MUSE: Manipulating Unified Framework for Synthesizing Emotions in Images via Test-Time Optimization

Yingjie Xia[1] Xi Wang[2] Jinglei Shi[1,3] Vicky Kalogeiton[2] Jian Yang[1]
[1] VCIP & TMCC & DISSec, College of Computer Science, Nankai University
[2] LIX, Ecole Polytechnique, IP Paris [3] Key Lab of SCCI, Dalian University of Technology

***Abstract*—Images evoke emotions that profoundly influence perception, often prioritized over content. Current Image Emotional Synthesis (IES) approaches artificially separate generation and editing tasks, creating inefficiencies and limiting applications where these tasks naturally intertwine, such as therapeutic interventions or storytelling. In this work, we introduce MUSE, the first unified framework capable of both emotional generation and editing. By adopting a strategy conceptually aligned with Test-Time Scaling (TTS) that widely used in both LLM and diffusion model communities, it avoids the requirement for additional updating diffusion model and specialized emotional synthesis datasets. More specifically, MUSE addresses three key questions in emotional synthesis: (1) *HOW* to stably guide synthesis by leveraging an off-the-shelf emotion classifier with gradient-based optimization of emotional tokens; (2) *WHEN* to introduce emotional guidance by identifying the optimal timing using semantic similarity as a supervisory signal; and (3) *WHICH* emotion to guide synthesis through a multi-emotion loss that reduces interference from inherent and similar emotions. Experimental results show that MUSE performs favorably against all methods for both generation and editing, improving emotional accuracy and semantic diversity while maintaining an optimal balance between desired content, adherence to text prompts, and realistic emotional expression. It establishes a new paradigm for emotion synthesis.**

***Index Terms*—emotional synthesis, diffusion models, token optimization, emotion guidance, test-time optimization.**

I. Introduction

"Conscience is the voice of the soul, the passions are the voice of the body." -Jean-Jacques Rousseau (FR.)

MEDIA, particularly images, have profoundly transformed the way people perceive the world in dissemination. People sometimes prioritize the emotion conveyed in images over content, a phenomenon so impactful that gives rise to the "post-truth" concept in political science. This emotional impact has led to fast progress in Image Emotional Synthesis (IES) across various fields [1], [2].

Current IES work typically splits into two streams: (1) *generation*, which creates novel imagery with a desired emotional profile, and (2) *editing*, which retouches existing images to alter their affect. Maintaining *separate* pipelines for these tasks inflates training cost, complicates deployment, and most critically, introduces inconsistency between the generation and refinement. A unified framework would collapse that divide, enabling continuous cycles of generation and adjustment that are vital for therapeutic art tools, adaptive storytelling engines, and emotion-aware communication platforms where creation and modification blur in practice.

Early IES methods conveyed emotions by editing low-level features, such as linking colors to affective words [3] or applying color constraints via emotion classifiers [4]. Later approaches explored high-level styles, using semantic spaces [5] or text-driven color guidance [6]. Recent IES approaches leverage diffusion [7], [8] following its success in various content generation. EmoGen [9] pioneered emotional diffusion generation by mapping emotion and semantic spaces, translating abstract emotions into concrete attributes for the generation task. EmotiCrafter [10] extends emotional generation to the continuous valence-arousal space by embedding emotions into text features. Emotional editing has also been advanced: EmoEdit [11] first employs GPT4V to extract emotion factor trees that identify key emotional attributes. Recently, EmoAgent [12] employs large language models (LLMs) to generate editing instructions and calls specific editing tools (IP2P [13]) for affective image manipulation.

Despite these advances, SOTA IES methods continue to face several key limitations: (1) The color- and style-based approaches [5], [6], [14], [15] often fail to accurately evoke specific emotions due to the strong inherent image emotion. (2) While the diffusion-based methods [16] often rely on specialized datasets, and may significantly change the content or even produce a totally different image. (3) Methods that implicitly [9] or explicitly [11] associate emotions with certain attributes will make some elements repeatedly appear in the synthesized images, reducing their semantic diversity. Finally, (4) existing approaches address either emotional generation [9], [10], [16] or editing [11], [15], creating inconsistent emotional representations and fragmenting user experience.

To address these issues, we propose MUSE, a unified framework for both emotional generation and editing (see Fig.1). Our idea is similar to the test-time scaling philosophy [17], we search for a better noise via reverse optimization of tokens during inference to achieve more precise and harmonious emotion control. It focuses on *how, when*, and *which* emotions should guide the synthesis: (1). **How:** We leverage prior knowledge from an off-the-shelf emotion classifier to guide the synthesis process. Gradients are back-propagated to progressively optimize predefined emotional tokens at the input. As a result, a longer denoising process is employed during inference, but leading to improved synthesis quality. (2). **When:** While scaling during the inference stage is beneficial, introducing emotional guidance too early or too late may disrupt image content or fail to convey the target emotion. To address this, we use semantic similarity as a supervisory signal to determine the optimal timing for applying guidance, striking

a balance between effective control and efficient generation. (3). **Which:** We propose a multi-emotion cross-entropy loss to reduce interference from both the inherent and similar emotions, improving the emotional synthesis accuracy and alleviating conflicts between the target and inherent emotions.

In summary, MUSE optimizes emotional tokens by minimizing a cross-entropy emotional loss, then introduces these tokens at the optimal timing to guide image emotional synthesis. Extensive experiments and analysis show that MUSE strikes a notable balance among desired content, textual adherence and emotional realism. It achieves superior emotion accuracy and human preference across both tasks against those SOTA emotion generation methods such as UG [18], SD3 [19], FLUX [20], EmoGen [9], and emotion editing ones like SDEdit (SDE) [21], AIF [6], IDE [22], PnP [23], Forgedit [24], and EmoEdit [11], *without* relying on specialized datasets or additional updates of the diffusion model. The contributions of our work can be summarized as follows:

- We propose the first unified framework for both emotional image generation and editing, which leverages a test-time optimization strategy to achieve more accurate and natural emotion manipulation without creating specialized dataset and model finetuning.
- We investigate the three key aspects of how, when, and which emotion to guide the synthesis by continuously optimizing emotional tokens. CLIP-based semantic similarity is employed to ensure alignment between generated images and text prompts, while a multi-emotion loss function guides the accurate expression of the target emotions. These factors together ensure a stable synthesis.
- We carried out comprehensive experiments on both emotional editing and generation tasks to demonstrate that MUSE outperforms recent State-Of-The-Art (SOTA) methods such as EmoGen [9] and EmoEdit [11] in terms of emotional accuracy (Emo-A/B/C), and visual realism (FID [25]), while also ensuring high semantic similarity (CLIP) and human preference (HPSv2 [26]). In addition, we reported superior performance in corresponding human evaluation studies against compared methods.

II. Related work

A. Diffusion Models for Image Synthesis

Diffusion models for image synthesis gradually add random noise to data and learn a model to reverse this process to recover data from the original distribution [27], [28]. Methods either operate in pixel space, e.g. Imagen [29] or in latent space for improved efficiency at higher resolutions such as DALLE-2 [30], Stable Diffusion (SD) [7]. To condition the generation, most works apply guidance [31], [32] during denoising. The authors in [31] introduce classifier-free guidance to guide the generation directly without an explicit pre-trained classifier. To further enhance controllability, recent works explore various plug-and-play modules. For example, IP-Adapter [33] injects features from reference images via cross-attention operation Moreover, structure-aware controllers such as ControlNet [34] and T2I-Adapter [35] incorporate spatial conditions like edges or depth maps to guide image layout. More recently, [17] proposed a test-time scaling framework that improves generation quality by searching for better noise inputs during inference, complementing the traditional denoising step scaling.

B. Visual Emotion Analysis

Visual Emotion Analysis (VEA) traditionally focused on designing handcrafted features for analyzing emotions in images [36]. Modern works [37]–[39] rely on DL for emotional recognition. Later works [40], [41] leveraged LLMs. They either proposed prompt-based fine-tuning to improve classification [40], or proposed language-supervised fusion of language and emotion features to boost the emotion capability of visual models [41]. Emotion analysis has also been explored in videos [42], [43] and online chatting [44], [45]. [42] proposed a Masked AutoEncoder-based video affective analysis method, employing lexicon verification and emotion-oriented masking for robust temporal emotion representation learning. For conversational emotion recognition, [46] introduced a transformer-based model with self-distillation to enhance multimodal interactions and representations. In addition to recognition methods, [47] introduced two psychological metrics (ECC and EMC) to quantify the severity of emotional misclassification, emphasizing emotion-aware evaluation.

C. Image Emotional Synthesis

Image Emotional Synthesis (IES) generates media aligned with emotions through image generation and editing. Editing methods usually rely on color [15], style [6] and diffusion [11], [16]. The first attempt [14] associates color moods with images via histograms, while [15] disentangles neutral high-level concepts from emotional low-level features via adversarial training. Style-based methods alter the overall image style, as in [48], which proposes a language-driven method transferring emotions based on style images and text instructions. Affective Image Filter (AIF) [6] employs a multi-modal transformer to translate abstract emotional cues from text into concrete visual elements such as color and texture. Diffusion models edit image emotions by altering the semantics of their content. [11] introduces emotion factor trees to identify attributes requiring modification by Vision-Language Models (VLMs) for emotion adjustment. Similarly, [9] pioneers emotional generation by linking emotions to specific VLM-focused attributes, but both works suggest that emotions are tied to fixed elements, thereby limiting semantic diversity. The work of Make Me Happier (MMH) [16] proposes a diffusion-based editing framework that preserves image semantics while modifying emotions, though it relies on simple emotion encodings that may limit expressive richness. EmotiCrafter [10] embeds Valence-Arousal into text features but struggles with arousal control and suffers from human-centric bias. EmoAgent [12] introduces a multi-agent framework that plans, edits, and critiques emotional attributes in a structured manner, but increases system complexity and relies on predefined emotional attributes. Despite these advances, a flexible framework handling both tasks with high emotional accuracy and quality remains absent, a gap our work addresses.

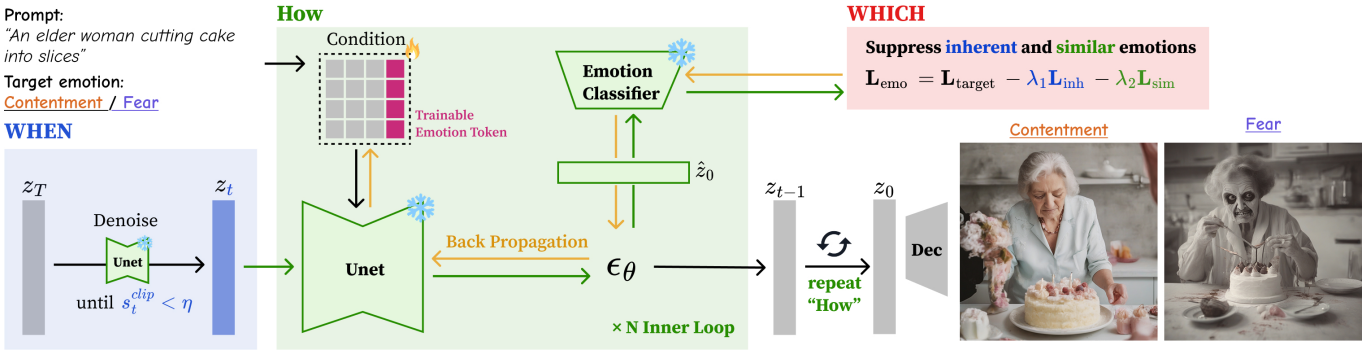


Fig. 1: Our unified framework, MUSE, addresses the *how*, *when*, and *which* questions in emotional image synthesis. It starts from a noise latent z_T that is either sampled from a Gaussian distribution (for generation) or obtained via inversion from an existing image (for editing). A CLIP model is employed to compute semantic similarity s_t^{clip} between the synthesized image and the text prompt for determining *when* to introduce the emotional guidance. To decide *how* to manipulate the emotion, we introduce a learnable emotion token in addition to the textual prompt tokens, during the inference stage. These tokens are optimized in an inner loop using an off-the-shelf emotional classifier that operates on the predicted denoised latent $\hat{z}_0$. Furthermore, we propose an emotional loss $\mathcal{L}_{\text{emo}}$ to ensure *which* emotions to enhance, which comprises the terms $\mathcal{L}_{\text{target}}$ for synthesizing the target emotion, and $\mathcal{L}_{\text{inh}}$, $\mathcal{L}_{\text{sim}}$ for suppressing inherent and similar emotions respectively.

III. Preliminary Knowledge

Diffusion for image synthesis involves reverting random noise to an image that follows the data distribution. It involves a forward addition of noise and a backward denoising process.

In the forward process, an image I_0 is initially processed through a VAE-like encoder to obtain a latent embedding $z_0 = \mathcal{E}(I_0)$. This embedding is then progressively destroyed by adding random noise over T steps:

$$z_t = \sqrt{\alpha_t} z_0 + \sqrt{1-\alpha_t}\epsilon, \quad \epsilon \sim \mathcal{N}(0, I) \quad , \tag{1}$$

where $t \in [1, T]$ represents the time step, and $\{\alpha_l\}_{l=1}^{T}$ is a scheduler that gradually decreases, controlling the extent of noise mixed with the data. In the backwards process, a network ϵ_θ is trained to reverse this diffusion process by predicting the added noise at each time step t:

$$\epsilon_\theta\left(z_t, t, c\right) \approx \frac{z_t - \sqrt{\alpha_t} z_0}{\sqrt{1-\alpha_t}} \quad , \tag{2}$$

where c is the text embedding derived from the text prompt P via the text encoder: $c = \mathcal{T}(P)$. Then the denoised embedding at time step $t-1$ can then be computed from the embedding at time step t as follows:

$$z_{t-1} = \frac{z_t - \left(\sqrt{1-\alpha_t}\right)\epsilon_\theta\left(z_t, t, c\right)}{\sqrt{\alpha_t}} \quad . \tag{3}$$

The network parameters θ are optimized by minimizing the Mean Squared Error (MSE) between the predicted noise and the ground truth [28]:

$$\theta^{'} = \arg\min_{\theta} \mathbb{E}_{z_0, t, \epsilon\sim\mathcal{N}(0,1)}\left[\left\|\epsilon_\theta\left(z_t, t, c\right) - \epsilon\right\|^2\right] \quad . \tag{4}$$

Once trained, the initial embedding distribution can be retrieved from a noise sample $z_T \sim q(z_T)$ in an iterative manner by Eq. 3. To apply condition information c into the generation, [32] propose *Classifier Guidance (CG)* that uses a pretrained classifier $p(c|x_t)$ to disturb the inference noise ϵ_θ, with a scalar $\omega{>}0$ controlling the guidance level of the condition c:

$$\hat{\epsilon}_\theta(x_t, c) = \epsilon_\theta(x_t, c) + (\omega+1)\nabla_{x_t}\log p(c|x_t) \quad . \tag{5}$$

To support more complex conditions, [31] proposes to replace the classifier with an implicit classifier: by dropping the condition during training, they employ a single network for both $\nabla_{x_t}\log p(x_t, c)$ and $\nabla_{x_t}\log p(x_t)$. This gives the *Classifier-Free Guidance (CFG)*, also controlled by ω:

$$\hat{\epsilon}_\theta(x_t, c) = \epsilon_\theta(x_t, c) + \omega\left(\epsilon_\theta(x_t, c) - \epsilon_\theta(x_t, \emptyset)\right) \quad . \tag{6}$$

IV. Methodology

We propose MUSE, a novel method for emotion-oriented image synthesis. As the overview shown in Fig. 1, it determines ***How***, ***When***, and ***Which*** emotions to incorporate in image synthesis. More specifically, it takes a text prompt or source image, along with an optimizable emotional tokens as inputs for either generation task or editing task. And a target emotion is set to be used in an off-the-shelf emotion classifier (Sec. IV-A). The optimization for emotional tokens begins when the semantic similarity of the synthesized image surpasses a preset threshold (Sec. IV-B), guided by a multi-emotion cross-entropy loss to ensure accurate emotion synthesis (Sec. IV-C).

A. Image Emotional Synthesis Framework

How to synthesize emotions during the denoising process of a diffusion model is challenging, especially with classifier-free guidance controlling textual adherence simultaneously. We propose using an emotion classifier as guidance to leverage its learned emotional priors. Previous methods, such as classifier guidance [32] or universal guidance [18], back-propagate directly to the predicted noise $\epsilon_\theta\left(z_t, t, c\right)$ as in Eq. 6. Yet, this often leads to unstable image synthesis, as the gradients derived from the emotional classifier may be adversarial to

those generated by the text prompts. Inspired by [49], we propose instead to optimize extra **emotional tokens** S for emotion synthesis.

Specifically, as shown in Fig. 1, the original text prompt P is embedded by a text encoder $\mathcal{T}$, and then concatenated with extra emotional tokens S, to obtain the emotionalized text embedding $c^{emo} = \mathcal{T}(P) \oplus S$; here $\oplus$ is concatenation of embeddings. Then, we compute the denoised embeddings $\hat{z}_t^{emo}$ conditioned on c^{emo} by replacing c in Eq. 6.

To well adapt an emotion classifier $\mathcal{D}_{emo}$ pre-trained with clean images to noisy data, we adopt the one-step inference similar to [18] to obtain the estimated clean embedding $\hat{z}_0^{emo}$ from each time step t as follows:

$$\hat{z}_0^{emo} = \frac{\hat{z}_t^{emo} - \left(\sqrt{1-\bar{\alpha}_t}\right)\epsilon_\theta\left(\hat{z}_t^{emo}, t, c^{emo}\right)}{\sqrt{\bar{\alpha}_t}}, \tag{7}$$

where $\bar{\alpha}_t = \prod_{l=1}^{t} \alpha_l$. We then optimize emotional tokens S via back-propagated gradients by minimizing the loss $\mathcal{L}_{\text{emo}}$ between the desired emotion y_{target} and the predicted emotion $\hat{y}_{\text{emo}}$=$\mathcal{D}_{\text{emo}}(\hat{z}_0^{\text{emo}})$ during the denoising process, as:

$$S^{'} = \arg\min_{S} \mathcal{L}_{emo}\left(y_{\text{target}}, p\left(\hat{y}_{emo} \mid \hat{z}_0^{emo}\right)\right), \tag{8}$$

where $\mathcal{L}_{emo}$ is the target loss function detailed in Sec. IV-C.

Our design of optimizing extra emotional tokens $S^{'}$ offers two main advantages: (1) It leverages prior knowledge from SOTA emotion classifiers directly, removing the need to create a large-scale emotional manipulation dataset. This is particularly beneficial given the subjectivity of emotions and the tremendous efforts required to create such a dataset. (2) Using both $S^{'}$ and P together as framework inputs allows ϵ_θ to simultaneously consider both semantic and emotional information from the prompts when estimating noise. This prevents gradient inconsistencies, avoids ignoring the textual information and ensures stable synthesis.

Finally, since the emotional tokens S only affect the model's input, there is no need for further complex operations on intermediate embeddings. The framework can be easily applied to both tasks by simply adapting the prompt and emotional token inputs (P, S) and noise samplings z_T.

Application in Generation: Existing emotion generation methods like EmoGen [9] synthesize emotions only according to the target emotion, while our method can achieve it both with and without text descriptions: For the generation with a text prompt, we set P textual description (e.g. "An elder woman cutting cake into slices") as in Fig. 1 and randomly initialize tokens S. For the generation with only target emotion, we simply set P = "an image of "+ target emotion. Both cases start from a random noise $z_T \sim \mathcal{N}(0, I)$.

Application in Editing: For the editing task, we use the description of the source image to set P and randomly initialize tokens S. The noise z_T is obtained from the source image by using the inversion in DDIM [50].

B. Supervision on Semantic Similarity

When to start emotional guidance during the denoising is another challenge for emotion synthesis:

- **Adding guidance too early.** During the early stage of denoising, the image content is not yet determined [51]. Using both the text prompt P and the optimized emotional tokens $S^{'}$ as inputs may cause the network to focus excessively on the target emotion but neglect the semantics. The final image has an accurate target emotion but risks semantic deviation.
- **Adding guidance too late.** In the later stages of denoising, where the main content of the image is almost decided, the model tends to concentrate only on details such as texture and subtle structures. At this point, strong inherent emotions present in the textual prompts or image content (e.g. an image of a park during the day has the inherent emotion of "contentment") may risk incorporating also the target emotion. This results in the final image with accurate semantic but emotional deviation.

Introducing the guidance of the target emotion at the **right timing** can suppress inherent emotions while preserving semantic accuracy. To achieve this without increasing the complexity of the denoising, we leverage the image encoder $\mathcal{E}_I^{\text{clip}}$ and text encoder $\mathcal{E}_T^{\text{clip}}$ from CLIP [13] to measure the semantic similarity s_t^{clip} (with a temperature factor $\tau = 0.07$) between the generated content and text prompts at each time step t:

$$s_t^{clip} = \frac{1}{\tau}\text{CLIP}(\mathcal{E}_I^{\text{clip}}(\mathcal{D}(\hat{z}_{t\to 0})), \mathcal{E}_T^{\text{clip}}(P)), \tag{9}$$

where $\hat{z}_{t\to 0}$ is the embedding iteratively derived from $\hat{z}_t$ to the time step $t = 0$ by using Eq. 3. In the reverse process, once s_t^{clip} exceeds a preset threshold η, it indicates that the image content has preliminarily formed, then the emotional guidance S is optimized and introduced:

$$c^{\text{emo}} = \begin{cases} \mathcal{T}(P) \oplus \varnothing & s_t^{clip} < \eta \\ \mathcal{T}(P) \oplus S^{'} & s_t^{clip} >= \eta \end{cases}, \tag{10}$$

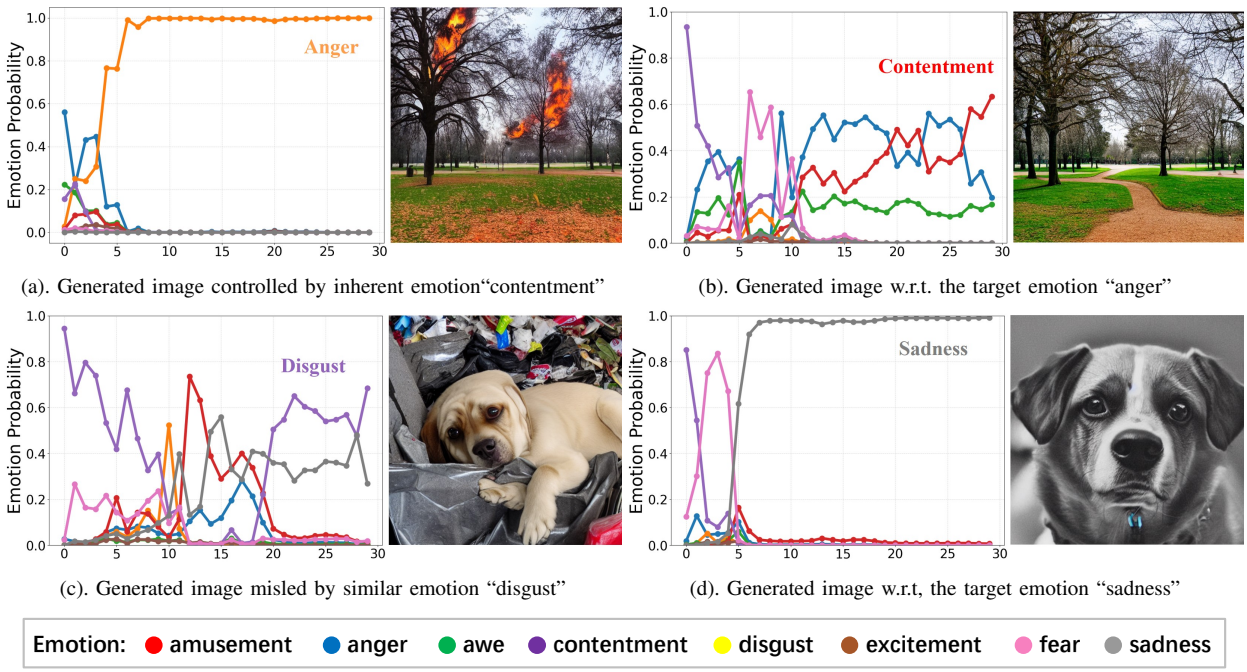
with $\varnothing$ is a void placeholder. This mechanism enables us to introduce target emotional guidance at the right stage, ensuring the convey of the target emotion.

C. Suppression of Undesired Emotions

Which emotion to guide the synthesis should be emphasized throughout the generation. Two types of undesired emotions affect the expression of the target emotion:

(1) Inherent Emotions derived from the source image or text prompt. E.g., an image or text related to "park" naturally carries the inherent emotion "contentment". Once semantic content is primarily formed, this inherent emotion continuously tries to take control of the expressed emotion (see Fig. 2(a)). To suppress the inherent emotion, we propose applying the semantic-similarity supervision mechanism from Sec. IV-B, obtaining the embedding $\hat{z}_{t'}$ at the time step t' before s_t^{clip} exceeds the threshold η (i.e. t'=t+1). We then predict the inherent emotion y_{inh} from $\hat{z}_{t'}$ with the emotional classifier, and apply a cross-entropy loss to suppress the emotional conflict:

$$\mathcal{L}_{\text{inh}} = \ell_{\text{ce}}(y_{\text{inh}}, p(\hat{y}_{\text{emo}} \mid \hat{z}_0^{\text{emo}})), \tag{11}$$



(a). Generated image controlled by inherent emotion"contentment"

(b). Generated image w.r.t. the target emotion "anger"

(c). Generated image misled by similar emotion "disgust"

(d). Generated image w.r.t, the target emotion "sadness"

Fig. 2: Visualization of suppressing similar and inherent emotions. (a) and (b) show the generation of a "park" image w.r.t the target emotion "anger", with and without inherent emotion suppression. (c) and (d) show the generation of a "dog staring at you" image w.r.t the target emotion "sadness", with and without similar emotion suppression. The x-axis denotes the number of inner emotional optimization loops, and the y-axis represents the emotion probability predicted by the classifier.

Backbone	Methods	Emo-A (%)↑	Emo-B (%) ↑	Emo-C (%)↑	FID ↓			CLIP ↑	HPSV2 ↑
					EmoSet [53]	FI_8 [54]	COCO [55]		
Transformer-Unet	PixArt-α [52]	23.87	25.29	25.54	63.16	63.90	51.91	30.33	0.22
SD3.0	SD3 [19]	18.28	21.07	21.41	76.68	79.70	42.21	31.51	0.18
FLUX1.0	FLUX [20]	18.93	21.80	20.73	80.67	82.53	52.23	30.76	0.24
SD1.5	SD [7]	15.00	15.81	16.55	64.34	67.18	41.02	28.72	0.18
	UG [18]	23.46	33.10	18.78	67.04	62.87	61.68	23.73	0.21
	EmoGen [9]	30.96	28.72	28.69	43.80	45.96	59.11	23.39	0.17
	MUSE	**68.38**	**59.38**	**32.23**	**43.53**	**44.18**	**26.52**	**30.33**	**0.24**
SDXL1.0	SDXL [56]	19.48	21.50	20.35	69.54	**69.42**	39.76	30.59	**0.26**
	UG [18]	18.03	17.45	16.10	131.66	120.74	126.14	26.14	0.18
	MUSE	**41.87**	**39.87**	**30.14**	**69.30**	69.87	**38.28**	**31.23**	0.24

TABLE I: **Comparison against the SOTA T2I emotion generation methods.** For all methods, when accessible, we evaluate both the SD [7] and SDXL [56] backbones. Both UG, EmoGen and MUSE take emotion as a label for supervision, while SD, SDXL, SD3 and FLUX textualize emotion as the prompt as in PixArt-α [52]. We use 3 classifiers to measure accuracy: Emo-A with the classifier pretrained by EmoSet [53], Emo-B with the training-agnostic classifier from [9], and Emo-C with the same classifier as Emo-A, but pretrained on FI_8 [54]. We also measure FID on 3 datasets, COCO [55], Emoset [53], and FI_8 [54]. We further resepcitvely quantify the text-image adherence and human preference with CLIP score (CLIP) and HPSV2 [26]. The best performance is highlighted in **bold** with a gray background, and the second-best performance is underlined.

where ℓ_{ce} is a cross-entroy loss. This loss term can suppress the inherent emotion in synthesized images (see Fig. 2(b)).

(2) Similar Emotions usually relate to the target emotion. Unlike other classification tasks where categories have distinct and mutually exclusive characteristics, categories of emotion exhibit relative similarity [47]. For example, "disgust" and "sadness" are similar emotions positioned adjacent to each other in Mikel's wheel psychological model [57]. Similar emotions can lead to a shift in the guided emotion toward them during the synthesis (see Fig. 2(c)). The similar emotions y_{sim} associated with the target emotion y_{target} can be obtained by querying the emotion wheel [57], and they can be suppressed using cross-entropy loss:

$$\mathcal{L}_{\text{sim}} = \ell_{\text{ce}}(y_{\text{sim}}, p(\hat{y}_{\text{emo}}|\, \hat{z}_0^{\text{emo}})). \tag{12}$$

In summary, to suppress these two types of undesired emotion and to force moving towards the target emotion, we propose the following multi-emotion loss used in Eq. 8:

$$\mathcal{L}_{\text{emo}} = \mathcal{L}_{\text{target}} - \lambda_1 \mathcal{L}_{\text{inh}} - \lambda_2 \mathcal{L}_{\text{sim}}, \tag{13}$$

where $\mathcal{L}_{\text{target}} = \max(0, \ell_{\text{ce}}(y_{\text{target}}, p(\hat{y}_{\text{emo}}|\, \hat{z}_0^{emo}))$ is loss in terms of target emotion, and λ_1, λ_2 are hyperparameters to control the suppression strength. During each step, an inner-loop optimization is performed on the emotional token, and

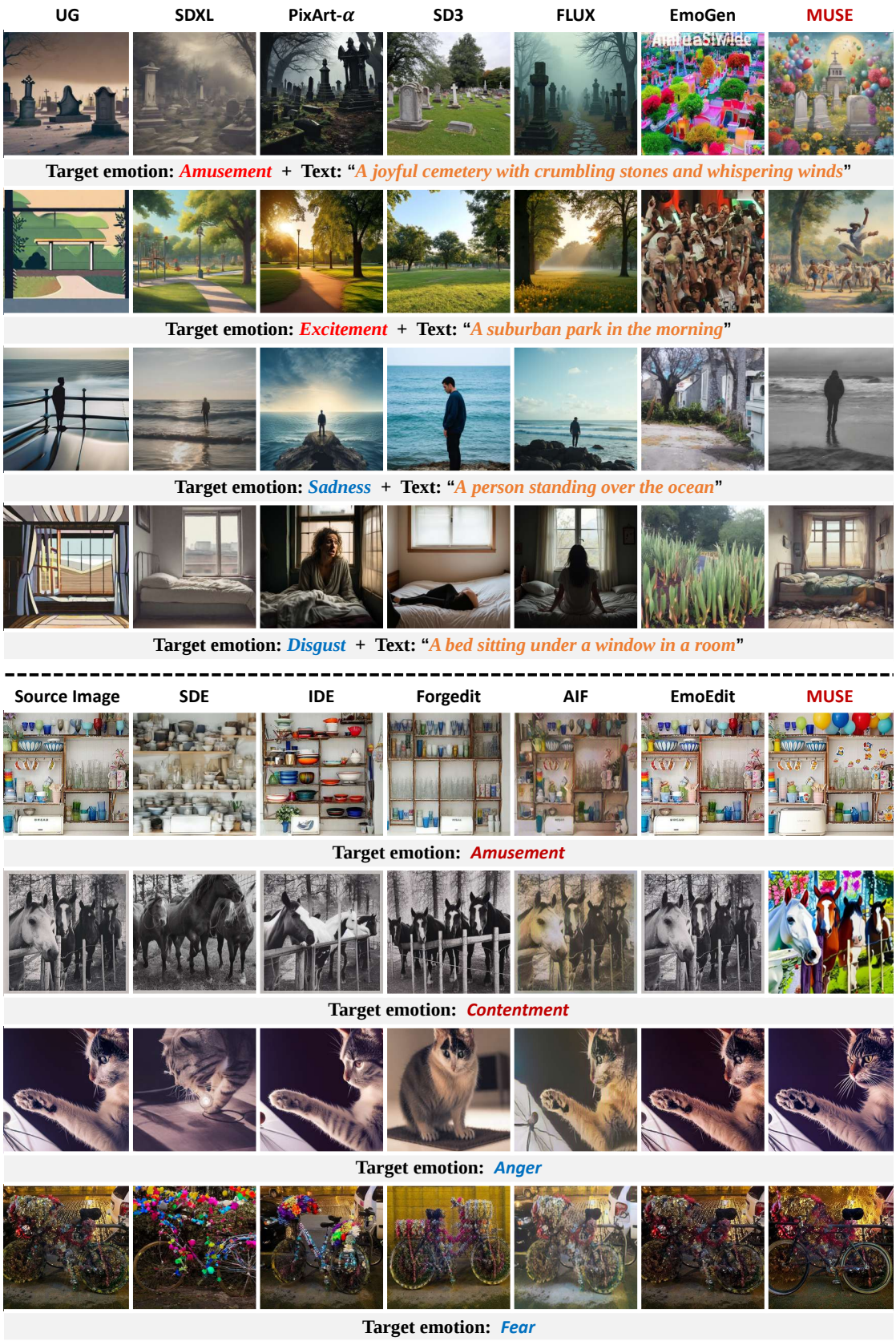


Fig. 3: Qualitative comparison of emotional image generation and editing methods. **Upper Part**: Emotion generation results from SD [7], UG [18], PixArt-α [52], SDXL [7], SD3 [19], FLUX [20], and EmoGen [9]. UG and MUSE are conditioned on both text prompts and emotion labels, while other methods rely solely on emotionally descriptive prompts. **Lower Part**: Emotion editing results from SDE [21], AIF [6], IDE [22], Forgedit [24], and EmoEdit [11]. All editing methods take a neutral input image and modify it according to the target emotion.

Methods	Emo-A (%)↑	Emo-B (%)↑	Emo-C (%)↑	CLIP ↑	HPSV2 ↑
SDE [21]	14.80	15.24	15.49	0.78	0.23
PnP [23]	13.37	13.50	13.63	**0.86**	0.23
AIF [6]	12.44	12.44	12.56	**0.86**	0.21
Forgedit [24]	13.50	13.19	13.69	0.76	0.23
IDE [22]	23.87	25.29	14.04	0.83	**0.24**
EmoEdit [11]	29.31	29.13	27.44	**0.86**	0.22
MUSE	**69.50**	**59.86**	**34.43**	0.74	0.23

TABLE II: Comparisons with the SOTA methods on the editing task. In all cases, the input condition is an image and the emotion label.

ended once the loss $\mathcal{L}_{\text{emo}}$ is lower than a preset threshold 10^{-4}.

V. Experiments

A. Dataset and Evaluation

Datasets: We carry out experiments on three datasets: EmoSet [53], COCO [55], and FI_8 [54]. EmoSet [53] consists of 118,102 images of eight emotion classes. The off-the-shelf classifier we use is pre-trained on this dataset. Following [7], [56], we also employ COCO [55] due to its diversity and detailed text descriptions by sampling image-text pairs from it. FI_8 [54] is used as an external validation set to evaluate image fidelity, which contains 22,697 images.

Metrics & Settings: For the generation task, we assess image fidelity using FID [25] and emotion accuracy with Emo-A, Emo-B and Emo-C. Emo-A is measured using EmoSet-pretrained SimEmo [40] classifier, while Emo-B involves a classifier directly from [9] to fairly show generalization in emotional accuracy. And Emo-C also measured with the same classifier as used by Emo-A, but pre-trained on FI_8 [54]. We use the CLIP score [13] to evaluate text-image alignment. And we use HPSV2 [26] to assess image quality and aesthetics. For the editing task, we use Emo-A, Emo-B, Emo-C, CLIP, and HPSV2. We use the pre-trained Stable Diffusion v1.5[1], Stable Diffusion XL v1.0[2] and CLIP model[3] as model backbone and semantic similarity measurement. We adopt the Adam optimizer [58] with a linearly decaying learning rate $lr = 0.01 \cdot \left(1 - \frac{t}{100}\right)$. To balance the loss terms, we set $\lambda_1 = 0.0005$ and $\lambda_2 = 0.0015$ for $\mathcal{L}_{\text{inh}}$ and $\mathcal{L}_{\text{sim}}$.

B. Comparison to the state of the art

For the generation task, we compare MUSE against backbones Stable Diffusion (SD) [7], SDXL [56], SD3 [19] and FLUX [20], as well as one T2I synthesis SOTA method PixArt-α [52]. These methods explicitly incorporate the target emotion through textual prompts. We also benchmark against Universal Guidance (UG) [18], and EmoGen [9], a SOTA emotional image generation method. For editing task, we compare MUSE with representative methods such as SDEit (SDE) [21], PnP [23], Forgedit [24], InstDiffEdit (IDE) [22], AIF [6] and EmoEdit [11]. And we built MUSE on SD and SDXL backbones for both tasks. We adopted the categorization of emotion from [57], which include four positive (amusement, awe, contentment, excitement) and four negative emotions (anger, disgust, fear, and sadness). We randomly sampled text captions from the COCO dataset [55], which provides diverse textual descriptions.

Emotional Generation: For the emotional generation task, it's crucial to align both the textual semantic, image content and target emotion. We generated 700 images per emotion (totally 5600 images) with MUSE and evaluate its performance with six metrics. We demonstrate quantitative and qualitative comparisons in Table I and upper part of Fig. 3, respectively. We can observe from Tab. I that MUSE built on both SD and SDXL backbones achieves the highest emotional evocation accuracy (Emo-A, B & C), surpassing the 2nd-best methods by about 30%, 20% and 10%. Moreover, it also proves the robustness of MUSE as it achieves the best accuracy with a training classifier (Emo-A), an agnostic classifier (Emo-B), and a cross-validated classifier (Emo-C). MUSE also gives the best FID for SD across all datasets, while MUSE-SDXL ranks 1st in two out of three datasets, with FI_8 slightly lagging behind. Universal Guidance [18] shows the worst FID, and EmoGen [9] exhibits degraded FID on COCO, likely due to overfitting to its emotional dataset. Both SD and SDXL deliver strong FIDs, consistent with their high-quality backbones. In terms of adherence, MUSE outperforms others on the CLIP metric, while methods like UG [18] and EmoGen [9] struggle with the text-emotion alignment, often overly focusing either on the text description or the desired emotion in the synthesized images (see Fig. 3). Additionally, as measured by HPSv2 [26], MUSE also exhibits good performance in human aesthetic evaluations.

The upper section of Fig. 3 demonstrates MUSE's superior capability in aligning emotional expression with semantic content. For example, in the 2nd row, our method successfully evokes the target emotion of "excitement" by incorporating "dancing crowds," while baseline methods merely focus on depicting the literal "park" scene. More notably, when handling semantic-emotional conflicts (1st row) where the concept of "cemetery" inherently contradicts "amusement" or "joyful" emotions, MUSE effectively introduces vibrant visual elements to convey the desired emotional tone. In contrast, comparison methods exhibit significant limitations: PixArt-α [52] and SDXL [56] completely fail to express the target emotion, while EmoGen [9] deviates from the original prompt semantics.

Emotional Editing: For the emotional editing task, we randomly select 200 image-text pairs per emotion (totally 1.6K images) from COCO [55]. We compare MUSE with SOTA approaches, including global editing methods SDEdit [21] and PnP [23], the style-based method AIF [6], and the emotion-specific editing method EmoEdit [11]. Additionally, we include editing methods based on the diffusion model such as Forgedit [24] and InstDiffEdit (IDE) [22]. The results are shown in Table II and the lower part of Fig. 3. We observe that AIF [6] PnP [23] and EmoEdit [11] achieve high CLIP similarity of 0.86, but suffer from emotion accuracy lower than 14%. EmoEdit [11] performs better, yet its emotional accuracy remains below 30%. IDE [22] aligns better images

[1] https://huggingface.co/runwayml/stable-diffusion-v1-5
[2] https://huggingface.co/stabilityai/stable-diffusion-xl-base-1.0
[3] https://huggingface.co/openai/clip-vit-large-patch14

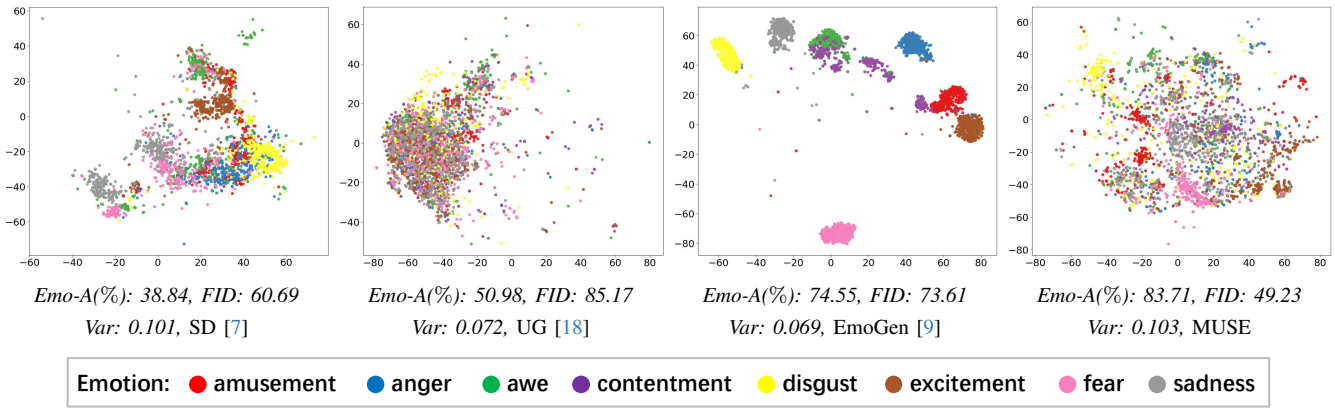


Fig. 4: T-SNE [59] visualization of feature distributions for non-textual emotional generation across four models. Each data point represents the CLIP semantic features of an image, with different colors indicating their corresponding emotion categories. The Var is the averaged intra-class variance of CLIP features. Compared to the other three methods (a), (b), and (c), MUSE achieves better separation of content within the same emotion category while exhibiting a more dispersed overall semantic distribution.

Guidance Classifier	Loss: $\mathcal{L}_{\text{target}}$ -	Optim ET†	Emo-A (%)↑	Emo-B (%)↑	Emo-C (%)↑	FID ↓ EmoSet	FID ↓ FI_8	FID ↓ COCO	CLIP ↑
EmoGen [9]	$\mathcal{L}_{\text{sim}}$ - $\mathcal{L}_{\text{inh}}$	✓	55.38	**72.38**	**33.42**	44.55	46.20	26.92	29.86
SimEmo* [40]	/	✓	66.78	58.05	30.91	44.66	44.82	26.64	29.34
SimEmo* [40]	$\mathcal{L}_{\text{sim}}$	✓	67.05	57.01	31.57	45.04	45.53	29.40	**30.42**
SimEmo* [40]	$\mathcal{L}_{\text{inh}}$	✓	67.04	58.23	31.95	45.76	46.39	26.83	30.20
SimEmo* [40]	$\mathcal{L}_{\text{sim}}$ - $\mathcal{L}_{\text{inh}}$	✗	12.94	12.72	13.26	60.55	67.43	42.37	30.00
SimEmo* [40]	$\mathcal{L}_{\text{sim}}$ - $\mathcal{L}_{\text{inh}}$	✓	**68.38**	59.38	32.23	**43.53**	**44.18**	**26.52**	30.33

† OET indicates Optimizing (or not) the Emotional Token. ∗ means classifier trained by ourselves.

TABLE III: **Ablation study**. The 1st row shows the performance of MUSE using the classifier from EmoGen [9], the 2nd-4th rows show the results of MUSE trained using different loss term combinations, and the 5th row is the results of MUSE when removing the emotional token optimization. The final row shows results using our proposed MUSE setting.

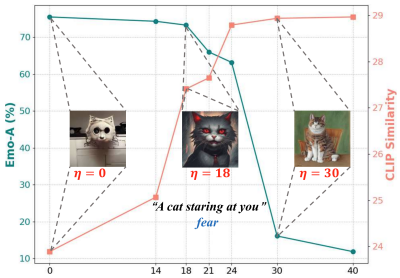


Fig. 5: Evolution of accuracy & similarity in terms of threshold η. Images shown are generated at early, optimal, and late stages of emotional guidance.

Guidance Timing	Emo-A (%)↑	Emo-B (%) ↑	Emo-C (%)↑	FID ↓ EmoSet	FID ↓ FI_8	FID ↓ COCO	CLIP ↑	HPSV2 ↑
$\eta = 0$	**74.30**	**60.85**	**36.62**	88.64	84.91	88.92	23.11	0.23
$\eta = 14$	73.26	59.60	35.76	69.46	66.50	44.53	23.90	0.22
$\eta = 18$	68.38	59.38	32.23	**43.53**	**44.18**	**26.52**	30.33	**0.24**
$\eta = 21$	65.97	55.22	31.63	62.11	68.55	51.10	30.08	0.23
$\eta = 24$	63.16	51.22	27.38	68.40	67.78	50.25	**30.41**	0.23
$\eta = 30$	16.09	14.02	13.80	68.19	67.43	42.37	30.00	0.23

TABLE IV: Generation ablation on η compared with SOTA methods.

with target emotions with emotional accuracy around 24%, but often introduces irrelevant elements (e.g., adding a vase on a cabinet or ribbons on a bicycle in the last row of Fig. 3). In contrast, MUSE produces more natural edits by balancing image content and emotions, achieving notably higher emotion accuracies (at least 7% accuracy gain) and a competitive aesthetic score of 0.23, despite the slightly lower CLIP score, thus highlighting its editing capabilities.

Semantic Visualization To provide a more direct observation of the semantics for generated images, we employed T-SNE [59] to visualize in Fig. 4 the distribution of 3200 generated images (400 images per emotion) in the CLIP semantic space for methods SD [7], UG [18] and EmoGen [9] and measure their corresponding averaged intra-class variance. EmoGen [9] explicitly maps each emotion to specific attributes, resulting in a semantic space that clearly displays several emotional clusters. This structure implies that images sharing the same emotion tend to exhibit similar visual elements, leading to reduced semantic diversity. On the other hand, SD [7] and UG [18] reveal different characteristics. SD [7] prioritizes the preservation of semantic diversity and image quality but tends to overlook emotion. This approach leads to higher variance and FID scores, while yielding lower emotional accuracy. In contrast, UG [18] places greater emphasis on emotion, at the expense of semantic diversity and image quality. Our MUSE model combines the strengths of SD [7], achieving high FID and diversity, while also attaining superior emotional accuracy without explicitly mapping emotions to

attributes, as in EmoGen [9].

Methods	Qua ↓	Sem ↓	Emo ↓	Div ↓	Adh ↓
SD [7]	**2.45**	2.45	2.83	2.50	2.89
UG [18]	2.53	2.62	2.71	2.60	2.54
EmoGen [9]	2.52	2.62	2.56	2.83	2.31
MUSE	2.50	**2.31**	**1.90**	**2.07**	**2.26**

TABLE V: User study, numbers are the average rank.

User Study: We generated 160 images (8 emotions × 5 text prompts) for SD [7], UG [18], EmoGen [9] and MUSE. In a blind evaluation, 15 participants ranked the images across five dimensions: image quality (Qua), semantic fidelity (Sem), emotional fidelity (Emo), content diversity (Div), and adherence between emotion and semantics (Adh). Results are shown in Tab. V. Although MUSE slightly trails SD [7] in image quality, it achieves the highest average rank in other criteria, especially emotional fidelity and diversity. This demonstrates that MUSE provides a more engaging visual experience and stronger emotional impact for viewers.

C. Ablation Study

Emotion Guidance Timing: To prove the importance of emotion guidance timing, i.e. when to start emotional token optimization, we randomly sample 200 textual prompts from the COCO dataset, generating one image per prompt for 8 emotions. We test different thresholds η=$\{0, 14, 18, 21, 24, 30\}$ for generation and show results in Tab. IV. Among them, η=$\{0, 14\}$ and η=$\{24, 30\}$ means early and late stages of guidance introduction respectively. Fig. 5 shows the evolution of emotion accuracy and CLIP scores with respect to η. Early guidance produces correct emotions but incorrect semantics, while late guidance shows inverse results. Only the guidance at the right timing (e.g. η=$\{18, 21\}$) yields a good balance between emotions and semantics. And η=18 gives the best balance of emotion accuracy and image quality. We also remove emotional token optimization (5th row in Tab. III), which is equivalent to η=50, where MUSE fails to synthesize correct emotions entirely.

Emotional Loss Term: To validate the effectiveness of our emotional loss in suppressing inherent and similar emotions, we test different combinations of loss terms in Eq. 13 (see 2nd-4th rows in Tab. III). The results show that adding $\mathcal{L}_{\text{inh}}$ and $\mathcal{L}_{\text{sim}}$ reduces interference from inherent and similar emotions, improving the final emotion evocation accuracy.

Guidance Classifier: To further investigate whether our method can effectively leverage the emotional priors learned by the classifier and to explore the influence of dataset and architecture, we conducted experiments on different classifiers, as shown in Tab. III. Specifically, we refer to the pretrained EmoGen [9] classifier as *EmoGen*, and to an identical classifier architecture adopted from [40] but pretrained on the same EmoSet dataset as *SimEmo*. For metrics, Emo-A and Emo-B are evaluated with the *EmoGen* [9] and *SimEmo* [40] classifiers, whereas Emo-C is assessed using a *SimEmo* model pretrained on the smaller FI_8 dataset (see Sec. V-A).

A comparison between the first (*EmoGen*) and last (*SimEmo*) rows of Tab. III shows that either classifier can effectively steer emotional-content synthesis. Nonetheless, the accuracy metrics (Emo-A/B) are consistently favored toward whichever classifier supplies the guidance signal, exposing certain architecture-dependent bias. These results illustrate the generality of our framework on different classifiers while also motivating evaluation under multiple architectures; accordingly, we adopt *SimEmo* as the default guide in our experiments. Finally, when the classifiers for guidance and testing share the same architecture but are pretrained on different datasets, MUSE demonstrates higher Emo-C than other methods (as shown in Tab. I), though the knowledge gap (Emo-C is trained under a smaller dataset) between the datasets leads to a significant drop in performance.

Given this inherent knowledge gap in datasets, methods like MUSE that can quickly leverage priors from each dataset without the need for retraining exhibit a clear advantage in emotional synthesis.

VI. Conclusion

We presented a unified framework for image emotion synthesis and editing, addressing three key aspects: how, when, and which emotion to be manipulated during synthesis. Specifically, we leverage the prior knowledge of external emotion classifiers and introduce a test-time optimization strategy during inference. Reverse optimization of emotional tokens is applied to maintain synthesis stability, while semantic similarity is used to determine the precise timing of emotional injection, ensuring semantic alignment. In addition, an multi-emotion loss function is employed to suppress undesired emotional interference and alleviate emotional conflicts. Experimental results demonstrate superior performance in both emotional accuracy and visual quality compared to existing methods for both task. Future work will focus on reducing knowledge gap between dataset and inference time. Moreover, as real-world emotional expression encompasses a far richer range than the eight discrete classes, leaving subtler or compound affects will also be a meaningful direction of our work.

References

[1] T. Van Gorp and E. Adams, *Design for emotion*. Elsevier, 2012. 1

[2] S. Mann and J. Cowburn, "Emotional labour and stress within mental health nursing," *J Psychiatr Ment Hlt.*, vol. 12, no. 2, pp. 154–162, 2005. 1

[3] X. Wang, J. Jia, and L. Cai, "Affective image adjustment with a single word," *Visual Comput.*, vol. 29, pp. 1121–1133, 2013. 1

[4] D. Liu, Y. Jiang, M. Pei, and S. Liu, "Emotional image color transfer via deep learning," *Pattern Recogn. Lett.*, vol. 110, pp. 16–22, 2018. 1

[5] S. Sun, J. Jia, H. Wu, Z. Ye, and J. Xing, "MSNet: A deep architecture using multi-sentiment semantics for sentiment-aware image style transfer," in *IEEE Int. Conf. Acoust., Speech, and Sign. Process.*, 2023, pp. 1–5. 1

[6] S. Weng, P. Zhang, Z. Chang, X. Wang, S. Li, and B. Shi, "Affective image filter: Reflecting emotions from text to images," in *IEEE Int. Conf. Comput. Vis.*, 2023, pp. 10 810–10 819. 1, 2, 6, 7

[7] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022, pp. 10 684–10 695. 1, 2, 5, 6, 7, 8, 9

[8] T. Brooks, A. Holynski, and A. A. Efros, "Instructpix2pix: Learning to follow image editing instructions," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2023, pp. 18 392–18 402. 1

[9] J. Yang, J. Feng, and H. Huang, "Emogen: Emotional image content generation with text-to-image diffusion models," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2024, pp. 6358–6368. 1, 2, 4, 5, 6, 7, 8, 9

[10] Y. He, S. Dang, L. Ling, Z. Qian, N. Zhao, and N. Cao, "Emoticrafter: Text-to-emotional-image generation based on valence-arousal model," *arXiv:2501.05710*, 2025. 1, 2

[11] J. Yang, J. Feng, W. Luo, D. Lischinski, D. Cohen-Or, and H. Huang, "Emoedit: Evoking emotions through image manipulation," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2025, pp. 24 690–24 699. 1, 2, 6, 7

[12] Q. Mao, H. Hu, Y. He, D. Gao, H. Chen, and L. Jin, "Emoagent: Multi-agent collaboration of plan, edit, and critic, for affective image manipulation," *arXiv:2503.11290*, 2025. 1, 2

[13] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *Int. Conf. on Mach. Learn.*, 2021, pp. 8748–8763. 1, 4, 7

[14] C.-K. Yang and L.-K. Peng, "Automatic mood-transferring between color images," *IEEE Comput. Graph. App.*, vol. 28, no. 2, pp. 52–61, 2008. 1, 2

[15] S. Zhu, C. Qing, C. Chen, and X. Xu, "Emotional generative adversarial network for image emotion transfer," *Expert Syst. Appl.*, vol. 216, p. 119485, 2023. 1, 2

[16] Q. Lin, J. Zhang, Y. S. Ong, and M. Zhang, "Make me happier: Evoking emotions through image diffusion models," *arXiv:2403.08255*, 2024. 1, 2

[17] N. Ma, S. Tong, H. Jia, H. Hu, Y.-C. Su, M. Zhang, X. Yang, Y. Li, T. Jaakkola, X. Jia *et al.*, "Inference-time scaling for diffusion models beyond scaling denoising steps," *arXiv:2501.09732*, 2025. 1, 2

[18] A. Bansal, H.-M. Chu, A. Schwarzschild, S. Sengupta, M. Goldblum, J. Geiping, and T. Goldstein, "Universal guidance for diffusion models," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2023, pp. 843–852. 2, 3, 4, 5, 6, 7, 8, 9

[19] P. Esser, S. Kulal, A. Blattmann, R. Entezari, J. Müller, H. Saini, Y. Levi, D. Lorenz, A. Sauer, F. Boesel *et al.*, "Scaling rectified flow transformers for high-resolution image synthesis," in *Int. Conf. on Mach. Learn.*, 2024. 2, 5, 6, 7

[20] B. F. Labs, "Flux," https://github.com/black-forest-labs/flux, 2024. 2, 5, 6, 7

[21] C. Meng, Y. He, Y. Song, J. Song, J. Wu, J.-Y. Zhu, and S. Ermon, "Sdedit: Guided image synthesis and editing with stochastic differential equations," in *Int. Conf. Learn. Represent.* 2, 6, 7

[22] S. Zou, J. Tang, Y. Zhou, J. He, C. Zhao, R. Zhang, Z. Hu, and X. Sun, "Towards efficient diffusion-based image editing with instant attention masks," in *AAAI Conf. Artif. Intell.*, vol. 38, no. 7, 2024, pp. 7864–7872. 2, 6, 7

[23] N. Tumanyan, M. Geyer, S. Bagon, and T. Dekel, "Plug-and-play diffusion features for text-driven image-to-image translation," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2023, pp. 1921–1930. 2, 7

[24] S. Zhang, S. Xiao, and W. Huang, "Forgedit: Text guided image editing via learning and forgetting," *arXiv:2309.10556*, 2023. 2, 6, 7

[25] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *Adv. Neural Inform. Process. Syst.*, 2017. 2, 7

[26] X. Wu, K. Sun, F. Zhu, R. Zhao, and H. Li, "Human preference score: Better aligning text-to-image models with human preference," in *IEEE Int. Conf. Comput. Vis.*, 2023, pp. 2096–2105. 2, 5, 7

[27] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *Int. Conf. on Mach. Learn.*, 2015, pp. 2256–2265. 2

[28] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Adv. Neural Inform. Process. Syst.*, vol. 33, pp. 6840–6851, 2020. 2, 3

[29] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans *et al.*, "Photorealistic text-to-image diffusion models with deep language understanding," *Adv. Neural Inform. Process. Syst.*, vol. 35, pp. 36 479–36 494, 2022. 2

[30] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," *arXiv:2204.06125*, vol. 1, no. 2, p. 3, 2022. 2

[31] J. Ho and T. Salimans, "Classifier-free diffusion guidance," in *Adv. Neural Inform. Process. Syst. worksh.* 2, 3

[32] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," *Adv. Neural Inform. Process. Syst.*, vol. 34, pp. 8780–8794, 2021. 2, 3

[33] H. Ye, J. Zhang, S. Liu, X. Han, and W. Yang, "Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models," *arXiv:2308.06721*, 2023. 2

[34] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2023, pp. 3836–3847. 2

[35] C. Mou, X. Wang, L. Xie, Y. Wu, J. Zhang, Z. Qi, and Y. Shan, "T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models," in *AAAI Conf. Artif. Intell.*, vol. 38, no. 5, 2024, pp. 4296–4304. 2

[36] S. Zhao, H. Yao, Y. Yang, and Y. Zhang, "Affective image retrieval via multi-graph learning," in *ACM Int. Conf. Multimedia*, 2014, pp. 1025–1028. 2

[37] J. Yang, D. She, M. Sun, M.-M. Cheng, P. L. Rosin, and L. Wang, "Visual sentiment prediction based on automatic discovery of affective regions," *IEEE Trans. Multimedia*, vol. 20, no. 9, pp. 2513–2525, 2018. 2

[38] W. Zhang, X. He, and W. Lu, "Exploring discriminative representations for image emotion recognition with cnns," *IEEE Trans. Multimedia*, vol. 22, no. 2, pp. 515–523, 2019. 2

[39] L. Xu, Z. Wang, B. Wu, and S. Lui, "Mdan: Multi-level dependent attention network for visual emotion analysis," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022, pp. 9479–9488. 2

[40] S. Deng, L. Wu, G. Shi, L. Xing, W. Hu, H. Zhang, and Y. Xiang, "Simple but powerful, a language-supervised method for image emotion classification," *IEEE Trans. Affect. Comput.*, vol. 14, no. 4, pp. 3317–3331, 2022. 2, 7, 8, 9

[41] J. Pan, J. Lu, and S. Wang, "A multi-stage visual perception approach for image emotion analysis," *IEEE Trans. Affect. Comput.*, 2024. 2

[42] Z. Zhang, P. Zhao, E. Park, and J. Yang, "Mart: Masked affective representation learning via masked temporal distribution distillation," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2024, pp. 12 830–12 840. 2

[43] H. Zhang and M. Xu, "Recognition of emotions in user-generated videos with kernelized features," *IEEE Trans. Multimedia*, vol. 20, no. 10, pp. 2824–2835, 2018. 2

[44] W. Nie, Y. Bao, Y. Zhao, and A. Liu, "Long dialogue emotion detection based on commonsense knowledge graph guidance," *IEEE Trans. Multimedia*, vol. 26, pp. 514–528, 2023. 2

[45] N. Lu, Z. Han, and Z. Tan, "A hypergraph based contextual relationship modeling method for multimodal emotion recognition in conversation," *IEEE Trans. Multimedia*, 2024. 2

[46] H. Ma, J. Wang, H. Lin, B. Zhang, Y. Zhang, and B. Xu, "A transformer-based model with self-distillation for multimodal emotion recognition in conversations," *IEEE Trans. Multimedia*, 2023. 2

[47] Z. Chenxi, S. Jinglei, N. Liqiang, and Y. Jufeng, "To err like human: Affective bias-inspired measures for visual emotion recognition evaluation," *Adv. Neural Inform. Process. Syst.*, 2024. 2, 5

[48] T.-J. Fu, X. E. Wang, and W. Y. Wang, "Language-driven artistic style transfer," in *Eur. Conf. Comput. Vis.*, 2022, pp. 717–734. 2

[49] R. Gal, Y. Alaluf, Y. Atzmon, O. Patashnik, A. H. Bermano, G. Chechik, and D. Cohen-or, "An image is worth one word: Personalizing text-to-image generation using textual inversion," in *Int. Conf. Learn. Represent.* 4

[50] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," in *Int. Conf. Learn. Represent.* 4

[51] A. Hertz, R. Mokady, J. Tenenbaum, K. Aberman, Y. Pritch, and D. Cohen-Or, "Prompt-to-prompt image editing with cross-attention control," in *Int. Conf. Learn. Represent.*, 2023. 4

[52] J. Chen, J. Yu, C. Ge, L. Yao, E. Xie, Z. Wang, J. T. Kwok, P. Luo, H. Lu, and Z. Li, "Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis," in *Int. Conf. Learn. Represent.*, 2024. 5, 6, 7

[53] J. Yang, Q. Huang, T. Ding, D. Lischinski, D. Cohen-Or, and H. Huang, "Emoset: A large-scale visual emotion dataset with rich attributes," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2023, pp. 20 383–20 394. 5, 7

[54] Q. You, J. Luo, H. Jin, and J. Yang, "Building a large scale dataset for image emotion recognition: The fine print and the benchmark," in *AAAI Conf. Artif. Intell.*, vol. 30, no. 1, 2016. 5, 7

[55] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Eur. Conf. Comput. Vis.* Springer, 2014, pp. 740–755. 5, 7

[56] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, and R. Rombach, "Sdxl: Improving latent diffusion models for high-resolution image synthesis," in *Int. Conf. Learn. Represent.* 5, 7

[57] J. A. Mikels, B. L. Fredrickson, G. R. Larkin, C. M. Lindberg, S. J. Maglio, and P. A. Reuter-Lorenz, "Emotional category data on images from the international affective picture system," *Behav. Res. Methods*, vol. 37, pp. 626–630, 2005. 5, 7

[58] D. P. Kingma, "Adam: A method for stochastic optimization," *arXiv:1412.6980*, 2014. 7
[59] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *J. Mach. Learn. Res.*, vol. 9, no. 11, 2008. 8