EmoEdit: Evoking Emotions through Image Manipulation

Jingyuan Yang[1], Jiawei Feng[1], Weibin Luo[1], Dani Lischinski[2], Daniel Cohen-Or[3], Hui Huang[1]*
[1]Shenzhen University [2]The Hebrew University of Jerusalem [3]Tel Aviv University
{jingyuanyang.jyy, fengjiawei0909, waibunlok, danix3d, cohenor, hhzhiyan}@gmail.com

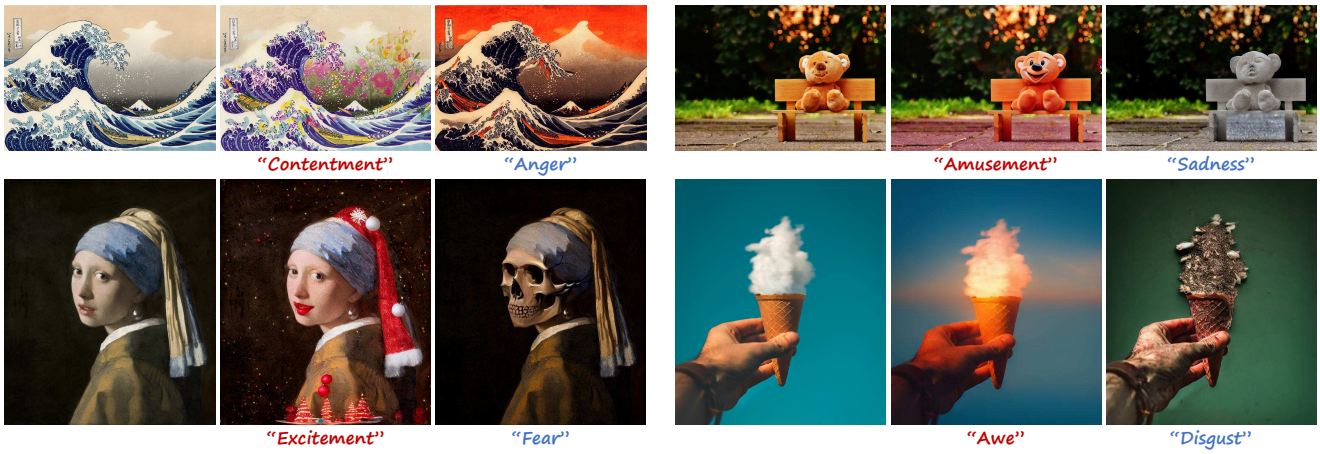


Figure 1. Affective Image Manipulation with EmoEdit, which seeks to modify a user-provided image to evoke specific emotional responses in viewers. Our method requires only emotion words as prompts, without necessitating detailed descriptions of the input or output image.

Abstract

Affective Image Manipulation (AIM) seeks to modify user-provided images to evoke specific emotions. This task is inherently complex due to its twofold objective: evoking the intended emotion while preserving image composition. Existing AIM methods primarily adjust color and style, often failing to elicit precise, profound emotional shifts. Drawing on psychological insights, we introduce EmoEdit, which extends AIM by incorporating content modifications to enhance emotional impact. Specifically, we construct EmoEditSet, a large-scale AIM dataset of 40,120 paired data through emotion attribution and data construction. To make generative models emotion-aware, we design an Emotion Adapter and train it using EmoEditSet. We further propose an instruction loss to capture semantic variations in each data pair. Our method is evaluated both qualitatively and quantitatively, demonstrating superior performance over state-of-the-art techniques. Additionally, we showcase the portability of our Emotion Adapter to other diffusion-based models, enhancing their emotion knowledge with diverse semantics. Code is available at: https://github.com/JingyuanYY/EmoEdit.

*Corresponding author

1. Introduction

"The emotion expressed by wordless simplicity is the most abundant."

–William Shakespeare

Emotions weave through our daily lives, deeply intertwined with our perception of the world. Among myriad factors, visual stimuli stands out as particularly influential in shaping emotional states. The emerging field of Visual Emotion Analysis (VEA) [29, 44, 47] explores the intricate connections between visuals and human emotions, with applications from advertising[12, 25] to robotics [9, 35]. In the arts, profound emotional responses have long been elicited via manipulation of visual elements. This prompts an intriguing question: whether and how can we deliberately alter an image to steer the viewer's emotional experience?

This paper addresses the task of Affective Image Manipulation (AIM), which adds or replaces emotional elements (*e.g.*, objects, scenes, facial expressions, actions) to evoke a specific emotion while keeping the rest of the input image unchanged., as depicted in Fig. 1. This task poses considerable challenges, requiring the automatic selection of appropriate visual elements suited for the image at hand, and integrating them in an emotionally resonant manner, while striving to maintain the original composition of the image.

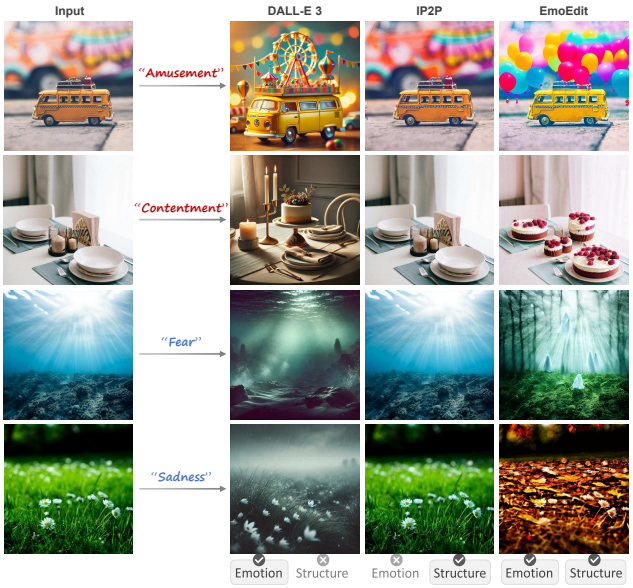


Figure 2. While DALL-E 3 conveys emotions well, IP2P remains faithful to original structure, neither approach satisfies both aspects. EmoEdit fills this gap by creating images with both emotion fidelity and structure preservation.

Current state-of-the-art generative models, while proficient in image editing, often fail to balance AIM's contradictory objectives. For instance, DALL-E 3 [1] conveys desired emotions but does not adhere to the original structure, as shown in Fig. 2. Conversely, InstructPix2Pix (IP2P)[3] preserves structure but lacks emotional expressiveness.

Previous approaches in AIM focused primarily on adjusting color and style [7, 17, 39, 41], but these methods struggle to evoke precise and significant emotion shifts, and some are limited to binary emotion categories (positive and negative). Psychological studies show that semantic content strongly influences emotional perception [4]. Inspired by this, we introduce *EmoEdit*, a novel approach that goes beyond color and style adjustments to perform substantive modifications of relevant content. Furthermore, we use the eight emotion categories introduced by Mikels *et al*. [21], to achieve more precise control over the emotions conveyed.

EmoEdit is a content-aware AIM framework capable of evoking emotions with diverse semantic modifications. Given the lack of paired data, we first generate EmoEditSet, in two stages: emotion attribution and data construction. Specifically, based on the recent large-scale EmoSet [47], we leverage a Vision-Language Model (VLM) to create eight emotion factor trees, each comprising several semantic summaries for a specific emotion. Images from several different sources [34, 50] are collected and then modified by InstructPix2Pix [3] with emotion factors to generate target candidates. Since data quality is crucial for AIM, we carefully filter target candidates using four evaluation metrics and human feedback. EmoEditSet ultimately comprises 40,120 image pairs, serving as a high-quality, semantic-diverse benchmark dataset for AIM.

To make diffusion models emotion-aware, we design the Emotion adapter, which integrates knowledge from EmoEditSet. Specifically, we leverage the structure of Q-Former [16], particularly its attention mechanisms, to facilitate interaction between the target emotion and the input image. To capture the semantic variations in each emotion data pair, we further introduce an instruction loss. EmoEdit is optimized with both instruction loss and diffusion loss, yielding images that preserve the original structure while remaining emotionally faithful and semantically diverse.

To ensure fair comparisons, we assembled an evaluation set of 405 images with a distribution distinct from the training set. Quantitative and qualitative evaluations are conducted against state-of-the-art editing techniques including: global, local, and emotion-related. Our evaluation metrics assess three key aspects: pixel-wise similarity, semantic similarity, and faithfulness to the target emotion. Finally, we demonstrate the portability of Emotion adapter to other diffusion-based models, spanning both editing and generation tasks.

In summary, our contributions are:

- EmoEdit, a content-aware AIM framework capable of generating emotion-evoking, contextually fitting, and structurally faithful variant of a user-provided input image, requiring only emotion categories as prompts.
- EmoEditSet, the first large-scale AIM dataset, featuring 40,120 image pairs labeled with emotion directions and content instructions, establishing a high-quality, semantically diverse benchmark.
- Emotion adapter, trained with diffusion loss and the proposed instruction loss, functions as a plug-and-play module to enhance generative models with emotion-awareness once trained.

2. Related work

2.1. Visual Emotion Analysis

Over the past two decades, researchers in VEA have focused on a pivotal inquiry: What evokes visual emotions? Responses have varied, ranging from low-level features like color and texture [14, 19, 30, 53] to high-level content and style [2, 30, 45, 46, 53]. Lee *et al*. [14] propose an approach to evaluate visual emotions by constructing prototypical color images for each emotion. Drawing inspiration from psychology and art theory, Machajdik *et al*. [19] extract and combine color and texture features for emotion classification. A significant milestone in this field was achieved by Borth *et al*. [2], who constructed a comprehensive visual sentiment ontology named SentiBank, where each concept is represented by an Adjective-Noun Pair (ANP). Rao *et*

al. [30] introduced MldrNet, a model that predicts emotions by combining pixel-level, aesthetic, and semantic features. To develop a comprehensive representation for emotion recognition, Zhang *et al*. [53] integrated content and style information. Yang *et al*. explored networks based on different visual stimuli [46], and further investigated the correlations between them [45]. Previous studies indicate a correlation between emotion and visual elements, but establishing a causal relationship requires image manipulation. By introducing specific visual elements into the input image and observing how the emotion varies, we can gain deeper insights into understanding emotions.

2.2. Diffusion-based Image Manipulation

Recent years have witnessed a meteoric rise in generative models, ranging from GANs [10], VAEs [13], and normalizing flows [31], to diffusion models [6, 11, 32]. Methods in diffusion-based image manipulation can be roughly grouped into global [20, 37, 42] and local [8, 22, 26, 38]. A pioneering work, SDEdit [20] leverages stochastic differential equations for guided image synthesis. PnP [37] enables semantic editing through condition-controlled guidance. To modify images in a more fine-grained manner, researchers have also explored local editing. Pix2Pix-Zero [26] allows for zero-shot editing of images without training on paired data. InsDiff [8] and InstructPix2Pix (IP2P) [3] allow users to edit specific regions of images by giving instructions. ControlNet [3] enables precise, localized image editing through spatial conditioning in diffusion models. Further, BlipDiffusion [15] integrates VLM with diffusion models for precise visual region modification. By plugging an additional module, adapter-based methods [23, 49] have successfully incorporated contextual information into diffusion models. Existing editing methods can effectively manipulate concrete concepts but face challenges when dealing with more abstract emotions. To address this, we design the Emotion adapter that makes diffusion models emotion-aware while maintaining their original structure.

2.3. Affective Image Manipulation

Most of the previous works in AIM can be grouped into color-based [5, 17, 27, 39, 43, 54] and style-based [7, 36, 41]. Yang *et al*. [43] pioneered the application of color transfer in image emotion manipulation, dividing the color spectrum into 24 distinct moods. Given an emotion word, Wang *et al*. [39] automatically adjust image color to meet a desired emotion and Liu *et al*. [17] further consider the semantic information by exploiting deep learning technique. Peng *et al*. [27] propose a method to transfer the color and texture of the target image to the source image, to modify emotions. Recent methods, such as CLVA [7] and AIF [41] reflect emotions derived from textual input by adjusting both the color and style of the original image. Most previous studies have focused on adjusting color and style to elicit specific emotions. However, psychologists have demonstrated that image content is a critical emotional stimuli [4]. In light of this, we introduce EmoEdit to evoke emotions with clear and diverse semantic modifications.

3. Method

We first construct EmoEditSet, the first large-scale AIM dataset (Fig. 3) and subsequently perform EmoEdit by training the Emotion adapter with the paired data (Fig. 4).

3.1. EmoEditSet

The absence of large-scale, high-quality datasets has largely hindered the advancement of AIM. Emotions are complex. Therefore, we seek to build EmoEditSet with diverse and representative examples for effective model training. In Fig. 3, we represent each emotion with various representative semantic summaries (emotion attribution) and collect, generate and filter emotion data pairs (data construction).

Emotion Attribution EmoSet is a recently introduced large-scale visual emotion dataset [47]. However, attribute labels within it are limited. While some important emotion elements are missing (*e.g*., firework, ghost), the single-word labels may restrict emotional expressiveness. Consequently, we conduct clustering on EmoSet to identify the common visual cues for each emotion in Fig. 3 (a). Given the significant correlation between emotions and semantics [4], we experimented with semantic embeddings generated by CLIP [28] and DINOv2 [24]. We observed that, for our purposes, CLIP captures visual semantics more effectively and eventually chose it for clustering. Several post-processing steps are implemented to eliminate clusters characterized by a low number of images, excessive pixel-wise similarity, and low emotion score.

Following clustering and filtering, we designate the remaining N clusters as *factors* for each emotion. We employ GPT-4V to assign a content summary to each factor and categorize the factors into four different types: object, scene, action, and facial expression. For instance, one of the factors shown in Fig. 3 (a) is summarized as "Books with flowers", and classified into the "Object" category. The emotion factor tree is structured hierarchically, where each of the semantic summaries at the leaf nodes can evoke the emotion at the root node. EmoSet comprises eight distinct emotions [21], for each of which a specific factor tree is constructed. These emotions include *amusement*, *awe*, *contentment*, *excitement*, *anger*, *disgust*, *fear*, and *sadness*, with the first four as *positive* and the last four as *negative*. For more details, please refer to the supplementary material.

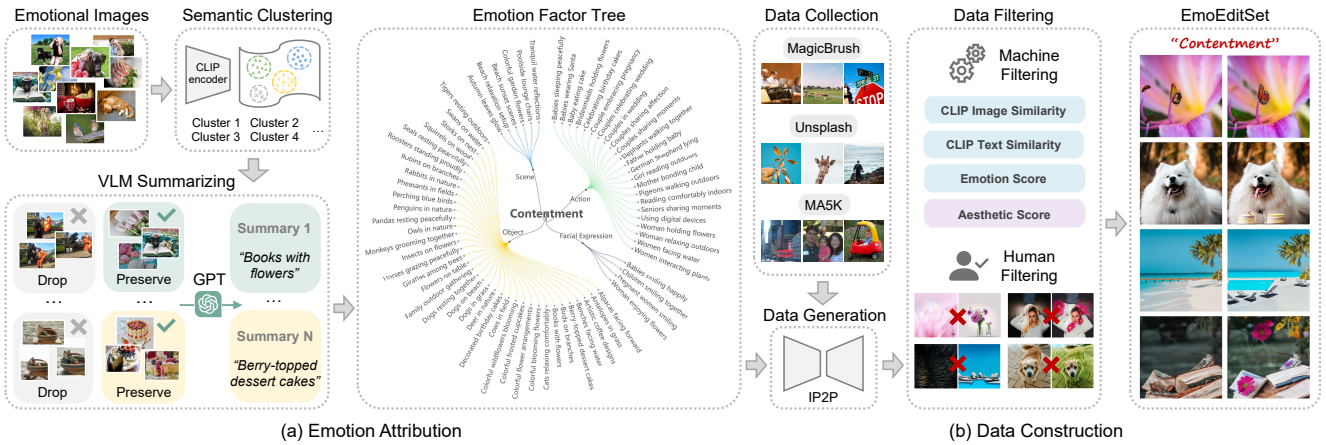


Figure 3. Overview of EmoEditSet. (a) Emotion Attribution: Emotion factor trees are built with various representative semantic summaries. (b) Data Construction: Through careful collection, generation and filtering, EmoEditSet is built with accurate and diverse paired data.

Data Construction Aiming for larger data scale and greater image diversity, we collect images from multiple sources, including MagicBrush [50], MA5K [34] and Unsplash*. While images in the first two sources are collected from social media, those from unsplash are more artistic. We utilize IP2P [3] to generate emotion data pairs, with instructions derived from the emotion factor trees. Given the knowledge gap between GPT-4V and Stable Diffusion [32], we combine factors with high semantic similarity, and eliminate those with low emotion score or high content abstractness to enhance generation results. For more details, please refer to the supplementary material. Taking contentment as an example, given an input image, IP2P receives instructions like "Add colorful butterfly", generating N images simultaneously.

We then filter the N target candidates through machine evaluation and human review. As depicted in Fig. 3 (b), we utilize CLIP image similarity [28], CLIP text similarity [28], Aesthetic score [33] and propose Emotion score as evaluation metrics. To ensure appropriate pixel-wise consistency between the source and target images, we employ CLIP image similarity with a threshold ranging from 0.75 to 0.9. Subsequently, CLIP text similarity is utilized to assess editing quality by comparing the text prompt with the target image, with a range of 0.25 to 1. Since our task aims to evoke emotions through image manipulation, predicted score after softmax for the intended category, *i.e.*, Emotion score, should exceed 0.3. We introduce Aesthetic score to select the optimal target image, considering content consistency and visual appeal. While these metrics greatly improve data quality, undesired edits still occur. As shown in Fig. 3 (b), some source-target pairs exhibit irrelevant structures, while certain target images are hard to interpret due to unreasonable or distorted content. Therefore, we incorporate manual review to further ensure the quality of the paired data.

Ultimately, EmoEditSet is constructed with a total of 40,120 image pairs, each formatted as a triplet: "source-emotion-target". There are 15,531 source images and 40,120 data pairs, indicating that most images are modified to evoke multiple emotions, with an average of 2.6 directions per image. An additional semantic instruction is provided for each image pair. In Fig. 3 (b), several semantic instructions can elicit *contentment*. Owing to the emotion factor tree, we can generate emotion-evoking targets with abundant semantic representations.

3.2. EmoEdit

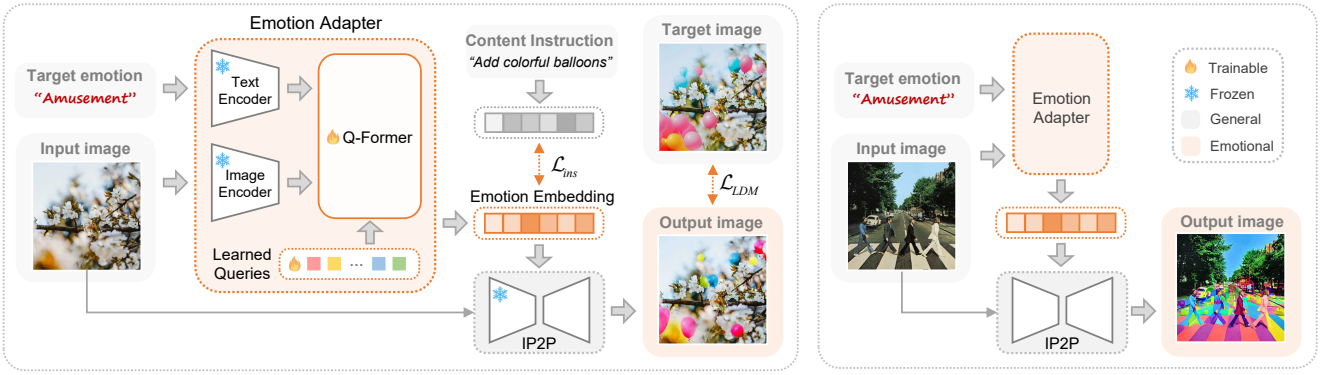
To evoke specific emotions from viewers through image manipulation, we propose EmoEdit in Fig. 4. Emotion adapter is trained using paired data from EmoEditSet, which can be directly plugged into existing diffusion-based methods to enhance their emotion knowledge.

3.2.1 Emotion Adapter

IP2P struggles to interpret and express emotions, as in Fig. 2, and we aim to embed this capability into the model. While directly fine-tuning IP2P with emotion data is a viable approach, it presents several drawbacks: (1) it is time-consuming and demands significant computational resources; (2) it lacks compatibility with methods other than IP2P. Recent studies have demonstrated that an *adapter* can be seamlessly plugged into existing diffusion models without modifying the original structure [23, 49, 51]. In this work, we investigate the feasibility of developing the Emotion adapter to make diffusion models emotion-aware.

AIM poses unique challenges compared to other editing tasks, since each emotion has various semantic representations. Consequently, automatically selecting the most ap-

*https://unsplash.com/data



(a) Training Process of EmoEdit (b) Inference Process of EmoEdit

Figure 4. Overview of EmoEdit. (a) EmoEdit trains the Emotion adapter with paired data from EmoEditSet, by optimizing instruction loss and diffusion loss. (b) Given a user-provided image, EmoEdit can modify the image to evoke the desired emotion with clear semantics.

propriate representation for an input image remains a significant issue. Q-Former [16] can leverage contextual information from one modality to enhance understanding in another. We apply this capability to build the Emotion adapter, where target emotion and the input image are encoded as e_t and e_i and then fused by self-attention and cross-attention mechanisms. Specifically, the learned queries q functions as an emotion dictionary. The self-attention mechanism first selects relevant semantics from this dictionary based on the target emotion (Eq. 1), while cross-attention identifies the most suitable representation by considering the emotional context A_s and the input image (Eq. 2):

$$A_s = softmax(\frac{[q;e_t]W_q^s([q;e_t]W_k^s)^T}{\sqrt{d_k}})[q;e_t]W_v^s, \quad (1)$$

$$A_c = softmax(\frac{A_sW_q^c(e_iW_k^c)^T}{\sqrt{d_k}})e_iW_v^c, \quad (2)$$

where W_q^s, W_k^s, W_v^s are the learned parameters in the self-attention block and W_q^c, W_k^c, W_v^c the cross-attention block, d_k is the dimension of keys. Given emotion dictionary, target emotion and input image, our Emotion adapter integrates information from these three sources iteratively and generate the most appropriate emotion embedding c_e.

3.2.2 Instruction Loss

During training, as in Fig. 4 (a), we optimize only the Emotion adapter while keeping the parameters in IP2P fixed. Diffusion loss [32] is introduced to train generative models by iteratively denoising random noise, facilitating high-quality sample generation and enhancing training stability. Thus, we apply diffusion loss during the optimization process to guide the learning of pixel-level emotional representations in the paired data:

$$\mathcal{L}_{LDM} = \mathbb{E}_{\mathcal{E}(x),c_i,c_e,\epsilon,t}\left[\|\epsilon - \epsilon_\theta(z_t, t, \mathcal{E}(c_i), c_e)\|_2^2\right], \quad (3)$$

where c_i represents the input image, $\mathcal{E}(x)$ denotes the latent encoder, ϵ refers to the added noise, ϵ_θ indicates the denoising network and z_t is the latent noise at time t.

By applying diffusion loss alone, however, EmoEdit overly emphasizes pixel-wise similarities, *i.e.*, color and texture, resulting in undesired large color blocks in Fig. 6. As in Fig. 3 (a), each emotion can be evoked by several factors. To capture the semantic variations in each emotion-data pair, we introduce an instruction loss that guides the training process with explicit instructions:

$$\mathcal{L}_{ins} = \frac{1}{N}\|c_e - \mathcal{E}_{txt}(t_{ins})\|_2^2, \quad (4)$$

where t_{ins} denotes the content instruction, $\mathcal{E}_{txt}$ indicates the text encoder and N is the normalization factor. By combining the two losses, EmoEdit is optimized for both pixel-level and semantic-level accuracy.

In the inference process, as shown in Fig. 4 (b), given a user-provided image and a target emotion, EmoEdit generates an image that is emotionally faithful, semantically clear, and structurally preserved. Notably, our Emotion adapter can be seamlessly plugged into other diffusion-based generative models beyond IP2P, as further validated in Fig. 8.

4. Experiments

4.1. Dataset and Evaluation

Dataset For fair comparisons, we compiled an inference set of 405 images distinct from the training set, sourced from user uploads available online*. Each image is targeted with 8 emotion directions, resulting in a total of 3,240 image pairs. To capture a broader spectrum of real data, our data comprises positive, negative and neutral images, categorized by a pre-trained emotion classifier on EmoSet [47].

*https://www.pexels.com/

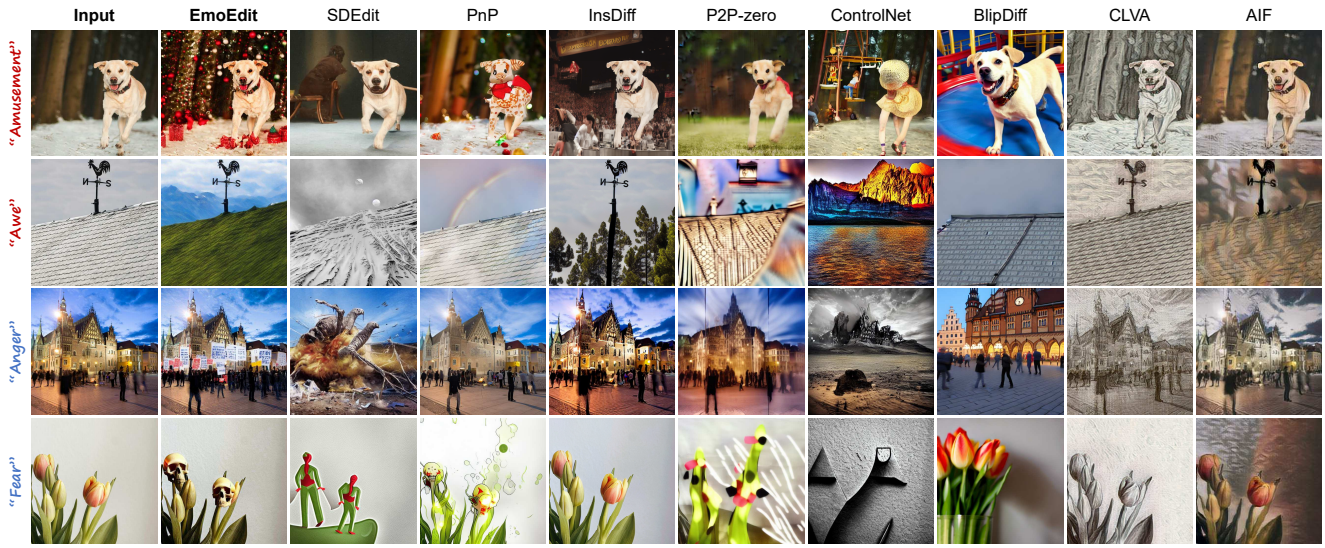


Figure 5. Comparison with the state-of-the-art methods, where EmoEdit surpasses others on emotion fidelity and structure integrity.

Table 1. Comparisons with the state-of-the-art methods on global editing, local editing and style-based AIM methods.

Method	PSNR ↑	SSIM ↑	LPIPS ↓	CLIP-I ↑	Emo-A ↑	Emo-S ↑
SDEdit [20]	15.43	0.415	0.459	0.638	38.21%	0.221
PnP [37]	14.41	0.436	0.381	**0.851**	23.83%	0.095
InsDiff [8]	10.75	0.318	0.505	0.796	19.22%	0.060
P2P-Zero [26]	13.76	0.420	0.546	0.685	20.31%	0.067
ControlNet [51]	11.98	0.292	0.603	0.686	36.33%	0.213
BlipDiff[15]	9.00	0.249	0.654	0.810	18.06%	0.045
CLVA [7]	12.61	0.397	0.479	0.757	14.04%	0.017
AIF [41]	14.05	0.537	0.493	0.828	12.74%	0.004
EmoEdit	**16.62**	**0.571**	**0.289**	0.828	**50.09%**	**0.335**

Evaluation Metrics Given the multiple objectives of AIM, we assess our method based on three aspects: pixel-level (PSNR, SSIM), semantic-level (LPIPS, CLIP-I) and emotion-level (Emo-A, Emo-S). PSNR measures the reconstruction quality by comparing the pixel-by-pixel similarity between the original and edited images, indicating the level of noise and distortion introduced during the manipulation process. SSIM [40] evaluates the structural similarity of images by comparing low-level structural information, luminance, and contrast. LPIPS [52] leverages deep learning models to assess the perceptual similarity between images. CLIP image similarity (CLIP-I) [28] measures the semantic consistency between the original and edited images, ensuring the edits align with human perception and maintain contextual relevance. Emotion accuracy (Emo-A) [48] assesses how well the edited image matches the targeted emotion, utilizing a pre-trained emotion classifier [47]. AIM is a more challenging task than emotion generation because it requires both preserving structure and evoking emotions. Therefore, we also introduce a new metric to evaluate the increase in predicted scores for the desired emotion type, called the Emotion Incremental Score (Emo-S). For more details, please refer to the supplementary material.

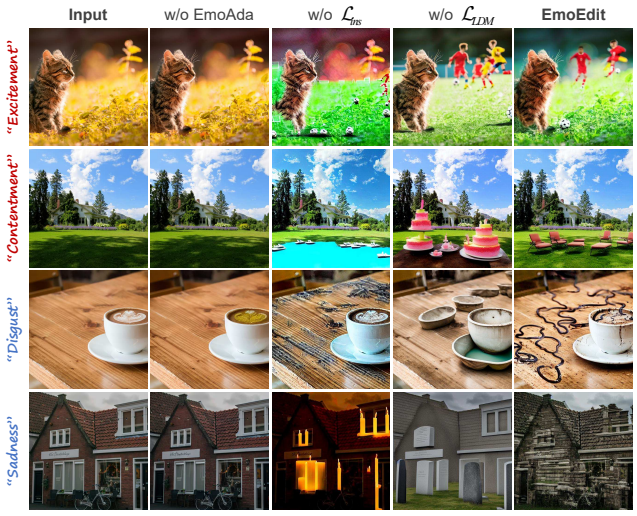


Figure 6. Ablation study on methodology. Emotion adapter, instruction loss and diffusion loss, are demonstrated to be vital.

4.2. Comparisons

Since EmoEdit is the first attempt at content-aware editing in AIM, we compare our method with the relevant state-of-the-art techniques. These include global editing methods: SDEdit [20], PnP [37], local editing methods: InsDiff [8], P2P-zero [26], ControlNet [51], BlipDiff [15] and style-based AIM methods: CLVA [7], AIF [41].

Qualitative Comparisons We present the qualitative results in Fig. 5. EmoEdit excels in both preserving structures and evoking emotions compared to other methods. Most

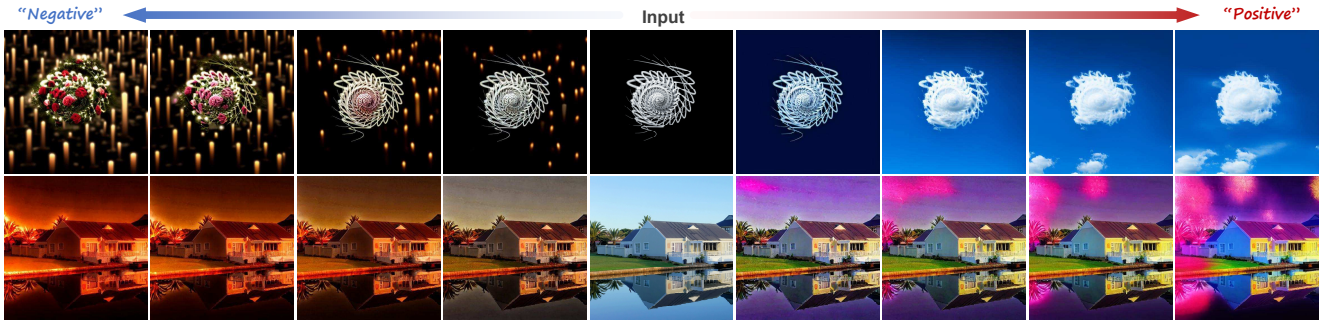


Figure 7. Ablation study on image guidance scale. EmoEdit can progressively edit an input image to different emotion polarities.

Table 2. User preference study. The numbers indicate the percentage of participants who vote for the result.

Method	Structure integrity ↑	Emotion fidelity ↑	Balance ↑
SDEdit [20]	11.71±8.91%	10.85±7.50%	5.07±6.08%
P2P-zero [26]	3.05±3.60%	5.06±5.93%	0.94±3.05%
BlipDiff [15]	15.12±16.13%	8.35±5.89%	4.88±10.45%
EmoEdit	**70.12±23.41%**	**75.73±16.44%**	**89.12±14.56%**

compared methods lack sufficient emotion knowledge, resulting in image distortion and severe artifacts. Methods like SDEdit and ControlNet possess a certain level of emotional understanding, attempting to show people with *awe* or *anger*. However, they struggle to choose contextually appropriate content for the given image. Both CLVA and AIF are emotion style transfer methods that transform realistic images into artistic styles, but the modifications on different emotions are hard to distinguish.

Quantitative Comparisons In Table 1, EmoEdit outperforms other methods across various metrics. Due to a lack of emotion knowledge, most methods perform poorly on emotion-level metrics, with only 38.21% versus 50.09% in Emo-A, and 0.221 versus 0.335 in Emo-S. One of the biggest challenges in AIM is balancing emotion fidelity with structure integrity. However, aside from emotion-level metrics, EmoEdit also achieves the best results in pixel-level metrics like PSNR and SSIM. For CLIP-I, EmoEdit ranks second, slightly behind PnP, likely because content changes can decrease semantic similarity. Among all compared methods, global editing techniques SDEdit and PnP perform relatively well, possibly because emotion effects are more global rather than local. CLVA and AIF aim to evoke emotions through artistic styles, which effectively preserves image structure but are restricted to elicit emotions in realistic images.

User Study We conducted a user study to assess whether humans prefer our method. We invited 41 participants of various ages, with each session lasting about 15 minutes. The study included 40 sets of images, each featuring an original image alongside four edited versions from different methods: SDEdit, P2P-zero, BlipDiffusion, and EmoEdit. Participants were shown a set of images and asked three questions: (1) Which image best preserves the structure? (2) Which image most strongly evokes the targeted emotion? (3) Which image achieves the best balance between structure and emotion? Participants could choose one out of the four options and we calculate the vote percentage for each question. Results in Table 2 show that EmoEdit is the most preferred choice in all questions. Despite the challenge of maintaining structure while conveying emotions, EmoEdit received the highest votes for both aspects, *i.e.* 70.12% and 75.73%. In terms of balance, EmoEdit shows a clear advantage with 89.12% support, confirming that our method aligns with human perception.

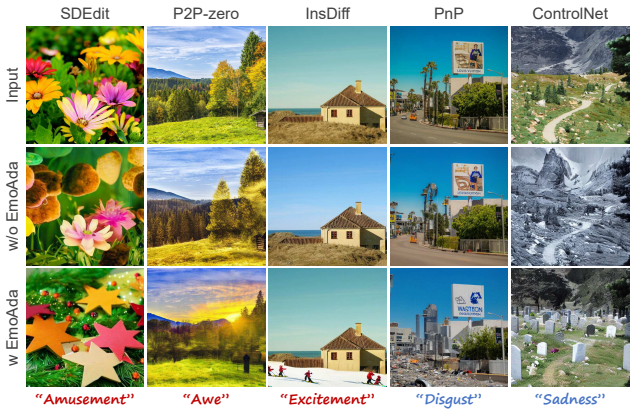


Figure 8. Emotion adapter can be effectively plugged into existing editing models to enrich their emotional knowledge.

4.3. Ablation Study

Methodology In Fig. 6, we evaluate the effectiveness of several key designs in EmoEdit, including Emotion adapter, instruction loss and diffusion loss. Without the Emotion adapter (w/o EmoAda), images remain nearly identical, highlighting its significance. Both diffusion loss and instruction loss are crucial for evoking emotions. While diffusion loss well preserves the original structure, instruction

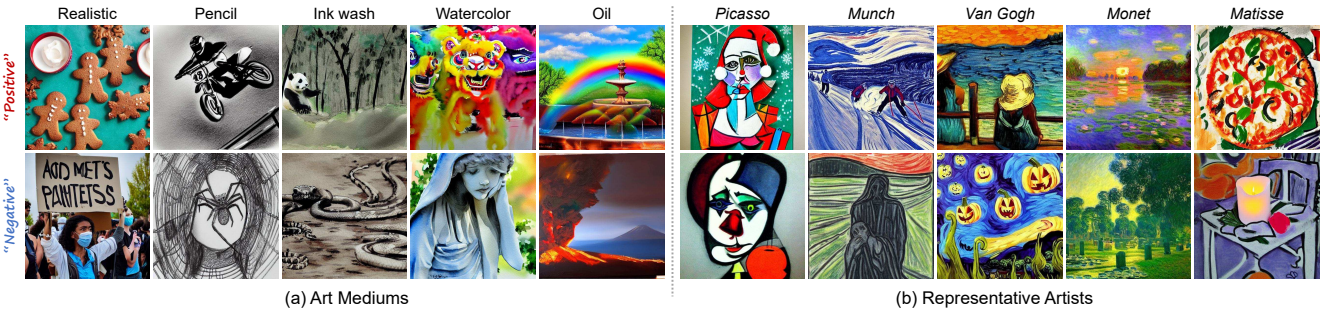


Figure 9. Emotion adapter can be extended to stylized image generation, preserving style and evoking emotions with clear semantics.

loss enhances semantic clarity. For instance, in the case of *contentment*, using diffusion loss adds water to the grass without clear semantics, while using instruction loss adds cakes to the grass, without structural preservation. EmoEdit adds lounge chairs, showcasing structure integrity, semantic clarity and contextual fitting.

Guidance Scale We show EmoEdit's editing results with variations in image guidance scale in Fig. 7. The middle image represents the input, while the left and right images illustrate different guidance scales for two emotion polarities: positive and negative. We observe that as the image guidance scale decreases, emotion intensity increases while structure preservation diminishes. Although evoking emotion and preserving structure are often contradictory, our method effectively balances them, as demonstrated in Table 1 and Table 2. Users can customize the level of manipulation by adjusting guidance scale to suit their preferences.

4.4. Applications

Once EmoEdit is trained, Emotion adapter can be directly plugged into various diffusion-based models to enhance their emotion awareness with diverse semantics, covering both image editing task and image generation task.

Emotion-enhanced Editing Models As most existing editing models in Fig. 5 lack emotion knowledge, we experiment to investigate whether Emotion adapter can augment their emotion capabilities in a plug-and-play manner. In Fig. 8, when Emotion adapter is plugged to the original method (w EmoAda), these models are capable of generating images with emotion fidelity and contextual fit. For example, ControlNet transforms the input image into black-and-white, while additional elements, such as thumb stones, are introduced after attaching the Emotion adapter.

Emotion-aware Stylized Image Generation Apart from editing task, Emotion adapter can also be extended to stylized image generation task in Fig. 9. Once trained, Emotion adapter functions as an emotional interpreter, encoding each emotion polarity (*i.e.*, positive, negative) into distinct semantic representations. We combine the Emotion adapter with Composable Diffusion [18], using 5 art mediums and 5 representative artists as style prompts. Results show that each generated images preserve art styles while evoke specified emotions with clear semantics. Taking *Monet* as an example, the sunset evokes *awe (positive)*, while the graveyard makes people feel *sadness (negative)*.

5. Conclusion

Discussion We present EmoEdit, a method designed to evoke emotions by modifying the content of user-provided images. EmoEditSet is first constructed with 40,120 image pairs, serving as a data foundation for AIM. Emotion adapter is proposed and trained using both diffusion loss and the designed instruction loss, which can be directly plugged into diffusion-based generative models to enhance their emotion-awareness. Our method is evaluated both qualitatively and quantitatively, demonstrating a strong balance between structure preservation and emotion faithfulness. User study and ablation study validate the effectiveness of EmoEdit, while applications further highlight the portability of Emotion adapter and the rich emotion knowledge embedded within it.

Limitations In the real world, there are numerous visual emotion factors beyond those in the constructed emotion factor trees. Besides, due to the inherent complexity of emotions, more than eight emotion categories exist. Our method is heavily dependent on EmoSet, suggesting that the limited data may introduce potential biases and constraints. Expanding the range of emotion factors and categories would improve the quality and diversity of the editing results. Since AIM is a highly human-centered task, future work should place greater emphasis on human interaction. Evaluation metrics, such as emotion accuracy, are largely dependent on the training data and are prone to bias, requiring more human feedback. Additionally, increased flexibility would allow users to tailor the editing process to suit their personal tastes and preferences.

References

[1] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, Wesam Manassra, Prafulla Dhariwal, Casey Chu, Yunxin Jiao, and Aditya Ramesh. Improving image generation with better captions. Technical report, OpenAI, 2023. 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 *ACM International Conference on Multimedia*, pages 223–232, 2013. 2

[3] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Conference on Computer Vision and Pattern Recognition*, pages 18392–18402, 2023. 2, 3, 4

[4] 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

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

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

[7] Tsu-Jui Fu, Xin Eric Wang, and William Yang Wang. Language-driven artistic style transfer. In *European Conference on Computer Vision*, pages 717–734, 2022. 2, 3, 6

[8] Zigang Geng, Binxin Yang, Tiankai Hang, Chen Li, Shuyang Gu, Ting Zhang, Jianmin Bao, Zheng Zhang, Han Hu, Dong Chen, and Baining Guo. Instructdiffusion: A generalist modeling interface for vision tasks. *arXiv preprint arXiv:2309.03895*, 2023. 3, 6

[9] Riccardo Gervasi, Federico Barravecchia, Luca Mastrogiacomo, and Fiorenzo Franceschini. Applications of affective computing in human-robot interaction: State-of-art and challenges for manufacturing. *Proceedings of the Institution of Mechanical Engineers, Part B: Journal of Engineering Manufacture*, 237(6-7):815–832, 2023. 1

[10] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in Neural Information Processing Systems*, 27, 2014. 3

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

[12] Jin-Ae Kang, Sookyeong Hong, and Glenn T Hubbard. The role of storytelling in advertising: Consumer emotion, narrative engagement level, and word-of-mouth intention. *Journal of Consumer Behaviour*, 19(1):47–56, 2020. 1

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

[14] Joonwhoan Lee and EunJong Park. Fuzzy similarity-based emotional classification of color images. *IEEE Transactions on Multimedia*, 13(5):1031–1039, 2011. 2

[15] Dongxu Li, Junnan Li, and Steven Hoi. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. *Advances in Neural Information Processing Systems*, 36, 2024. 3, 6, 7

[16] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*, pages 19730–19742. PMLR, 2023. 2, 5

[17] Da Liu, Yaxi Jiang, Min Pei, and Shiguang Liu. Emotional image color transfer via deep learning. *Pattern Recognition Letters*, 110:16–22, 2018. 2, 3

[18] Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B Tenenbaum. Compositional visual generation with composable diffusion models. In *European Conference on Computer Vision*, pages 423–439. Springer, 2022. 8

[19] Jana Machajdik and Allan Hanbury. Affective image classification using features inspired by psychology and art theory. In *ACM International Conference on Multimedia*, pages 83–92, 2010. 2

[20] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021. 3, 6, 7

[21] Joseph A Mikels, Barbara L Fredrickson, Gregory R Larkin, Casey M Lindberg, Sam J Maglio, and Patricia A Reuter-Lorenz. Emotional category data on images from the international affective picture system. *Behavior Research Methods*, 37(4):626–630, 2005. 2, 3

[22] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *Conference on Computer Vision and Pattern Recognition*, pages 6038–6047, 2023. 3

[23] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Association for the Advance of Artificial Intelligence*, pages 4296–4304, 2024. 3, 4

[24] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 3

[25] F Javier Otamendi and Dolores Lucia Sutil Martín. The emotional effectiveness of advertisement. *Frontiers in Psychology*, 11:2088, 2020. 1

[26] Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. Zero-shot image-to-image translation. In *Special Interest Group on Computer Graphics and Interactive Techniques*, pages 1–11, 2023. 3, 6, 7

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

[28] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763, 2021. 3, 4, 6

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

[30] Tianrong Rao, Xiaoxu Li, and Min Xu. Learning multi-level deep representations for image emotion classification. *Neural Processing Letters*, 51:2043–2061, 2020. 2, 3

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

[32] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 3, 4, 5

[33] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. 4

[34] Jing Shi, Ning Xu, Yihang Xu, Trung Bui, Franck Dernoncourt, and Chenliang Xu. Learning by planning: Language-guided global image editing. In *Conference on Computer Vision and Pattern Recognition*, pages 13590–13599, 2021. 2, 4

[35] Micol Spitale and Hatice Gunes. Affective robotics for well-being: A scoping review. In *2022 10th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, pages 1–8. IEEE, 2022. 1

[36] 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 *International Conference on Acoustics, Speech, and Signal Processing*, pages 1–5, 2023. 3

[37] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *Conference on Computer Vision and Pattern Recognition*, pages 1921–1930, 2023. 3, 6

[38] Bram Wallace, Akash Gokul, and Nikhil Naik. Edict: Exact diffusion inversion via coupled transformations. In *Conference on Computer Vision and Pattern Recognition*, pages 22532–22541, 2023. 3

[39] Xiaohui Wang, Jia Jia, and Lianhong Cai. Affective image adjustment with a single word. *The Visual Computer*, 29: 1121–1133, 2013. 2, 3

[40] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. 6

[41] Shuchen Weng, Peixuan Zhang, Zheng Chang, Xinlong Wang, Si Li, and Boxin Shi. Affective image filter: Reflecting emotions from text to images. In *International Conference on Computer Vision*, pages 10810–10819, 2023. 2, 3, 6

[42] Chen Henry Wu and Fernando De la Torre. Unifying diffusion models' latent space, with applications to cyclediffusion and guidance. *arXiv preprint arXiv:2210.05559*, 2022. 3

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

[44] Jufeng Yang, Dongyu She, Yu-Kun Lai, Paul L Rosin, and Ming-Hsuan Yang. Weakly supervised coupled networks for visual sentiment analysis. In *Conference on Computer Vision and Pattern Recognition*, pages 7584–7592, 2018. 1

[45] 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

[46] 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. 2, 3

[47] 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 *International Conference on Computer Vision*, pages 20383–20394, 2023. 1, 2, 3, 5, 6

[48] Jingyuan Yang, Jiawei Feng, and Hui Huang. Emogen: Emotional image content generation with text-to-image diffusion models. *arXiv preprint arXiv:2401.04608*, 2024. 6

[49] 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, 4

[50] Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. Magicbrush: A manually annotated dataset for instruction-guided image editing. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 4

[51] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *International Conference on Computer Vision*, pages 3836–3847, 2023. 4, 6

[52] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018. 6

[53] 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. 2, 3

[54] 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