SAME OBJECT, DIFFERENT EMOTIONS: OBJECT-CENTRIC EMOTIONAL CONTENT GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Text-to-image models can generate semantically coherent images, yet often struggle to make a specified object convey a desired emotion. Existing methods typically rely on scene-level cues or distort the object to strengthen emotional expression, compromising object identity and realism. We address object-based emotion-controllable image generation, where both the target object and the overall image should express the intended emotion while remaining visually plausible. To support this task, we construct Obj-EmoSet, which augments affective images with object categories, basic content descriptions, and emotion-aware descriptions. We then propose an Emotion-Aware Visual Planner that translates abstract emotions into object-compatible states, actions, and scene conditions. A Visual Realism Refiner improves composition, spatial relationships, and textures while retaining emotional cues, and an Emotion Calibration Adapter aligns emotional semantics with visual features. Together, these components enable object-level emotional expression while preserving semantic consistency and visual realism. Extensive qualitative and quantitative experiments demonstrate strong performance in object-level emotion control, semantic consistency, and image realism, advancing emotion-aware generation from coarse global modulation toward fine-grained object-centric control.

1 INTRODUCTION

Existing text-to-image models have demonstrated strong capabilities in semantic understanding and content generation, yet remain limited in modeling affective semantics. Emotion is an abstract visual concept that can be conveyed through both global scene-level cues and the visual state of specific objects. Therefore, emotional image generation requires not only establishing an appropriate overall atmosphere, but also enabling the target object itself to express the intended emotion while preserving its semantic identity and visual realism.

Although existing emotional generation methods are capable of producing visual content with clear emotional tendencies, they still exhibit two major limitations. First, most existing approaches primarily focus on emotion expression at the global image level and have not yet achieved effective **emotion control based on target-object semantics** Dang et al. (2025); Yang et al. (2024). Second, to enhance emotional discriminability, these methods often introduce emotion-related visual features to strengthen affective expression, which may compromise the realism and naturalness of the generated content Yang et al. (2026); Black Forest Labs (2025); Chen et al. (2024b); Liu et al. (2025). Therefore, how to preserve the semantic stability of the target object while enabling both the object itself and the overall image to consistently convey the desired emotion, without sacrificing visual realism and naturalness, remains a critical challenge. As illustrated in Fig. 1, existing generative models still exhibit clear limitations in object-centric emotion control. For example, when generating combinations such as "Disgust + Chair" or "Fear + Stone," these models often rely on human facial expressions or contextual scene cues to convey the target emotion, rather than expressing the emotion through the target object itself. PixArt-α, in particular, tends to introduce exaggerated, anthropomorphic, or fantastical modifications to the object, such as transforming a chair into a throne-like structure or a stone into a monster-like form, thereby compromising semantic fidelity and visual realism. In comparison, EmoGen better preserves the object category, but its ability to express the specified emotion remains limited. Overall, existing methods still fall short in terms of **object–emotion consistency, semantic preservation, and content realism**. Emotional image

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

generation aims to transform abstract emotional semantics into concrete, plausible, and perceptible visual content. Unlike conventional visual attributes, emotional expression is typically determined by the joint effects of image content and object state. Therefore, the key challenge of emotional image generation lies not only in establishing an overall affective atmosphere, but also in ensuring that the target object itself conveys an emotion consistent with the desired affective condition.

Existing affective image datasets, such as EmoSet Yang et al. (2023), primarily provide image-level emotion labels and content descriptions for the entire image. They generally lack explicit modeling of object categories and their emotion-related states, making them insufficient for directly supporting object-based controllable emotional generation. To address this limitation, we construct a new dataset, termed **Obj-EmoSet**, based on EmoSet. Obj-EmoSet simultaneously provides object categories, basic content descriptions, and emotion-aware content descriptions.

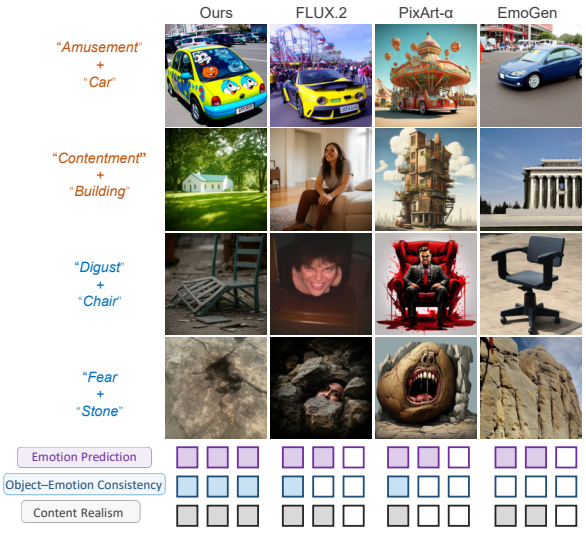


Figure 1: FLUX.2 is realistic but weak in object-level emotion. PixArt-α uses exaggerated cues, reducing realism. EmoGen preserves semantics but shows limited emotional expression. Our method achieves object-level emotion control while maintaining semantic consistency and realism.

However, even with such fine-grained object-level annotations, accurately grounding abstract emotions in specific objects remains challenging. Emotion is inherently abstract and complex, and its visual manifestation does not correspond to a fixed or unique representation. The same emotion, when associated with different objects, may be expressed through substantially different states, actions, shapes, or visual relationships. Consequently, there exists a significant **Affection gap** between abstract emotional semantics and concrete visual manifestations. Bridging this gap requires transforming emotions into perceptible and generatable visual states that are compatible with the intrinsic properties of the target object, which is a key challenge for object-level emotional control.

To bridge the **Affection gap** between abstract emotions and concrete visual representations, we propose an **Emotion-Aware Visual Planner**. This module translates abstract emotional concepts into object-related visual conditions, including object states, actions, composition, lighting, and atmosphere, enabling emotions to be concretely manifested in the target object and its surrounding scene. However, excessively emphasizing emotional cues may lead to overly abstract visual representations and compromise image realism. To address this issue, We further introduce a **Visual Realism Refiner** to improve composition, spatial relationships, and textures while preserving emotional cues, balancing visual realism and emotional expressiveness. The **Emotion Calibration Adapter** aligns emotional semantics with visual features for more natural emotional expression. The main contributions of this work are summarized as follows:

1. We investigate the problem of **object-based emotion-controllable content generation** and construct **Obj-EmoSet**, which provides object categories, basic content descriptions, and emotion-aware content descriptions, offering fine-grained supervision for object-level emotion modeling.

2. We propose an **Emotion-Aware Visual Planner** that translates abstract emotions into actionable visual conditions associated with target objects, establishing an explicit mapping from emotional semantics to concrete visual representations.

3. We introduce a **Visual Realism Refiner** and an **Emotion Calibration Adapter**, and adapt the base generative model with real-world emotional images. Together, they preserve object-level semantic consistency while improving emotional expressiveness and visual realism.

2 Related Work

2.1 Text-to-Image Content Generation

Text-to-image generation aims to synthesize visual content from natural-language descriptions. Recent large-scale generative models have enabled high-quality open-domain generation, with latent diffusion models (LDMs) improving efficiency by operating in a compressed latent space and using cross-attention to align textual and visual features Rombach et al. (2022a). However, text prompts alone often provide insufficient control over spatial structure, local appearance, and subject identity. Existing methods therefore introduce additional conditioning signals. ControlNet uses structural cues such as edges, depth, and poses Zhang et al. (2023); IP-Adapter incorporates reference images for appearance and semantic guidance Ye et al. (2023); and DreamBooth learns subject-specific representations from a few examples to improve identity consistency Ruiz et al. (2023).

These methods are effective when explicit visual constraints are available. Emotional generation is more ambiguous, as emotions such as amusement, fear, sadness, and awe do not correspond to fixed visual targets and instead depend on object states, actions, composition, and scene context. Our task focuses on object-level emotional generation without additional visual references at inference time. The model must preserve the target object and its stable attributes while inferring plausible visual states consistent with both the target emotion and the object's properties.

2.2 Controllable Image Editing and Content Preservation

Controllable image editing aims to modify specified semantic attributes while preserving visual content that is irrelevant to the edit. A key challenge is therefore to achieve a favorable trade-off between editability and content preservation. Prior work addresses this challenge through different forms of conditioning and structural constraints. Prompt-to-Prompt improves spatial consistency in text-guided editing by constraining cross-attention maps Hertz et al. (2022). InstructPix2Pix learns to perform image editing directly from natural-language instructions Brooks et al. (2023), while Blended Latent Diffusion localizes edits using spatial masks Avrahami et al. (2023). Subject-specific and reference-conditioned approaches further preserve subject identity and appearance while allowing selected visual attributes to be modified Ruiz et al. (2023); Ye et al. (2023).

Emotional editing, however, introduces a distinct form of ambiguity. Unlike conventional semantic attributes, an emotion does not correspond to a fixed visual transformation: the same target emotion may require substantially different changes depending on the object being edited. As a result, overly conservative editing may fail to convey the intended emotion, whereas aggressive editing can distort object identity or structure. To address this tension, we decompose object appearance into *stable object attributes* and *controllable emotional states*. Our approach preserves category, identity, and basic structure, while selectively modifying visual factors that are relevant to the target emotion.

2.3 Emotional Image Generation and Manipulation

Emotion-controllable image generation aims to synthesize or edit visual content that expresses a specified emotion. Unlike explicit controls such as color, pose, or layout, emotion is an abstract semantic condition that can be conveyed through diverse visual cues, including scene properties, object states, and their interactions Kuniecki et al. (2017); Pilarczyk & Kuniecki (2014). Early work mainly focused on domains with explicit affective cues, such as facial expressions, talking faces, and landscapes Azari & Lim (2024); Xu et al. (2023); Park & Lee (2020). With the development of text-to-image diffusion models, recent methods extend emotion control to open-domain generation. EmoGen maps abstract emotions to visual-semantic representations for emotional content generation Yang et al. (2024), while EmoCtrl jointly models content and emotion conditions Yang et al. (2026). UniEmo further explores unified emotion understanding and generation Zhu et al. (2026).

Emotion-aware editing has also been studied. EmoEdit modifies emotionally relevant image content while preserving the original composition Yang et al. (2025), and Make Me Happier uses an input image to constrain subject and scene semantics during emotion-eliciting editing Lin et al. (2025). These editing-based approaches benefit from structural and identity information already provided by the source image. Despite this progress, existing methods primarily target image-level emotional consistency and offer limited control over the emotional state of a specific object. Emotional expression may be dominated by background, lighting, or global appearance, while unconstrained semantic changes can cause identity drift, structural inconsistency, or implausible object states. In contrast, we model emotion at the object-state level, preserving core object characteristics while

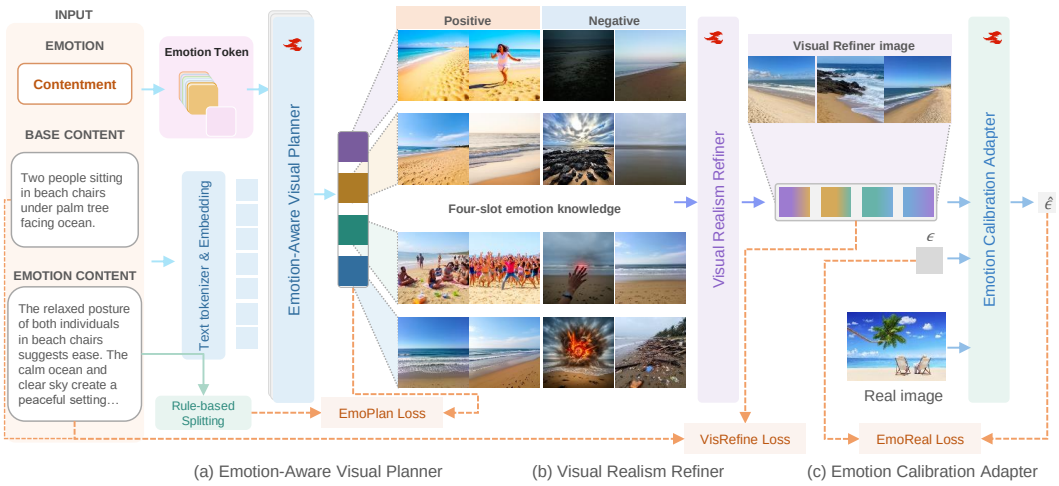


Figure 3: Overview of Same Object, Different Emotions. (1) emotion-aware visual planning for mapping abstract emotions to object-level visual conditions; (2) visual realism refinement for balancing emotional expressiveness and image realism; and (3) an emotion realism adapter for enhancing realistic object–emotion correspondence. Visualization of multi-slot emotion-aware content during inference, given "Ocean" as the input and "Contentment" as the target emotion.

restricting emotional changes to plausible states. This enables more consistent subjects, distinctive emotional expression, and realistic object-level transformations.

3 Method

3.1 Data Construction

Existing datasets primarily focus on emotion labels or emotion attributes, but remain limited in their semantic descriptions of visual content. A key requirement for object-based emotion control is to characterize how the same object changes in state or appearance under different emotions. Such paired descriptions are essential for fine-grained emotion control and consistency modeling of a fixed subject. To this end, we build on EmoSet, a large-scale emotion dataset, and construct a dual-description format centered on the same visual subject. This design enables the model to learn the correspondence between emotional semantics and object-specific visual states while preserving object identity and core semantic content, thereby providing more effective supervision for controllable emotion generation of the same object. As a result, the semantic difference between *Base Content* and *Emotion Content* is mainly concentrated on emotion-related states and visual attributes, providing a more explicit training signal for learning the correspondence between target emotions and object-level visual changes. After filtering, the resulting training candidate set contains 112,197 samples. The objects corresponding to each emotion in Obj-EmoSet are shown in Fig. 2. Further details are available in the supplementary material.

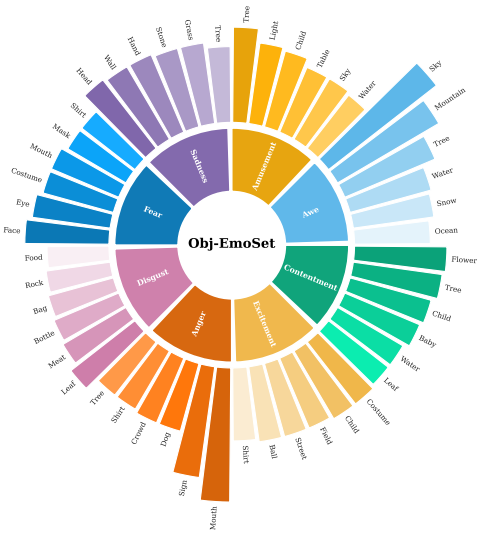


Figure 2: Top-6 Object Distributions Across Eight Emotion Categories in Obj-EmoSet.

3.2 Same Object, Different Emotions

Emotion-Aware Visual Planner Based on our customized dataset, Obj-EmoSet, we propose an object-centric controllable content generation framework (Same Object, Different Emotions). Emotion is inherently complex and abstract. Inspired by prior studies on visual emotion understand-

ing Ding et al. (2026), we introduce an Emotion-Aware Visual Planner (Plan), which explicitly decomposes abstract emotional concepts into structured multi-slot states, thereby narrowing the affective gap between high-level emotional semantics and low-level visual representations, as shown in Fig. 3. Specifically, the module first converts the target emotion into object-related visual semantic states and further organizes them into an executable visual plan. We characterize the target emotion from four complementary dimensions: subject or scene state, action or spatial arrangement, composition and lighting, and atmosphere and style. Given a target emotion e and a base description s, Plan predicts a structured visual plan: $\mathrm{p} = \mathrm{f}_{\mathrm{plan}}(e, s; \theta_P)$, (0) where

$$p = \left\{ p^{\mathrm{state}}, p^{\mathrm{action}}, p^{\mathrm{comp}}, p^{\mathrm{style}} \right\}. \tag{1}$$

Specifically, p^{state} describes observable subject- or scene-level states, including facial expression, pose, shape, and local appearance changes; p^{action} describes subject actions, object interactions, and spatial arrangements; p^{comp} specifies composition, viewpoint, and lighting conditions; and p^{style} complements these attributes with global atmosphere and visual style. When reliable evidence for a certain type of visual information is unavailable, the corresponding slot is assigned `N/A`, preventing the model from introducing unnecessary or implausible attributes merely to complete the predefined structure. To represent target emotions as stable control variables, we introduce a dedicated emotion token for each of the eight emotion categories(T). Let the token sequence of the target visual plan be

$$y^P = \left(y_1^P, \ldots, y_T^P \right). \tag{2}$$

Plan maps abstract emotions to structured visual states. We therefore introduce a slot-based planning loss, $\mathcal{L}_{\mathrm{plan}}$, to supervise visual plan:

$$\mathcal{L}_{\mathrm{plan}} = -\frac{1}{T} \sum_{t=1}^{T} \log P_{\theta_P} \left(y_t^P \mid y_{<t}^P, e, s \right). \tag{3}$$

This objective directly learns the mapping from the target emotion and stable content condition to structured visual states.

Visual Realism Refiner Existing text-based emotion-driven models, however, often produce emotionally expressive results at the cost of visual realism, leading to outputs that appear unrealistic or overly abstract. This issue partly stems from the inherent abstraction and ambiguity of textual emotion descriptions. To alleviate this problem, we further propose a Visual Realism Refiner (Vis), which reorganizes and constrains the emotional conditions prior to emotion-driven content generation, enabling abstract emotional semantics to be translated into more explicit and visually plausible generation conditions.

Through this process, the structured visual plan is transformed into a natural-language condition that better preserves the original content semantics while aligning more closely with real-world visual characteristics. Given the target emotion e, the base description s, and the visual plan p generated by Plan, Vis produces a refined text condition:

$$c = f_{\mathrm{vis}}(e, s, p; \theta_R), \tag{4}$$

where c denotes a complete image-generation prompt that can be directly fed into the diffusion model. This design decouples visual-state reasoning from natural-language realization, allowing Plan to determine what visual attributes should change, while Vis determines how these changes should be organized into realistic and executable text conditions. Let the token sequence of the refined prompt be

$$y^R = \left(y_1^R, \ldots, y_M^R \right). \tag{5}$$

We introduce the Vis loss to supervise Vis in reconstructing the visual structure of the textual description.

$$\mathcal{L}_{\mathrm{vis}} = -\frac{1}{M} \sum_{t=1}^{M} \log P_{\theta_R} \left(y_t^R \mid y_{<t}^R, e, s, p \right). \tag{6}$$

Compared with directly using emotional text for generation, this process effectively mitigates the excessive interference of emotional semantics with content semantics, enabling the model to preserve the consistency of subject appearance, spatial layout, and scene details while maintaining the intended emotional atmosphere. Ultimately, the Vis-refined representation achieves a better balance between emotional expressiveness and visual realism.

Emotion Calibration Adapter To generate images with both strong emotional expressiveness and high visual realism, we propose Emotion Calibration Adapter (Cali), which enables the generated images to convey emotions in a more natural and visually coherent manner.

Given the text condition c generated by Vis and a random noise variable z, the image generation process is formulated as

$$I = f_{\text{Cali}}\left(c, z; \theta_G, \Delta\theta_L\right), \tag{7}$$

where θ_G denotes the frozen parameters of the FLUX.2 base model, and $\Delta\theta_L$ represents the trainable LoRA parameter updates.

Rather than fine-tuning the entire generator, we freeze the base Transformer and text encoder, and adapt only the low-rank parameters to modulate the response of the Transformer to emotion-related textual conditions. This adaptation enhances the model's ability to capture emotion-relevant visual distributions, including subject states, composition, illumination, color, and overall atmosphere, thereby producing natural images with a balanced expression of emotional content.

During training, the input image is first encoded into a latent representation x using the image encoder. Given a randomly sampled timestep t and Gaussian noise ϵ, an intermediate noisy latent state is constructed as

$$x_t = (1 - \sigma_t)x + \sigma_t\epsilon, \qquad \epsilon \sim \mathcal{N}(0, I). \tag{8}$$

Conditioned on the noisy latent x_t, timestep t, and text condition c, the FLUX.2 Transformer predicts the conditional velocity field. The target velocity is defined as

$$v^* = \epsilon - x. \tag{9}$$

Accordingly, the adaptation objective is formulated as

$$\mathcal{L}_{\text{Cali}} = \mathbb{E}_{x,c,t,\epsilon}\left[\left\|v_{\theta_G,\Delta\theta_L}(x_t, t, c) - (\epsilon - x)\right\|_2^2\right]. \tag{10}$$

4 Experiment

4.1 Evaluation

Evaluation Metrics We evaluate generated images in terms of emotion alignment, semantic consistency, and generation quality. For emotion alignment, we use Emotion Accuracy (Emo-A) Yang et al. (2024) and Emotion–Content Joint Accuracy (EC-A) to assess emotion fidelity and joint emotion-content alignment. For semantic consistency, Subject Similarity (Subj.) measures subject preservation, DINOv2 similarity evaluates visual consistency Oquab et al. (2023), MLLM measures high-level semantic alignment Wu et al. (2024), and Semantic Clarity (Sem-C) evaluates semantic recognizability Yang et al. (2024). For generation quality, MLLM-P measures perceptual quality Varga (2025), while Style evaluates style consistency Somepalli et al. (2024).

4.2 Comparisons

We compare our method with representative and competitive methods from four categories: (1) Foundation text-to-image models: Stable Diffusion 1.5 (SD1.5) Rombach et al. (2022b), SDXL Base 1.0 Podell et al. (2024), PixArt-α Chen et al. (2024a), FLUX.1 Black Forest Labs (2024), and FLUX.2 Black Forest Labs (2025); (2) Personalized generation methods: Textual Inversion (TI) Gal et al. (2023); (3) LLM- and multimodal-enhanced generation methods: LLM4GEN Liu et al. (2025) and OmniGen2 Wu et al. (2025); and (4) Emotional image generation methods: EmoGen Yang et al. (2024) and EmotiCrafter Dang et al. (2025).

Image-level Evaluation As shown in Tables 1, our method achieves strong overall performance in emotion expression, semantic consistency, and image realism. It attains the best Emo-A and EC-A scores, demonstrating superior alignment with the target emotion. For semantic consistency, our method ranks second on Subj., MLLM, and DINOv2, indicating that the enhanced emotional expression is achieved while maintaining strong semantic fidelity. For image realism, our method achieves the best performance on both MLLM-P and Style, suggesting higher perceptual quality and reduced over-stylization, and thus producing more natural and realistic images.

User study We further conduct a user study to evaluate human preferences for different emotion-aware image generation methods. A total of 48 participants with diverse ages and backgrounds were

Table 1: Image-level comparison across emotion alignment, semantic consistency, and generation quality. Bold indicates the best result, and underline indicates the second-best result.

Method	Emotion Alignment		Semantic Consistency				Generation Quality	
	Emo-A ↑	EC-A ↑	Subj. ↑	MLLM ↑	Sem-C ↑	DINOv2 ↑	MLLM-P ↑	Style ↓
Stable Diffusion 1.5	23.48	22.97	0.5437	0.2744	0.4876	0.3680	0.6564	36.35
SDXL Base 1.0	30.70	29.75	0.5832	0.4001	0.5038	0.4228	0.4661	63.80
FLUX.1	33.84	33.18	0.6400	0.5311	**0.5381**	0.5023	0.6578	39.80
FLUX.2	37.51	36.22	0.6227	0.5452	0.5299	0.5141	0.7466	26.47
Textual Inversion	39.61	27.45	0.4868	0.1701	0.4655	0.3057	0.6008	48.36
PixArt-alpha	41.24	39.33	**0.7287**	**0.6354**	0.4911	**0.6236**	0.3761	85.79
OmniGen2	38.22	35.95	0.6194	0.5306	0.5165	0.4720	0.5932	48.98
LLM4GEN	28.70	28.08	0.7039	0.4892	0.4600	0.5686	0.2899	99.60
EmoGen	21.11	19.31	0.4850	0.1804	0.4837	0.3452	0.5093	58.86
EmotiCrafter	22.92	15.68	0.6388	0.1512	0.4430	0.5000	0.3242	96.34
Ours	**53.61**	**49.50**	0.7207	0.5501	0.5283	0.6224	**0.7529**	**25.93**

Table 2: User preference study. Values indicate the percentage (%) of participants who preferred the result under each evaluation criterion.

Method	Emotion Evoking ↑	Object-Emotion Consistency ↑	Content Realism ↑
PixArt-alpha	35.69	33.33	14.51
FLUX.2	9.41	7.06	18.82
EmoGen	3.92	4.71	11.37
Ours	**50.98**	**54.90**	**55.29**

invited, with each session lasting approximately 20 minutes. The study consists of 25 emotion–content condition pairs, each associated with images generated by four representative emotion-aware image generation methods, including EmoGen, EmoCtrl, and two other competing approaches. Participants evaluated the generated images along three dimensions: **Emotion Evoking**, **Object–Emotion Consistency**, and **Content Realism**.

As shown in Table 2, our method achieves the highest user preference across all three evaluation dimensions. Specifically, our method obtains a preference rate of 50.98% in **Emotion Evoking**. For **Object–Emotion Consistency**, our method achieves the highest preference rate of 54.90%, showing a clear advantage over PixArt-alpha. In addition, our method ranks first in **Content Realism**, with a preference rate of 55.29%. These results demonstrate that our method can more accurately convey the target emotion while maintaining high content realism and better consistency between objects and emotions.

Object-level Evaluation To evaluate object-centric emotion control, we first use Segment Anything to segment Kirillov et al. (2023) the target objects and leverage GroundingDINO-based Liu et al. (2024) object detection together with segmentation confidence scores to retain reliable object regions. We then employ Emo-A and EC-A to quantify the alignment between the generated results and the target emotions, while Subject Cons. and DINOv2 are used to assess subject consistency and visual representation consistency of the target objects before and after emotion manipulation.

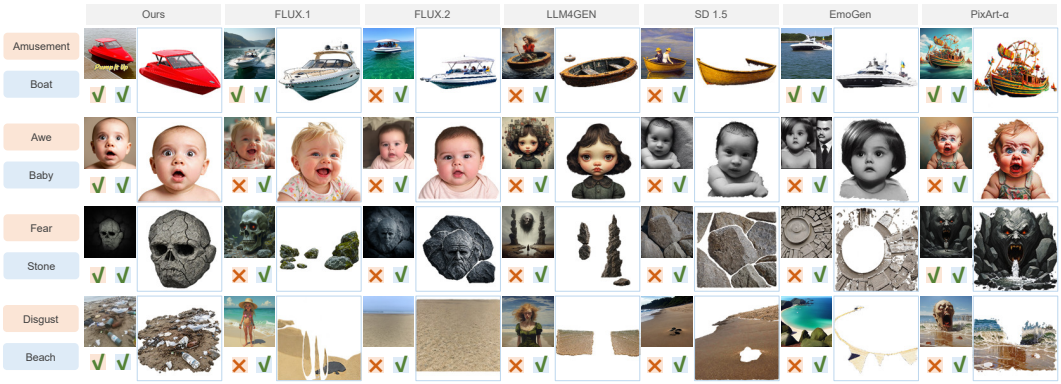


Figure 4: Qualitative comparison of object-level emotion-controllable generation. Visualization of the segmented target objects and their controllability under different target emotions.

Table 3: Object-level comparison across emotion alignment, and semantic consistency.

Method	Emo-A ↑	EC-A ↑	Subject Cons. ↑	DINOv2 ↑
SD 1.5	21.75	21.46	0.5596	0.3501
FLUX.1	24.61	24.17	0.6334	0.4585
FLUX.2	27.48	26.97	0.6434	0.4684
LLM4GEN	22.36	21.77	0.6441	0.4851
PixArt-alpha	30.64	27.76	0.6822	**0.5444**
EmoGen	16.15	14.23	0.5094	0.3129
Ours	**47.36**	**44.58**	**0.6887**	0.5330

Table 4: Ablation study of model configurations. Bold indicates the best result, and underline indicates the second-best result.

Method	Emotion Alignment		Semantic Consistency				Generation Quality	
	Emo-A ↑	EC-A ↑	Subj. ↑	MLLM ↑	Sem-C ↑	DINOv2 ↑	MLLM-P ↑	Style ↓
Baseline	29.57	28.62	0.6178	0.3086	0.5357	0.5331	0.6133	40.43
Plan	**53.10**	**50.78**	**0.7682**	**0.4747**	0.5444	**0.6827**	0.7610	24.83
Vis	17.72	17.03	0.5447	0.1206	**0.5677**	0.3800	**0.8563**	**10.60**
Plan+Vis	53.06	50.17	0.7270	0.4705	0.5396	0.6201	0.7785	22.50
Plan+Vis+Cali	**53.10**	50.47	0.7289	0.4661	0.5348	0.6256	0.7432	27.11

As shown in Fig. 4, the qualitative results show that our method effectively controls object-level emotional expression while maintaining background tones consistent with the target emotion, leading to more natural and realistic outputs. We further evaluate the segmented subject regions in terms of emotion alignment and content preservation, as shown in Table 3. Our method significantly outperforms the competing methods on both Emo-A and EC-A, demonstrating stronger object-level emotional controllability. Meanwhile, the content-preservation metrics remain high, indicating that emotional cues do not substantially alter the subject's original semantic and visual characteristics. Overall, our method achieves accurate emotional control while preserving subject consistency and visual realism.

4.3 Ablation Study

We conduct ablation studies to evaluate the contributions of the three proposed modules: the Emotion-Aware Visual Planner (Plan), the Visual Realism Refiner (Vis), and the Emotion Calibration Adapter (Cali). As shown in Fig. 5 and Table 4, the FLUX.2 baseline exhibits limited object-level emotional expressiveness and relatively modest overall visual quality. Plan primarily enhances emotion-related semantic content through slot-based augmentation, enabling the model to better capture and represent affective concepts in the generated images. Quantitatively, introducing Plan yields clear improvements in Emo-A and EC-A, while also improving Subj., MLLM, and DINOv2. These results demonstrate that Plan more effectively aligns emotional semantics with visual content. In contrast, Vis mainly improves the visual structure, fine-grained details, and overall composition of the generated images, thereby en-

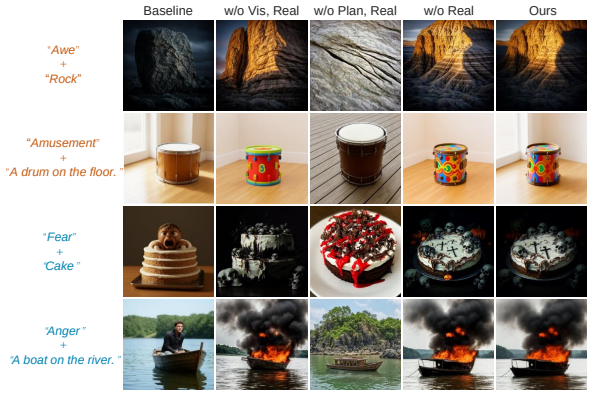


Figure 5: Ablation study on Plan, Vis, and Real.



Figure 6: Generation results conditioned on the same object category and different emotions.

hancing perceptual quality and realism. This effect is also reflected in its strong performance on MLLM-P and Style. Building on these improvements, the further incorporation of Cali provides a better balance between emotional expression and visual realism, resulting in more natural-looking generated images.

4.4 Qualitative Experiment

Emotion-guided Object Generation. As shown in Fig. 6, our method not only enhances the overall emotional expression by modulating color, illumination, and scene atmosphere, but also effectively controls the visual state of the object according to the target emotion, achieving a favorable balance between emotional expressiveness and visual realism. Negative emotions are naturally conveyed through cues such as darker lighting, fire, and degraded environments, whereas positive emotions are expressed through brighter colors and more vivid scenes. The results demonstrate that our method achieves a coherent balance among emotion alignment, object-level control, and image realism.

Object-Level Emotion-Guided Stylization. We evaluate the joint controllability of artistic style and object-level emotion across multiple styles, including Matisse, Picasso, Watercolor, and Oil Painting. As shown in Fig. 7, our method preserves the structure and semantic content of the target object while adapting the overall visual appearance to the specified artistic style and modulating the object's color, texture, and surrounding atmosphere according to the target emotion. For the same target emotion, different artistic styles produce distinct style-specific emotional expressions, demonstrating that our method can jointly control artistic style and object-level emotion without substantially compromising object identity or semantic consistency. Overall, our method effectively coordinates stylistic variation and object-level emotional expression, enabling joint controllable generation with strong content preservation.

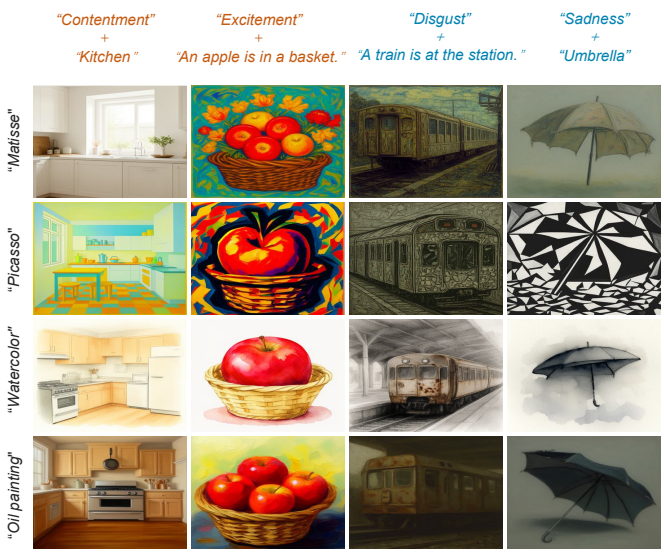


Figure 7: Object-Centric Emotion Control under Diverse Artistic Styles.

5 Conclusion

To address the limitations of existing emotion-aware image generation methods in object-level emotion control, semantic preservation, and visual realism, we propose an object-centric emotion-controllable generation framework that enables joint emotion control at both the image and object levels. The framework consists of three functional Emotion-Aware Visual Planner, Visual Realism Refiner and an Emotion Calibration Adapter—which translate abstract emotions into object-related visual states and further refine composition, illumination, texture, and overall atmosphere to improve the naturalness and realism of emotional expression. Extensive qualitative and quantitative experiments demonstrate that the proposed method achieves more accurate target emotion control while preserving subject identity, semantic content, and visual structure. Compared with existing methods, our approach not only produces more distinctive emotion-aware representations around target objects, but also reduces excessive stylization, resulting in more natural variations in color, lighting, scene atmosphere, and object state that are better aligned with real-world emotional imagery. Overall, the proposed framework achieves a better balance among emotion controllability, semantic consistency, and image realism, providing an effective solution for fine-grained object-centric emotional image generation.

REFERENCES

Omri Avrahami, Ohad Fried, and Dani Lischinski. Blended latent diffusion. *ACM Transactions on Graphics*, 42(4):149:1–149:11, 2023. doi: 10.1145/3592450.

Bita Azari and Angelica Lim. Emostyle: One-shot facial expression editing using continuous emotion parameters. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 6385–6394, 2024.

Black Forest Labs. FLUX. `https://github.com/black-forest-labs/flux`, 2024. Software/model resource.

Black Forest Labs. FLUX.2: Frontier visual intelligence. `https://bfl.ai/blog/flux-2`, 2025. Official model release/blog resource.

Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18392–18402, 2023.

Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. In *International Conference on Learning Representations*, 2024a.

Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. In *International Conference on Learning Representations*, 2024b.

Shengqi Dang, Yi He, Long Ling, Ziqing Qian, Nanxuan Zhao, and Nan Cao. EmotiCrafter: Text-to-emotional-image generation based on valence-arousal model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15218–15228, 2025.

Tingting Ding, Jingyuan Yang, and Hui Huang. AGER: Attribute-guided emotion reasoning for visual emotion analysis. *IEEE Transactions on Affective Computing*, 2026. doi: 10.1109/TAFFC.2026.3705363.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *International Conference on Learning Representations*, 2023.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *2023 IEEE/CVF international conference on computer vision (ICCV)*, pp. 3992–4003. IEEE, 2023.

Michał Kuniecki, Kinga B. Wołoszyn, Aleksandra Domagalik, and Joanna Pilarczyk. Effects of scene properties and emotional valence on brain activations: A fixation-related fMRI study. *Frontiers in Human Neuroscience*, 11:429, 2017. doi: 10.3389/fnhum.2017.00429.

Qing Lin, Jingfeng Zhang, Yew-Soon Ong, and Mengmi Zhang. Make me happier: Evoking emotions through image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16367–16376, 2025.

Mushui Liu, Yuhang Ma, Zhen Yang, Jun Dan, Yunlong Yu, Zeng Zhao, Zhipeng Hu, Bai Liu, and Changjie Fan. LLM4GEN: Leveraging semantic representation of LLMs for text-to-image generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 5523–5531, 2025. doi: 10.1609/aaai.v39i5.32588.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, pp. 38–55. Springer, 2024.

Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023.

Chanjong Park and In-Kwon Lee. Emotional landscape image generation using generative adversarial networks. In *Proceedings of the Asian Conference on Computer Vision*, 2020.

Joanna Pilarczyk and Michał Kuniecki. Emotional content of an image attracts attention more than visually salient features in various signal-to-noise ratio conditions. *Journal of Vision*, 14(12):4, 2014. doi: 10.1167/14.12.4.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *International Conference on Learning Representations*, 2024.

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

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

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22500–22510, 2023.

Gowthami Somepalli, Anubhav Gupta, Kamal Gupta, Shramay Palta, Micah Goldblum, Jonas Geiping, Abhinav Shrivastava, and Tom Goldstein. Measuring style similarity in diffusion models. *arXiv preprint arXiv:2404.01292*, 2024.

Domonkos Varga. Comparative evaluation of multimodal large language models for no-reference image quality assessment with authentic distortions: A study of openai and claude.ai models. *Big Data and Cognitive Computing*, 9(5):132, 2025.

Chenyuan Wu, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, Junjie Zhou, Ze Liu, Ziyi Xia, Chaofan Li, Haoge Deng, Jiahao Wang, Kun Luo, Bo Zhang, Defu Lian, Xinlong Wang, Zhongyuan Wang, Tiejun Huang, and Zheng Liu. OmniGen2: Exploration to advanced multimodal generation. *arXiv preprint arXiv:2506.18871*, 2025.

Xun Wu, Shaohan Huang, and Furu Wei. Multimodal large language model is a human-aligned annotator for text-to-image generation. *arXiv preprint arXiv:2404.15100*, 2024.

Chao Xu, Junwei Zhu, Jiangning Zhang, Yue Han, Wenqing Chu, Ying Tai, Chengjie Wang, Zhifeng Xie, and Yong Liu. High-fidelity generalized emotional talking face generation with multi-modal emotion space learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6609–6619, 2023.

Jingyuan Yang, Qirui Huang, Tingting Ding, Dani Lischinski, Daniel Cohen-Or, and Hui Huang. EmoSet: A large-scale visual emotion dataset with rich attributes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20383–20394, 2023.

Jingyuan Yang, Jiawei Feng, and Hui Huang. EmoGen: Emotional image content generation with text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6358–6368, 2024.

Jingyuan Yang, Jiawei Feng, Weibin Luo, Dani Lischinski, Daniel Cohen-Or, and Hui Huang. EmoEdit: Evoking emotions through image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24690–24699, 2025.

Jingyuan Yang, Weibin Luo, and Hui Huang. EmoCtrl: Controllable emotional image content generation. In *Proceedings of the ACM International Conference on Multimedia*, 2026. Accepted/published metadata should be updated with final pages and DOI when available.

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.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3836–3847, 2023.

Yijie Zhu, Lingsen Zhang, Zitong Yu, Rui Shao, Tao Tan, and Liqiang Nie. UniEmo: Unifying emotional understanding and generation with learnable expert queries. *IEEE Transactions on Image Processing*, 35:5165–5180, 2026. doi: 10.1109/TIP.2026.3691017.