EmoGen: Emotional Image Content Generation with Text-to-Image Diffusion Models

Jingyuan Yang, Jiawei Feng, Hui Huang*
Shenzhen University
{jingyuanyang.jyy, fengjiawei0909, hhzhiyan}@gmail.com

Figure 1. Emotional Image Content Generation (EICG). Given an emotion category, our network produces images that exhibit unambiguous meanings (*semantic-clear*), reflect the intended emotion (*emotion-faithful*) and incorporate varied semantics (*semantic-diverse*).

Abstract

Recent years have witnessed remarkable progress in image generation task, where users can create visually astonishing images with high-quality. However, existing text-to-image diffusion models are proficient in generating concrete concepts (dogs) but encounter challenges with more abstract ones (emotions). Several efforts have been made to modify image emotions with color and style adjustments, facing limitations in effectively conveying emotions with fixed image contents. In this work, we introduce Emotional Image Content Generation (EICG), a new task to generate semantic-clear and emotion-faithful images given emotion categories. Specifically, we propose an emotion space and construct a mapping network to align it with the powerful Contrastive Language-Image Pre-training (CLIP) space, providing a concrete interpretation of abstract emotions. Attribute loss and emotion confidence are further proposed to ensure the semantic diversity and emotion fidelity of the generated images. Our method outperforms the state-of-the-art text-to-image approaches both quantitatively and qualitatively, where we derive three custom metrics, i.e., emotion accuracy, semantic clarity and semantic diversity. In addition to generation, our method can help emotion understanding and inspire emotional art design. Project page: https://vcc.tech/research/2024/EmoGen.

1. Introduction

"What I cannot create, I do not understand."
–Richard Feynman

Emotions, often elusive yet profoundly influential, shape our actions, foster connections, and spark passions. With the prevalence of social medias, users tend to share specially crafted images to express their feelings. Aiming to find out

*Corresponding author

people's emotional responses towards different stimuli, Visual Emotion Analysis (VEA) is an intriguing yet challenging task in computer vision [34, 50, 51]. Recent years have witnessed rapid development in this field, bringing potential applications such as opinion mining [48], market advertising [6] and mental healthcare [17].

Thanks to the advent of diffusion models [7, 16, 38], unprecedented progress has been made in text-to-image generation, where users can generate high-quality images with crafted prompts or personalized objects [10, 39, 56]. Existing text-to-image diffusion models, are often excel in generating *concrete* concepts (*e.g.*, *cat*, *house*, *mountain*) but face limitations when tasked with more *abstract* ones (*e.g.*, *amusement*, *anger*, *sadness*). In reality, however, photographic works are not necessarily targeted on specific entities, but are often composed to convey certain feelings.

A natural question arises: *What if machines could create images that not only please our eyes but also touch our hearts?* Generating emotions is very challenging. Emotions are abstract while images are concrete, leaving the affective gap [13] hard to surmount. To bridge the gap, several efforts have been made to modify visual emotions by adjusting colors and styles, *i.e.*, image emotion transfer [30, 41, 47]. These methods, however, meet difficulties in evoking emotions correctly and significantly, *i.e.*, 29% emotion accuracy [47], as fixed image contents limit emotional variations. Moreover, we cannot generate emotional images solely from colors and styles. What truly triggers emotion? Psychological studies show that visual emotions are often evoked by specific semantics [1, 3, 4].

In this paper, we propose Emotional Image Content Generation (EICG), a new task to generate semantically clear and emotionally faithful visual contents conditioned on a given emotion category, as shown in Figure 1. Semantic clarity demands an unambiguous representation of visual contents, while emotion faithfulness entails generating images evoke the intended emotions. Contrastive Language-Image Pre-training (CLIP) [31] is a large-scale vision-language model with rich semantics. However, we observe in Figure 2 that CLIP space can not well capture emotional relationships. Therefore, we introduce an emotion space, which groups similar emotions together while keeping dissimilar ones apart. While emotion space excels in representing emotions, CLIP space exhibits a powerful semantic structure. To align emotion space with CLIP space, we propose a mapping network, interpreting abstract emotions with concrete semantics.

EmoSet [53] is a recently proposed large-scale visual emotion dataset with rich attributes. The Latent Diffusion Model (LDM) loss [38] is often utilized to optimize concrete entities with single and explicit semantics, posing a challenge in capturing the diversity within each emotion. To address this, we introduce an attribute loss to ensure semantic clarity and diversity, by leveraging the attribute labels in EmoSet. Recognizing that not all objects are affective, emotion confidence is further proposed to ensure the emotion fidelity of the generated contents.

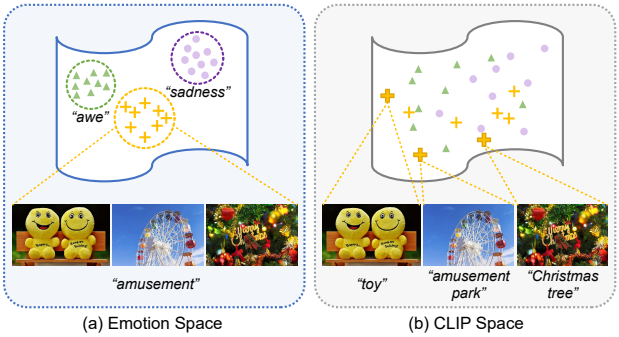


Figure 2. Despite (b) CLIP space demonstrates a powerful semantic structure, it struggles to effectively capture emotional relationships within (a), the proposed emotion space.

To estimate the generation quality of EICG, three evaluation metrics are specially designed: emotion accuracy, semantic clarity and semantic diversity. As EICG aims to create emotional contents, we design emotion accuracy to measure the alignment between intended and perceived emotions in the generated images. People are prone to evoke emotions only when the contents are easily recognizable. Thus we propose semantic clarity to assess the unambiguity of the generated image content. Additionally, in view of the assorted emotion stimuli, we devise semantic diversity to quantify the content richness under each emotion. We evaluate our method through both qualitative and quantitative analyses, surpassing the state-of-the-art text-to-image generation approaches across five metrics. Ablation studies are performed to verify the network design, and user studies are conducted to resonate our method with human viewers. Besides generation task, our method can also be applied to decompose emotion concepts, transfer emotional contents and fuse different emotions, which may be helpful to understand emotions and create emotional art design.

In summary, our contributions are:

- We introduce Emotional Image Content Generation, a novel task to generate emotion-faithful and semantic-clear image contents. We also derive three custom metrics to estimate the generation performance.
- We develop a mapping network to align the proposed emotion space to the powerful CLIP space, where attribute loss and emotion confidence are further designed to ensure the semantic richness and emotion fidelity.
- We evaluate our method against the state-of-the-art text-to-image approaches and demonstrate our superiority. Potential applications are exhibited for emotion understanding and emotional art design.

2. Related work

2.1. Visual Emotion Analysis

Researchers have been involved in VEA for over two decades, ranging from early traditional approaches [2, 23, 26] to recent deep learning ones [35, 51, 52, 58]. Given the inherent abstractness and complexity of visual emotion, researchers aim to identify the most influential elements, which range from low-level features like color, texture and style [23, 26, 35, 58] to high-level semantics [2, 35, 51, 52, 58]. Lee *et al*. [23] propose a scheme to evaluate emotional response from color images by reasoning the prototypical color for each emotion and the input images. As a milestone, Machajdik *et al*. [26] extract representative low-level features in composition, including color and texture, to predict visual emotions. Besides low-level features, Borth *et al*. [2] propose Adjective-Noun Pair (ANP) and build a visual concept detector named Sentibank. With the help of deep learning techniques, Rao *et al*. [35] construct MldrNet to extract emotional clues from pixel-level, aesthetic-level and semantic-level. To form a more discriminative emotional representation, Zhang *et al*. [58] integrate high-level contents and low-level styles. Yang *et al*. propose network to mine emotions from multiple objects [52] as well as object-scene correlations [51]. Existing work often treat VEA as a classification task, *i.e*., input an image and predict the emotion within it. Can we reverse this process? In other words, can we generate an image targeting on the given emotion word? Only by creating emotional images can we demonstrate the significance of visual elements, leading to a better understanding of emotions.

2.2. Text-to-Image Generation

Text-to-image generation aims to convert textual descriptions into corresponding realistic images. Existing generative models can be grouped into GANs [12, 24, 59], VAEs [9, 20, 55], flow-based [37], energy-based [22] and diffusion-based [7, 16, 38, 39, 56]. Diffusion models are witnessed impressive and rapid progress in recent years, where methods like GLIDE [28], DALLE2 [32], Imagen [40] can generate diverse, photo-realistic and high-quality images. Notably, Stable diffusion [38] is one of the most popular diffusion models, owing to its stable training and the capability for fine-grained control, supported by an active user community. For customized generation, several diffusion-based text-to-image methods are introduced, where methods vary from learning a new embedding [8, 10] and finetuning the network parameters [21, 39, 46]. Textual inversion [10] and DreamArtist [8] learn new concepts with a few user-provided images in the word embedding space, without further training on diffusion models. While DreamBooth [39] finetunes all the parameters to learn a new concept, Custom diffusion [21] only updates the key and value parameters in the cross attention layers. Further, ELITE [46] speeds up the running time with accurate generation results by adopting a global and local mapping network. Existing text-to-images models are capable of generating concrete concepts [7, 24, 56], or personalized ones [10, 21, 39], but face difficulties in generating more abstract ones. In reality, photographic works are not necessarily composed of targeted concepts, but often aim to convey specific feelings. Thus, how to generate emotion-evoking images remains a pressing and critical challenge.

2.3. Image Emotion Transfer

Image style transfer [11] aims to render the semantic content under different styles, producing visually stunning results [19, 33, 45, 54]. Similarly, image color transfer [36] seeks to adjust and harmonize the color characteristics of one image to match another [18, 29]. Specifically, color and style choices can strongly influence the emotions of an image [27]. By adjusting low-level visual elements, image emotion transfer aims to modify the emotional tone of the input image, including the color-based methods [5, 25, 30, 44, 49, 60] and the style-based ones [41, 47]. Yang and Peng *et al*. [49] makes the first attempt to transfer image colors. Wang *et al*. [44] present a system to modify the image color according to a given emotion word, and Liu *et al*. [25] further advance it with deep learning techniques. Peng *et al*. [30] introduce a new approach to alter the emotion of an input image by guiding its color and texture under the target image. More recently, to reflect emotions in styles, Sun *et al*. [41] and Weng *et al*. [47] bring promising results on emotion-aware image style transfer. Nevertheless, the alteration of visual emotions through colors and styles is limited due to fixed content, resulting in subtle emotional changes, *i.e*., 29% emotion accuracy in [47]. Psychological studies suggest that visual emotions can be elicited by specific semantics [3]. Thus, we propose a novel method to generate emotional image contents with clear semantics.

3. Method

3.1. Emotion Representation

Emotion Space EICG is a challenging task, which requires both semantic clarity and emotion fidelity. How to generate an image with distinct and emotional semantics? CLIP [31] is developed to align image and text modalities, where semantically related features are located in close proximity to each other. While CLIP shows impressive semantic representation capabilities, it struggles to effectively capture emotional relationships. As demonstrated in Figure 2, we can observe that sample points with emotional similarities are distantly separated within the CLIP space due to their differing semantics, *e.g*., *toy*, *amusement park* and *Christmas tree*. To better depict emotional relation-

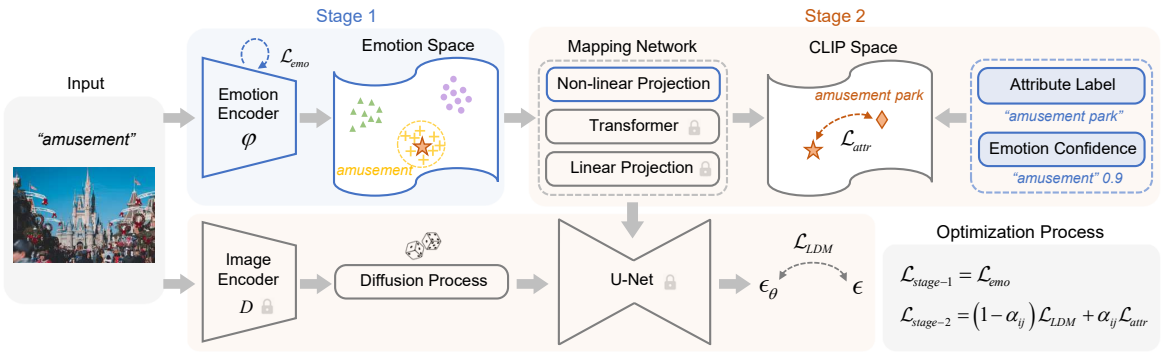


Figure 3. Training process of our network. Emotion representation (stage 1) learns a well-behaved emotion space and emotion content generation (stage 2) maps this space to CLIP space, aiming to generate image contents with emotion fidelity, semantic clarity and diversity.

ships, we introduce the emotion space, a latent space that clusters similar emotions together while keeping dissimilar ones apart. EmoSet [53] is a large-scale dataset with rich attributes, where each image is labeled with an emotion. Using aligned image-emotion pairs, we construct an encoder φ with ResNet-50 [14] to capture emotion representations. To train the encoder, we devise an emotion loss by implementing the widely-used Cross-Entropy (CE) loss, following the previous work [50, 51]:

$$\mathcal{L}_{emo} = -\sum_{i=1}^{C} y_{emo} \log \frac{\exp(\varphi(x,i))}{\sum_{i=1}^{C} \exp(\varphi(x,i))}, \quad (1)$$

where x represents the input image, y_{emo} denotes the emotion label and C stands for the total number of emotion categories. Once the loss function converges, emotion space is established. Parameters in emotion encoder remain fixed in the following emotional content generation process.

During inference, each emotion cluster is represented by a Gaussian distribution with learned parameters, *i.e.*, mean and standard deviation. For example, when taking *amusement* as input, we randomly sample a data point from corresponding Gaussian distribution to serve as its emotion representation, as shown in Figure 3. We have confirmed that Gaussian distribution suits emotion clusters well and the random sampling process induces diversity to EICG.

3.2. Emotional Content Generation

Mapping Network While emotion space is emotionally separable, CLIP space captures rich semantics. Existing text-to-image models entail clear and specific semantics as input, making CLIP space indispensable in the generation process. Consequently, establishing the mapping between emotion space and CLIP space becomes a crucial challenge. Intuitively, we attempt to build the mapping network using fully connected layers, following previous work [33, 43].

However, as depicted in Figure 2, clustered feature points in the emotion space are expected to disperse in the CLIP space to capture diverse semantics. Therefore, we utilize a Multilayer Perceptron (MLP) to build the mapping network, incorporating non-linear operations, *i.e.*, RELU, to facilitate the separation process. The non-linear projection F is succeeded by a CLIP text transformer t_θ, yielding textual embedding for U-Net. The end-token embedding of the transformer's output is passed through a fully-connected layer, producing the CLIP text feature. Particularly, to better preserve the prior knowledge in the CLIP space, parameters in the transformer and linear projection are kept frozen, while parameters in non-linear projection are learned, as depicted in Figure 3.

Attribute Loss Existing text-to-image diffusion models often employ Latent Diffusion Model (LDM) loss [38] for optimization process [10, 39, 56]:

$$\mathcal{L}_{LDM} = \mathbb{E}_{z,x,\epsilon,t}\left[\|\epsilon - \epsilon_\theta\left(z_t, t, t_\theta\left(F\left(\varphi\left(x\right)\right)\right)\right)\|_2^2\right], \quad (2)$$

where ϵ represents the added noise, ϵ_θ denotes the denoising network and z_t indicates the latent noised to time t.

In these cases, target concepts typically involve concrete entities (*e.g.*, *dog*, *car*, *flower*) or personalized objects (*e.g.*, someone's *corgi*). These concepts often exhibit consistency on semantic level and share certain similarities on pixel level. However, emotions are abstract concepts, where multiple semantics coexist under one specific category. Learning emotions solely with LDM loss may pose some challenges. For one thing, each emotion might collapse to a specific semantic point, *e.g.*, *amusement* collapsing to *amusement park*, losing intra-class diversity. In reality, semantics within one emotion are diverse, where single point cannot fully capture. Moreover, since LDM loss is designed to reconstruct the input image, it primarily focuses on learning and preserving pixel-level commonalities such as color and texture. In Figure 4 (a), with LDM loss alone, CLIP embedding for *amusement* is prone to be *colorful*, without exhibiting explicit and diverse semantics. We can conclude that it is hard to achieve robust emotion representations in CLIP space by implementing LDM loss alone.

In the pursuit of clear and diverse contents, semantics guidance is essential for the generation process. Thanks to

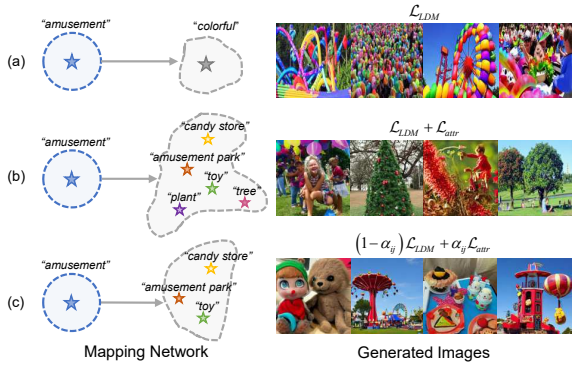


Figure 4. Motivation for loss function design. Compare to (a) LDM loss alone, (b) attribute loss enhances semantic clarity while (c) emotion confidence ensures emotion accuracy.

the rich attribute annotations in EmoSet, we select the mid-level attributes, *i.e.*, object class and scene type to guide the generation process. With this semantic guidance, we formulate an attribute loss to guarantee that the generated image contents possess clear and diverse semantics. For clarity, emotions are easily triggered in people only when visual contents are represented in an unambiguous manner. Considering the varied emotional stimuli in reality, attribute loss guides the network to learn multiple semantics under one specific emotion. Our attribute loss is devised on CLIP space, by calculating the cosine similarities $f(\cdot)$ and optimizing a symmetric CE loss over the similarity scores [31]:

$$\mathcal{L}_{attr} = -\sum_{j=1}^{K} y_{attr} \log \frac{\exp(f(v_{emo}, \tau_\theta(a_j)))}{\sum_{j=1}^{K} \exp(f(v_{emo}, \tau_\theta(a_j)))}, \quad (3)$$

$$f(p, q) = \frac{p \cdot q}{\|p\| \, \|q\|}, \quad (4)$$

where a_j denotes the j member in the attribute set, τ_θ represents the text encoder, v_{emo} implies the learned CLIP embedding and K indicates the total number of the attributes. With the attribute loss, each sample point is converging towards the correct semantic and distancing itself from the incorrect ones. Through the combination of attribute loss and LDM loss, we can effectively map each emotion to clear and diverse semantics, as demonstrated in Figure 4 (b).

Emotion Confidence However, it is worth noting that some of the semantics in Figure 4 (b) appear emotionally neutral, *e.g.*, *plant* and *tree*. Since attributes are annotated objectively, not all the attributes in EmoSet are emotional. Therefore, we propose emotion confidence to measure the correlations between emotions and semantic attributes. Initially, we gather all images associated with attribute j in EmoSet and send them to a pre-trained emotion classifier. Each image is predicted as an emotion vector $p(\cdot)$ and we

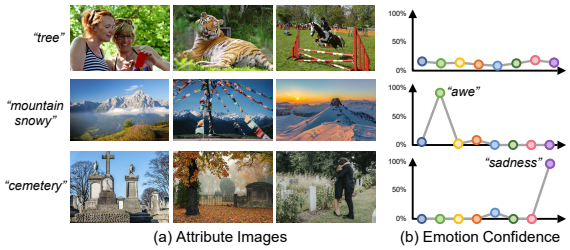


Figure 5. Illustration of emotion confidence. Each (a) attribute is represented by (b) a distribution of confidence on eight emotions.

sum all images up to get the emotional distribution d_j for attribute j. Each emotion i within this distribution is assigned a corresponding emotion confidence α_{ij}:

$$\alpha_{ij} = \frac{1}{N_j} \sum_{n=1}^{N_j} p(x_n, i), \quad (5)$$

where x_n represents the input image and N_j denotes the total image number in attribute j. We further illustrate the above process in Figure 5 with visual representations. When *mountain snowy* appears, people are more likely to experience *awe* and *cemetery* often elicits *sadness*. In contrast, the presence of *tree* in every emotion category suggests its lack of emotional specificity. Some attributes are emotion-related while others are not, which can be beneficial for generating emotional contents. We then use emotion confidence to balance between LDM loss and attribute loss:

$$\mathcal{L} = (1 - \alpha_{ij}) \mathcal{L}_{LDM} + \alpha_{ij} \mathcal{L}_{attr}, \quad (6)$$

where i represents the emotion category y_{emo} and j denotes the attribute type y_{att}. The greater the emotion confidence α_{ij} is, the stronger the impact attribute j has on the specific emotion i. Low confidence suggests a weak connection between the attribute and emotion, signaling that the network should learn more from the pixel-wise LDM loss. When higher confidence occurs, the network should prioritize the semantic meaning of the image, *i.e.*, the attribute loss. With this design, our network can adapt to a wide range of cases, generating image contents that are both semantically explicit and emotionally faithful, as shown in Figure 4 (c).

4. Experiments

4.1. Dataset and Evaluation

Dataset EmoSet [53] is a large-scale visual emotion dataset with rich attributes, comprising a total of 118,102 images. To investigate the connections between emotions and specific contents, we create a subset from EmoSet by preserving images with object/scene labels. Each image is labeled with both emotion and attribute labels, guiding the optimization process of emotion loss and attribute loss. Notably, the wide range of attribute labels assures for learning diverse and representative emotional contents.

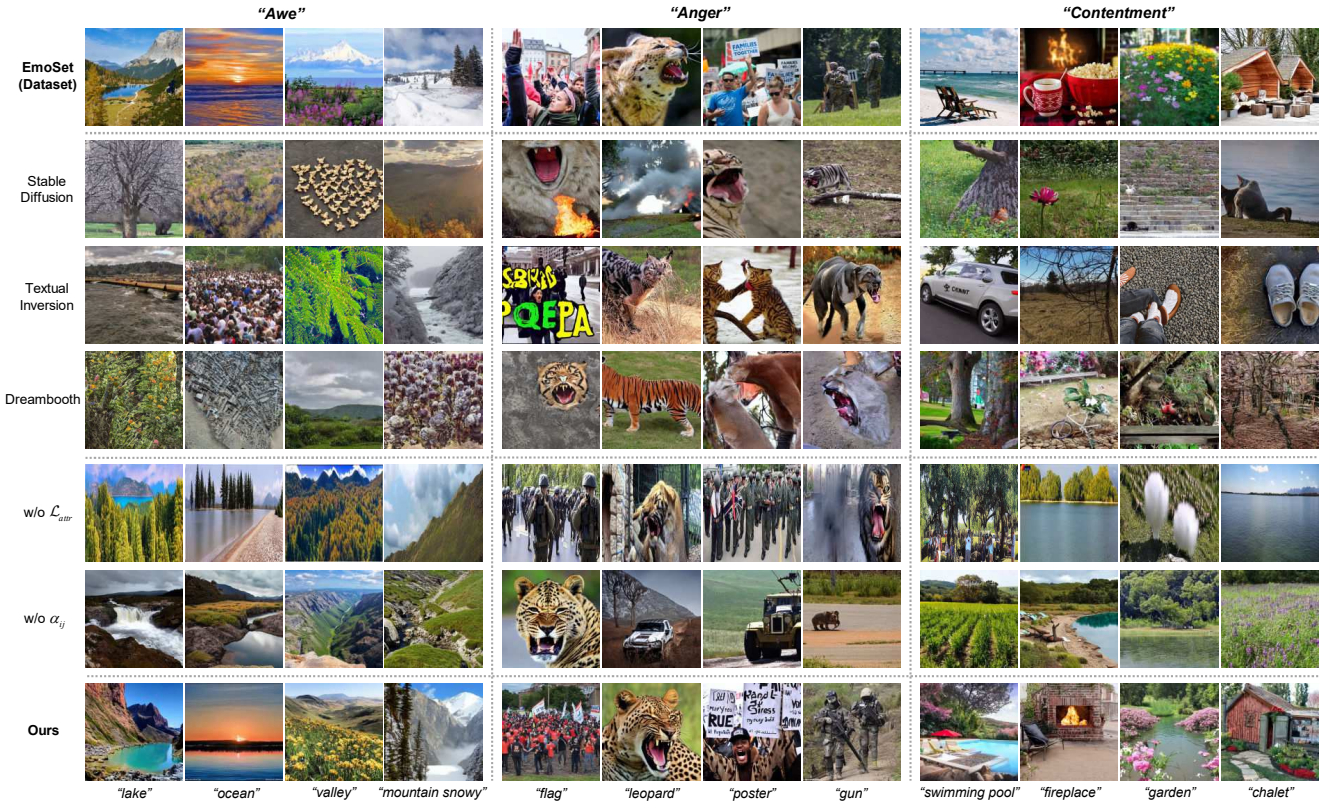


Figure 6. Qualitative comparisons with the state-of-the-art text-to-image generation approaches and ablation studies of our method.

Evaluation Metrics To comprehensively evaluate the performance of different methods on EICG task, we utilize commonly used metrics (FID, LPIPS) and design some specific ones (Emo-A, Sem-C, Sem-D). 1) **FID**: Frechet Inception Distance (FID) [15] quantifies the distribution distance between generated and real images, providing an estimate of image fidelity. 2) **LPIPS**: Similar to [42], we employ LPIPS [57] to assess the overall image diversity, with higher values indicating better performance. 3) **Emo-A**: Since EICG aims at creating emotion-evoking images, we design emotion accuracy to assess the emotional alignment between the targeted emotions and the generated images. 4) **Sem-C**: People are easily to evoke emotions under recognizable contents. We thus introduce semantic clarity to assess the explicitness of generated image contents. 5) **Sem-D**: Emotions are complex, where each can be triggered by multiple factors. To cover a diverse range of potential scenes or objects, we derive semantic diversity to estimate the content richness associated with each emotion. For more details, please refer to the supplementary materials.

4.2. Comparisons

As our method is the first attempt in EICG, we compare it with the most relevant and state-of-the-art text-to-image generation techniques: Stable diffusion [38], Textual inversion [10] and Dreambooth [39]. While Stable diffusion is a general image generation pipeline, Textual inversion and Dreambooth specialize in customized image generation.

Table 1. Comparisons with the state-of-the-art methods and ablation studies on emotion generation task, involving five metrics.

Method	FID ↓	LPIPS ↑	Emo-A ↑	Sem-C ↑	Sem-D ↑
Stable Diffusion [38]	44.05	0.687	70.77%	0.608	0.0199
Textual Inversion [10]	50.51	0.702	74.87%	0.605	0.0282
DreamBooth [39]	46.89	0.661	70.50%	0.614	0.0178
Ours	**41.60**	**0.717**	**76.25%**	**0.633**	**0.0335**
w/o F	57.54	0.713	71.12%	0.615	0.0261
w/o $\mathcal{L}_{attr}$	51.13	0.707	65.75%	0.592	0.0270
w/o α_{ij}	43.30	0.714	74.88%	0.591	0.0263

Qualitative Comparisons In Figure 6, our method is qualitatively compared with the state-of-the-art methods across three emotion categories, *i.e.*, *awe*, *anger* and *contentment*. Generation results of the rest five emotions can be found in the supplementary materials. Take *awe* as an example, all the three compared methods tend to produce images with dense textures and dim colors, which suggests that representations for each emotion may collapse to a single feature point. For *anger* and *contentment*, both Stable diffusion and Dreambooth distort the visual representations, *e.g.*, *tiger* and *bicycle*, and generate some contents with ambiguous semantics. Though Textual inversion preserves some semantic fidelity, it generates emotion-agnostic

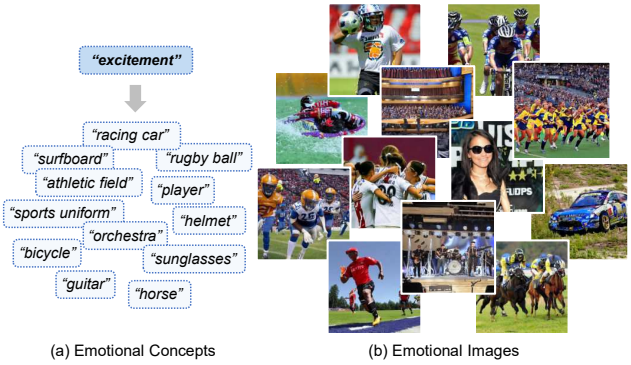
Table 2. User preference study. The numbers indicate the percentage of participants who prefer our results over those compared methods, given the same emotion category as input.

Method	Image fidelity ↑	Emotion faithfulness ↑	Semantic diversity ↑
Stable Diffusion	67.86±15.08%	73.66±11.80%	87.88±9.64%
Textual Inversion	79.91±16.92%	72.75±16.90%	85.66±10.51%
DreamBooth	77.23±14.00%	80.79±8.64%	81.68±17.06%

contents such as *shoes* and *cars*. Since these methods are crafted to learn customized concepts, challenges may arise when handling complex and diverse emotional images. Rather than generating *plants* and *trees*, our method can provide diverse and emotion-evoking image contents for *awe* through *lakes*, *oceans*, *valleys* and *snow-covered mountains*. In *anger*, our approach extends beyond mere *beasts*, encompassing *flags*, *posters*, and *guns*. Owing to attribute loss and emotion confidence, our method can effectively capture the rich and varied semantics while maintaining emotion faithfulness in EmoSet.

Quantitative Comparisons As shown in Table 1, the proposed method surpasses the compared methods across all five evaluation metrics. Particularly, better performance on FID and LPIPS indicates our method can generate images with higher fidelity and diversity, effectively capturing the characteristics of the training data. All methods achieve comparable results on emotion accuracy. From Figure 6, we observe that comparison methods are prone to fall into singular or incorrect emotion representations. Even such generation results are still separable in eight classes, they do not conform to human emotional cognition. This suggests that relying solely on Emo-A may be insufficient for EICG task. Therefore, we additionally introduce Sem-C and Sem-D to estimate the content clarity and diversity, where our method exhibits a clear advantage over other methods.

User Study Besides qualitative and quantitative comparisons, we also conduct a user study to determine whether our method is preferred by humans and to understand how people perceive emotions. We invite 14 participants from different social backgrounds and each test session lasts about 30 minutes. In the first part, generation results are evaluated on three dimensions: image fidelity, emotion faithfulness and semantic diversity. Each question presented to the participants includes two sets of images conveying the same emotion, drawn from our method and one of the comparison methods. The participants are then asked: *which group is more realistic? which group evokes a stronger sense of [emotion type]? which group is more diverse?* As illustrated in Table 2, our method attains the top rankings compared to the other three methods, particularly excelling in semantic diversity. We aim to explore the factors influencing visual emotions in the second part.



(a) Emotional Concepts (b) Emotional Images

Figure 7. Emotion decomposition. Each emotion word is broken down into a set of (a) emotional concepts, reflecting the semantics in (b) generated images.

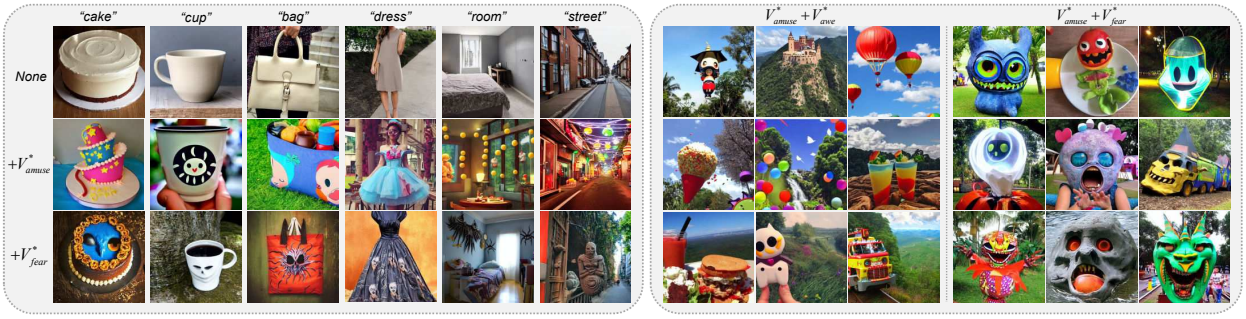
Table 3. Comparisons with the state-of-the-art methods on emotion transfer task, involving three metrics.

Method	Emo-A ↑		CLIP-img ↑		CLIP-txt ↑	
	amusement	*fear*	*amusement*	*fear*	*amusement*	*fear*
Stable Diffusion	51.54%	56.67%	**0.929**	0.825	0.257	0.251
Textual Inversion	60.82%	40.00%	0.902	0.792	0.270	0.259
Ours	**72.16%**	**63.33%**	0.913	**0.841**	**0.276**	**0.270**

Participants are shown an emotional image generated by our method and are asked: *which emotion best describes the image? why do you feel such emotion?* Compared to the 76.25% machine predicted one in Table 1, 82.14% emotion accuracy is achieved by user voting, where generated images are more emotion-evoking towards human participants. Additionally, 88.39% of the responses indicate that emotions are predominantly triggered by the content/semantic. This underscores how our task, EICG, is closely aligned with human cognition.

4.3. Ablation Study

We examine the efficacy of each network design, encompassing the non-linear mapping network F, the attribute loss $\mathcal{L}_{attr}$ and the emotion confidence α_{ij}. In Table 1, without nonlinear mapping network, emotion representations are aggregated, which fails to restore the real image set (high FID) and lacks semantic diversity (low Sem-D). Attribute loss is introduced to enhance semantic clarity and diversity, whose absence leads to performance drops in Sem-C and Sem-D. Besides, as shown in Figure 6, generated images exhibit semantic distortions when attribute loss is absent (w/o $\mathcal{L}_{attr}$) and display explicit contents with attribute loss (w/o α_{ij}). While image contents become clear and diverse with attribute loss, it is only with emotion confidence that we can effectively filter out emotion-agnostic semantics and generate images that evoke specific emotions (Ours).



(a) Emotion Transfer (b) Emotion Fusion

Figure 8. Emotion creation. (a) transfers emotion representations (*i.e.*, *amusement*, *fear*) to a series of neutral contents while (b) fuse two emotions (*i.e.*, *amusement-awe*, *amusement-fear*) together, which may be helpful for emotional art design.

4.4. Applications

Emotion Decomposition Emotions, serving as abstract concepts, pose a challenge for generative models to understand. Our method provides an opportunity to comprehend visual emotions by identifying the most relevant semantic contents for each emotion. To be specific, we visualize the semantics that are most closely aligned with our emotion representations in CLIP space. Each concept in Figure 7 (a), such as *surfboard*, *bicycle* and *athletic field*, is very likely to elicit *excitement*, where the corresponding images are presented in Figure 7 (b). Upon viewing such images, we identify the semantics and instinctively link them to specific emotions. These emotional concepts exhibit diversity, explicitness, and a strong capacity to evoke emotions. By decomposing visual emotions, we can not only generate emotional images with various semantics but also gain a deeper understanding of emotion evocation process. The results reveal the close relationship between emotions and semantics, in accordance with the psychological studies [3].

Emotion Transfer Once we identify emotional contents, the next step is to explore how we can use it to create meaningful and compelling designs. In addition to emotional content, there are also neutral ones. As shown in Figure 8 (a), we combine the common neutral objects/scenes with emotional representations learned by our method. Surprisingly, we find that these representations effectively preserve emotional semantics and seamlessly integrate them with new concepts. Taking *amusement* as an example, it preserves several semantics including *amusement park*, *picnic*, *princess*, *balloon* and *beautiful lanterns*. In Table 3, our method is quantitatively compared with the state-of-the-art methods on emotion transfer task, specializing in *room*, where our method can well-preserve semantics and effectively elicit emotions. Crucially, these creations can evoke explicit and strong emotions across various neutral semantics, suggesting the potential of our method in image editing, image transfer and emotional art design.

Emotion Fusion Additionally, we explore the possibilities of combining different emotion representations to evoke multiple emotions. In Figure 8 (b), we combine *amusement* and *awe* (positive-positive) as well as *amusement* and *fear* (positive-negative), bringing some intriguing observations. In the combination of *amusement* and *awe*, we observe elements associated with *amusement*, such as *toys*, *balloons*, and *ice-creams*, alongside awe-inspiring elements like the blue *sky*, *mountains*, *ocean*, and *city views*. Particularly, one may feel both fear and amusement when viewing the funny and horrible face. When we fuse emotions, we are essentially combining their corresponding visual contents.

5. Conclusion

Discussion In this paper, we introduce a new task named EICG and derive three specially designed metrics. We propose an emotion space and align it with the CLIP space, incorporating attribute loss and emotion confidence to ensure semantic clarity, semantic diversity and emotion fidelity. Experimental results indicate that our method surpasses the state-of-the-art text-to-image diffusion models both qualitatively and quantitatively, where user study confirms its superiority. Additionally, we outline potential applications for EICG and present some initial but promising results.

Limitations Emotions can be evoked by various visual factors such as color, style and content. In this paper, we focus on investigating the most influential factor, *i.e.*, contents. Moreover, the relationships between emotions and content is not strictly binary. In this paper, we simplify this connection by assuming content to be either emotional or emotion-agnostic. However, in reality, it is hard to assign *rose* to a single emotion category. *White rose* may evoke *sadness* while *red rose* can elicit *amusement*, making it hard to decide whether *rose* is emotional or not.

Acknowledgments: This work was supported in parts by NSFC (62302312, U21B2023), Guangdong Science and Technology Program (2023A1515011440), DEGP Innovation Team (2022KCXTD025), Shenzhen Science and Technology Program (KQTD20210811090044003, RCJC20200714114435012), Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ) and Scientific Development Funds of Shenzhen University.

References

[1] Damian Borth, Tao Chen, Rongrong Ji, and Shih-Fu Chang. Sentibank: large-scale ontology and classifiers for detecting sentiment and emotions in visual content. In *Proceedings of the 21st ACM International Conference on Multimedia*, pages 459–460, 2013. 2

[2] Damian Borth, Rongrong Ji, Tao Chen, Thomas Breuel, and Shih-Fu Chang. Large-scale visual sentiment ontology and detectors using adjective noun pairs. In *Proceedings of the 21st ACM International Conference on Multimedia*, pages 223–232, 2013. 3

[3] Tobias Brosch, Gilles Pourtois, and David Sander. The perception and categorisation of emotional stimuli: A review. *Cognition and Emotion*, 24(3):377–400, 2010. 2, 3, 8

[4] Linda Camras. Emotion: a psychoevolutionary synthesis, 1980. 2

[5] Tianlang Chen, Wei Xiong, Haitian Zheng, and Jiebo Luo. Image sentiment transfer. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 4407–4415, 2020. 3

[6] Domenico Consoli. A new concept of marketing: The emotional marketing. *BRAND. Broad Research in Accounting, Negotiation, and Distribution*, 1(1):52–59, 2010. 2

[7] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021. 2, 3

[8] Ziyi Dong, Pengxu Wei, and Liang Lin. Dreamartist: Towards controllable one-shot text-to-image generation via contrastive prompt-tuning. *arXiv preprint arXiv:2211.11337*, 2022. 3

[9] Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. Make-a-scene: Scene-based text-to-image generation with human priors. In *European Conference on Computer Vision*, pages 89–106. Springer, 2022. 3

[10] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022. 2, 3, 4, 6

[11] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2414–2423, 2016. 3

[12] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 3

[13] Alan Hanjalic. Extracting moods from pictures and sounds: Towards truly personalized tv. *IEEE Signal Processing Magazine*, 23(2):90–100, 2006. 2

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 4

[15] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in Neural Information Processing Systems*, 30, 2017. 6

[16] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 2, 3

[17] Elaine Hsieh and Brenda Nicodemus. Conceptualizing emotion in healthcare interpreting: A normative approach to interpreters' emotion work. *Patient Education and Counseling*, 98(12):1474–1481, 2015. 2

[18] Yifei Huang, Sheng Qiu, Changbo Wang, and Chenhui Li. Learning representations for high-dynamic-range image color transfer in a self-supervised way. *IEEE Transactions on Multimedia*, 23:176–188, 2020. 3

[19] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019. 3

[20] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 3

[21] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1931–1941, 2023. 3

[22] Yann LeCun, Sumit Chopra, Raia Hadsell, M Ranzato, and Fujie Huang. A tutorial on energy-based learning. *Predicting Structured Data*, 1(0), 2006. 3

[23] Joonwhoan Lee and EunJong Park. Fuzzy similarity-based emotional classification of color images. *IEEE Transactions on Multimedia*, 13(5):1031–1039, 2011. 3

[24] Wentong Liao, Kai Hu, Michael Ying Yang, and Bodo Rosenhahn. Text to image generation with semantic-spatial aware gan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18187–18196, 2022. 3

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

[26] Jana Machajdik and Allan Hanbury. Affective image classification using features inspired by psychology and art theory. In *Proceedings of the 18th ACM International Conference on Multimedia*, pages 83–92, 2010. 3

[27] Saif Mohammad and Svetlana Kiritchenko. Wikiart emotions: An annotated dataset of emotions evoked by art. In *Proceedings of the eleventh International Conference on Language Resources and Evaluation*, 2018. 3

[28] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 3

[29] Magnus Oskarsson. Robust image-to-image color transfer using optimal inlier maximization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 786–795, 2021. 3

[30] Kuan-Chuan Peng, Tsuhan Chen, Amir Sadovnik, and Andrew C Gallagher. A mixed bag of emotions: Model, predict, and transfer emotion distributions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 860–868, 2015. 2, 3

[31] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763, 2021. 2, 3, 5

[32] 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

[33] Harsh Rangwani, Lavish Bansal, Kartik Sharma, Tejan Karmali, Varun Jampani, and R Venkatesh Babu. Noisytwins: Class-consistent and diverse image generation through stylegans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5987–5996, 2023. 3, 4

[34] Tianrong Rao, Xiaoxu Li, and Min Xu. Learning multi-level deep representations for image emotion classification. *Neural Processing Letters*, pages 1–19, 2016. 2

[35] Tianrong Rao, Xiaoxu Li, and Min Xu. Learning multi-level deep representations for image emotion classification. *Neural Processing Letters*, 51:2043–2061, 2020. 3

[36] Erik Reinhard, Michael Adhikhmin, Bruce Gooch, and Peter Shirley. Color transfer between images. *IEEE Computer Graphics and Applications*, 21(5):34–41, 2001. 3

[37] Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *International Conference on Machine Learning*, pages 1530–1538, 2015. 3

[38] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 2, 3, 4, 6

[39] 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 the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22500–22510, 2023. 2, 3, 4, 6

[40] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 3

[41] Shikun Sun, Jia Jia, Haozhe Wu, Zijie Ye, and Junliang Xing. Msnet: A deep architecture using multi-sentiment semantics for sentiment-aware image style transfer. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1–5. IEEE, 2023. 2, 3

[42] Zhentao Tan, Menglei Chai, Dongdong Chen, Jing Liao, Qi Chu, Bin Liu, Gang Hua, and Nenghai Yu. Diverse semantic image synthesis via probability distribution modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7962–7971, 2021. 6

[43] Guy Tevet, Brian Gordon, Amir Hertz, Amit H Bermano, and Daniel Cohen-Or. Motionclip: Exposing human motion generation to clip space. In *European Conference on Computer Vision*, pages 358–374. Springer, 2022. 4

[44] Xiaohui Wang, Jia Jia, and Lianhong Cai. Affective image adjustment with a single word. *The Visual Computer*, 29: 1121–1133, 2013. 3

[45] Zhizhong Wang, Lei Zhao, and Wei Xing. Stylediffusion: Controllable disentangled style transfer via diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7677–7689, 2023. 3

[46] Yuxiang Wei, Yabo Zhang, Zhilong Ji, Jinfeng Bai, Lei Zhang, and Wangmeng Zuo. Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. *arXiv preprint arXiv:2302.13848*, 2023. 3

[47] 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 the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10810–10819, 2023. 2, 3

[48] Ali Yadollahi, Ameneh Gholipour Shahraki, and Osmar R Zaiane. Current state of text sentiment analysis from opinion to emotion mining. *ACM Computing Surveys (CSUR)*, 50(2): 1–33, 2017. 2

[49] Chuan-Kai Yang and Li-Kai Peng. Automatic mood-transferring between color images. *IEEE Computer Graphics and Applications*, 28(2):52–61, 2008. 3

[50] Jufeng Yang, Dongyu She, Yu-Kun Lai, Paul L Rosin, and Ming-Hsuan Yang. Weakly supervised coupled networks for visual sentiment analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7584–7592, 2018. 2, 4

[51] Jingyuan Yang, Xinbo Gao, Leida Li, Xiumei Wang, and Jinshan Ding. Solver: Scene-object interrelated visual emotion reasoning network. *IEEE Transactions on Image Processing*, 30:8686–8701, 2021. 2, 3, 4

[52] Jingyuan Yang, Jie Li, Xiumei Wang, Yuxuan Ding, and Xinbo Gao. Stimuli-aware visual emotion analysis. *IEEE Transactions on Image Processing*, 30:7432–7445, 2021. 3

[53] Jingyuan Yang, Qirui Huang, Tingting Ding, Dani Lischinski, Danny Cohen-Or, and Hui Huang. Emoset: A large-scale visual emotion dataset with rich attributes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20383–20394, 2023. 2, 4, 5

[54] Shuai Yang, Liming Jiang, Ziwei Liu, and Chen Change Loy. Pastiche master: Exemplar-based high-resolution portrait style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7693–7702, 2022. 3

[55] Chenrui Zhang and Yuxin Peng. Stacking vae and gan for context-aware text-to-image generation. In *2018 IEEE Fourth International Conference on Multimedia Big Data (BigMM)*, pages 1–5. IEEE, 2018. 3

[56] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In

Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 3836–3847, 2023. 2, 3, 4

[57] 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 the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018. 6

[58] Wei Zhang, Xuanyu He, and Weizhi Lu. Exploring discriminative representations for image emotion recognition with cnns. *IEEE Transactions on Multimedia*, 22(2):515–523, 2019. 3

[59] Minfeng Zhu, Pingbo Pan, Wei Chen, and Yi Yang. Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5802–5810, 2019. 3

[60] 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