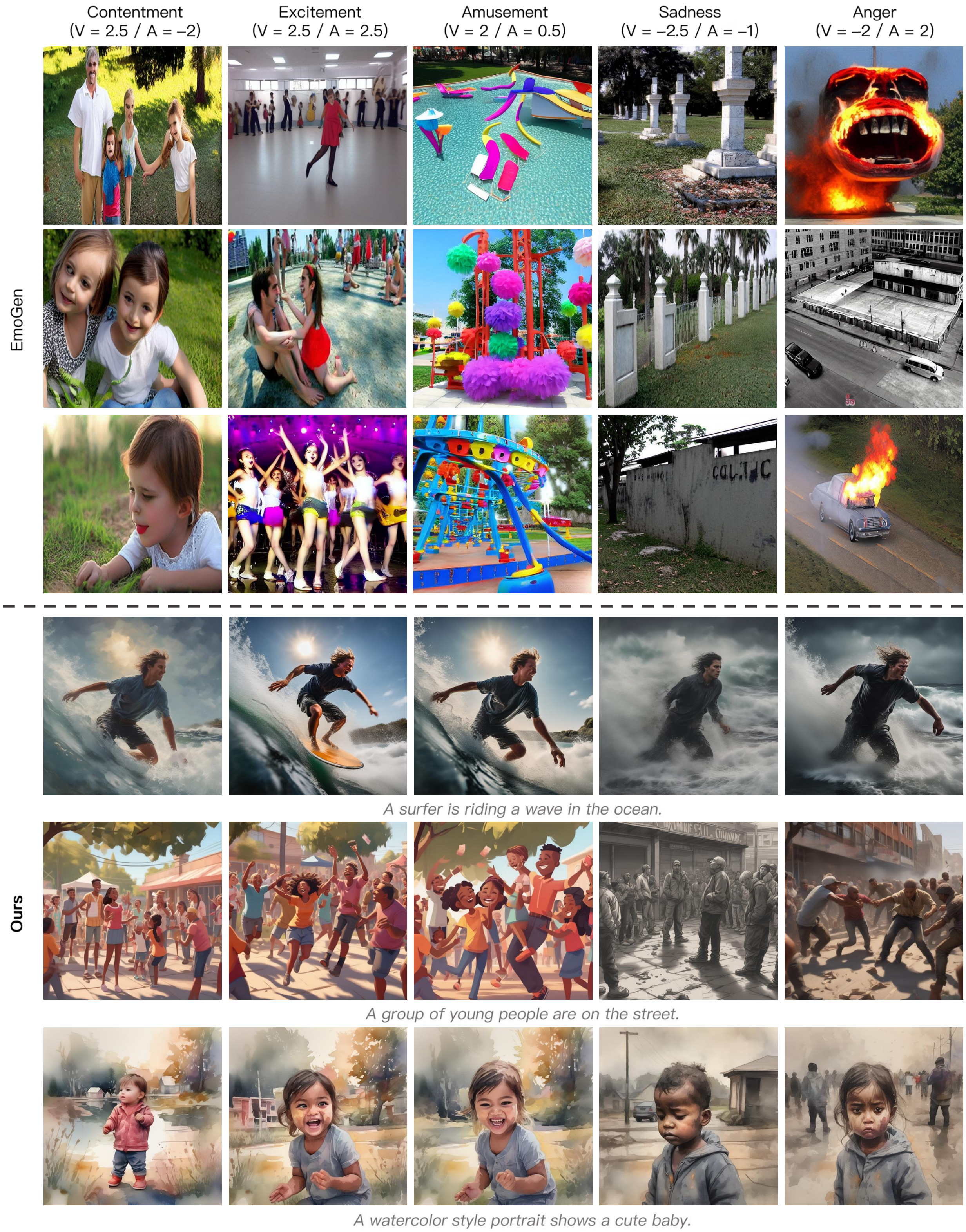
$$L_V = \frac{1}{M}\sum_{v_i,a_i} \text{LPIPS}(I_{v_i,a_i}, I_{v_i+h,a_i}) \tag{5}$$

$$L_A = \frac{1}{M}\sum_{v_i,a_i} \text{LPIPS}(I_{v_i,a_i}, I_{v_i,a_i+h}) \tag{6}$$

where $I_{v,a}$ is the generated image with specific V/A values input, M is the number of comparisons, and h is the increment in V/A values. Lower LPIPS-Continuous scores indicate smoother transitions between images.

4. User study

We provide a more detailed description of the task and procedure of the user study.

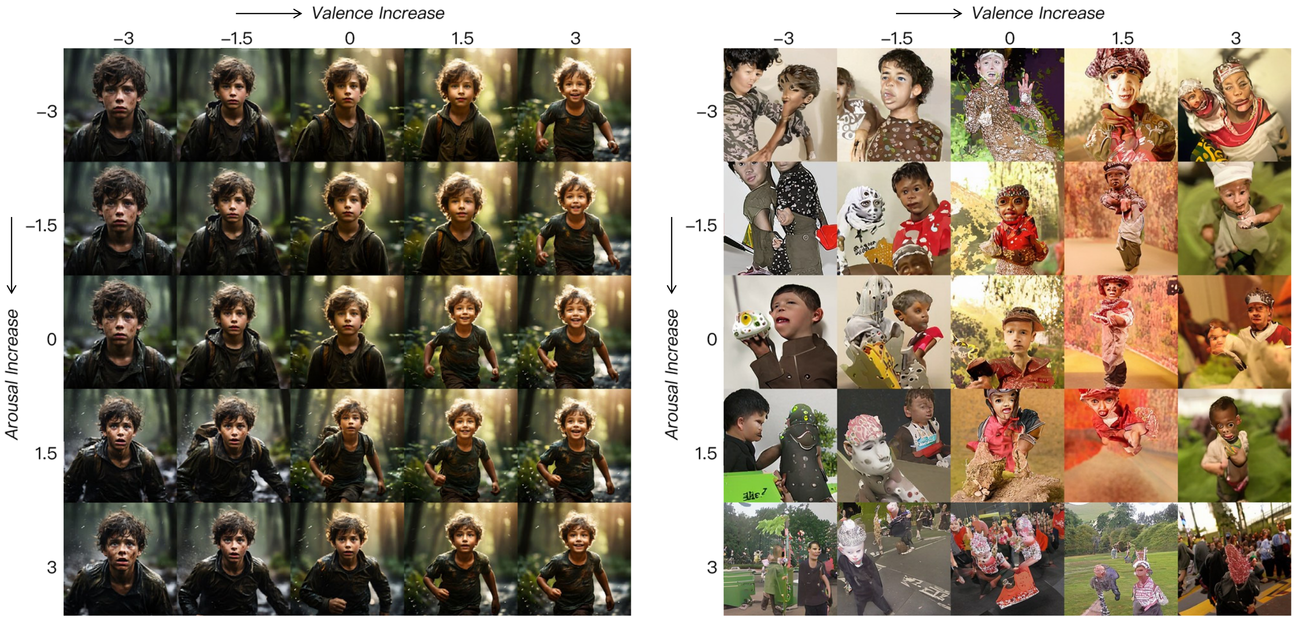


A surfer is riding a wave in the ocean.

A group of young people are on the street.

A watercolor style portrait shows a cute baby.

Figure 2. Comparisons with EmoGen. EmoGen focuses solely on generating images based on emotional categories. However, it lacks the capability to generate images based on subtle emotion change and does not support text-to-image generation. Our method addresses this challenge by enabling control over both free-text prompt and continuous emotion.



(a) Playground v2.5

(b) VAR-CLIP

Figure 3. Comparison of generation results using the prompt “A boy is running in the forest.” across different model architectures. (a) Playground v2.5 successfully generates images with our emotion-embedding network. (b) VAR-CLIP fails when integrating our method.

4.1. Task

We conducted a comprehensive user study with 20 college students to evaluate the effectiveness of our emotional content generation approach. The study assessed four key metrics: (1) **emotion ranking accuracy**, which assessed how accurately participants ranked images based on Arousal or Valence; (2) **emotion rating accuracy**, which evaluated how closely participants' ratings of arousal and valence matched the expected values; (3) **emotion consistency**, measuring the alignment of emotional changes across images with the valence-arousal axis; and (4) **emotion smoothness**, assessing the smoothness of emotional transitions between consecutive images.

4.2. Procedure

Prior to the formal experiment, participants received training on the concepts of arousal and valence. They then completed a qualification task, ranking five images to demonstrate their understanding. The main experiment consisted of two studies via an online questionnaire.

Study I: Ranking and Rating Assessment. Participants evaluated two collections of images specifically designed to test perception of Arousal (A) and Valence (V) values. Each collection contained 20 image sets (10 from our model and 10 from the baseline), with each set containing five images generated using different A or V values. Participants reordered randomly presented images based on perceived emotional intensity and rated each image on a scale from -3 to 3. We employed Kendall's τ_b correlation coefficient to measure the alignment between participant rankings and intended emotional values, with values ranging from -1 (completely reversed order) to 1 (identical order). Absolute error was used to quantify rating accuracy.

Study II: Consistency and Smoothness Evaluation. Participants assessed 20 image sets (10 per method), each containing 25 images generated with V-A values gradually transitioning from -3 to +3. Using a 5-point Likert scale, participants rated: (1) how well emotional changes aligned with the valence-arousal framework, and (2) the smoothness of emotional transitions between images. We analyzed these ratings through comparative mean and variance analysis between our method and the baseline.

5. Comparsion with EmoGen

We qualitatively compare our approach to EmoGen [6], a state-of-the-art model for emotional image generation that utilizes discrete emotion labels to control image outputs. For a fair comparison, we selected five emotion categories—excitement, amusement, contentment, anger, and sadness—and mapped them onto the Valence-Arousal Cartesian space to generate images.

As illustrated in Figure 2, EmoGen generates images solely based on single emotion labels without the ability to incorporate specific content through text prompts. This limitation results in a strong association between emotions and particular semantic elements; for example, ”sadness” often results in images featuring tombstones, while ”anger” typically includes elements like fire. Such tight coupling restricts the model’s flexibility and contextual adaptability.

In contrast, our model maintains a faithful representation of the input text prompts, ensuring semantic accuracy while flexibly modulating emotional expressions. By leveraging a continuous emotion model within the V-A framework, our method allows for more nuanced and context-sensitive image generation. Even when restricted to discrete emotions, our approach surpasses existing models by providing greater control over both content and emotion.

6. Compatibility

Our framework embeds V-A values into textual features while keeping the SDXL’s image generation module fixed. By directly leveraging SDXL’s established textual feature space, our approach ensures compatibility with SDXL-based architectures, allowing integration with various SDXL derivatives (e.g. Playground v2.5 [3]).

Compatibility with SDXL Variants. We tested our method with Playground v2.5 [3], an SDXL-based model enhanced for aesthetic quality. Utilizing the pretrained emotion-embedding network with $\alpha = 1$, we successfully embedded emotional information. As shown in Figure 3 (a), our VA embedding effectively controls both emotion and content. Additionally, Playground v2.5’s aesthetic optimizations result in higher-quality generated images.

Incompatibility with non-SDXL Variants. We conducted experiments with VAR-CLIP [8], a VAR-based text-to-image model, to evaluate our method’s cross-architecture compatibility. Unlike SDXL, which utilizes penultimate-layer CLIP features, VAR-CLIP employs final-layer pooled CLIP feature as textual feature. Despite our attempts to retrain the emotion embedding network, the resulting generations fail to demonstrate effective emotion control (Figure 3 (b)). The results indicate that our V-A integration may be incompatible with non-SDXL architectures.

7. Emotion Interpolation

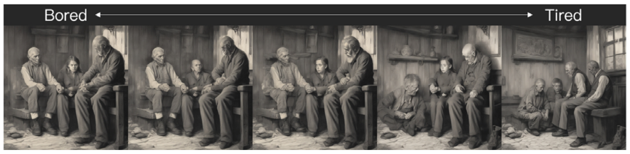


Figure 4. Interpolation between Bored and Tired.

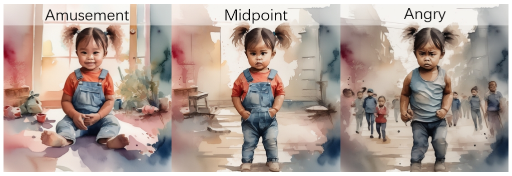


Figure 5. In-between state of Amusement and Angry.

We provide additional interpolation results between two discrete emotions, Bored (V = -1.0, A = -2.1) and Tired (V = -0.8, A = -2.075), in Figure 4. The result shows our method can achieve continuous emotional transitions. Continuous emotion control could deliver richer expressivity for film production or personalized content creation. In Figure 5, we interpolate between Amusement (V = 2.0, A = 1.0) and Anger (V = –2.0, A = 2.0) at the midpoint (V = 0, A = 1.5). The midpoint generated image exhibits an “in-between” state.

References

[1] 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. 2

[2] 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. 2

[3] Daiqing Li, Aleks Kamko, Ehsan Akhgari, Ali Sabet, Linmiao Xu, and Suhail Doshi. Playground v2. 5: Three insights towards enhancing aesthetic quality in text-to-image generation. *arXiv preprint arXiv:2402.17245*, 2024. 5

[4] 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

[5] 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. 2

[6] 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. 4

[7] 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. 1

[8] Qian Zhang, Xiangzi Dai, Ninghua Yang, Xiang An, Ziyong Feng, and Xingyu Ren. Var-clip: Text-to-image generator with visual auto-regressive modeling. *arXiv preprint arXiv:2408.01181*, 2024. 5

[9] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep fea-

tures as a perceptual metric. In *Proceedings of CVPR*, pages 586–595, 2018. 2