Moodifier: MLLM-Enhanced Emotion-Driven Image Editing

Jiarong Ye, and Sharon X. Huang, *Member, IEEE*

***Abstract*—Emotion-driven image editing is valuable for creative workflows, but precise control is difficult because the same emotion can correspond to different visual cues across domains (e.g., facial expressions, fashion, jewelry, and home décor). We present an integrated pipeline with three components. (1) MoodArchive: a dataset of 8M+ images with hierarchical emotion annotations generated by LLaVA-NeXT and partially validated by human evaluators. (2) MoodifyCLIP: a vision-language model fine-tuned on MoodArchive to map abstract emotions to context-aware visual attributes. (3) Moodifier: a training-free editing system (no additional finetuning for editing) that combines MoodifyCLIP with multimodal large language models (MLLMs) to generate where-to-edit / how-to-edit guidance, and applies attention-controlled diffusion to perform targeted emotional transformations while preserving identity and structure. Experiments on emotion understanding (zero-shot classification and retrieval) and emotion editing (automatic metrics and human preference study) show improved emotional fidelity with competitive content preservation compared to prior editing baselines. A license-filtered subset of MoodArchive, the MoodifyCLIP model, and a demo of the Moodifier image editor are released on our project page.[1] We will release the complete redistributable subset, the permitted metadata and annotation index, and the Moodifier source code upon paper acceptance.**

***Index Terms*—Affective Computing, Emotion-driven Image Editing, Multi-modal Diffusion Models**

I. Introduction

Imagine transforming a cocktail dress from *elegant* to *joyful* with a single click, or previewing how subtle material changes make a pendant feel more *romantic* or *exciting*. As illustrated in Fig. 1, precise emotion-driven editing can streamline creative iteration in design and content production[2]. Beyond aesthetic exploration, such controls are valuable for practical workflows, including character design in animation, emotion-aware product presentation in e-commerce, and rapid asset repurposing in marketing.

Despite this potential, current image editors provide limited *emotion-level* control. The core difficulty is that emotions are abstract and highly context-dependent: the same target emotion (e.g., "excitement") may correspond to different visual attributes across domains, such as color/pattern changes in fashion, facial muscle cues in portraits, or material/texture highlights in jewelry. As a result, existing text-guided editors (e.g., [1]–[3]) often lack explicit emotion grounding and region-level control, leading to either overly generic adjustments or unstable edits that degrade structure and identity, which forces time-consuming trial-and-error.

Hence we propose **Moodifier**, a two-stage framework that bridges abstract emotions to precise, structure-preserving visual edits. *Where to edit:* given a target emotion, an MLLM analyzes the source image and generates context-specific editing instructions together with spatial attention maps that localize emotion-relevant regions. *How to edit:* to convert abstract emotional intent into concrete, context-aware visual attributes, we introduce **MoodifyCLIP**, a vision-language model fine-tuned for emotional understanding. Unlike standard CLIP models trained on short, factual captions (e.g., COCO [4] and Flickr [5]), MoodifyCLIP leverages **MoodArchive** (Fig. 2), which provides large-scale affective images with hierarchical emotional annotations (Fig. 3).

Conditioned on MoodifyCLIP-processed prompts and the MLLM attention maps, Moodifier performs attention-controlled diffusion editing: it inverts the source image into latent space and modifies only emotion-relevant attributes while preserving the remaining content. As shown in Fig. 1, Moodifier generalizes across domains, preserving character identity, maintaining garment structure, and keeping object geometry stable while producing contextually appropriate emotional variations.

To summarize, as illustrated in Fig. 5 (3), we present an end-to-end pipeline for emotion-driven image editing with three components:

1) MoodArchive. A dataset of 8M+ images with hierarchical emotional annotations generated by LLaVA-NeXT [6] and partially validated by human evaluators. MoodArchive provides large-scale, structured supervision for learning emotion-relevant visual cues beyond coarse emotion labels.

2) MoodifyCLIP. A vision–language model fine-tuned on MoodArchive to map abstract emotions to context-aware visual attributes, enabling emotion-aligned representation learning for downstream retrieval and editing.

3) Moodifier. A training-free editing system that combines MoodifyCLIP, MLLMs, and diffusion-based attention control to perform targeted emotional transformations while preserving structure and identity.

II. Related Works

Visual Emotion Datasets. Early affective computing work often adopts discrete emotion categories (e.g., Ekman [7])

J. Ye and S.X. Huang are with the College of Information Sciences and Technology, Pennsylvania State University, University Park, PA 16802, USA. E-mails are: (jxy225@psu.edu; suh972@psu.edu).

[1]Project site: https://emomoodify.github.io/app/. Only images whose applicable licenses and source terms permit redistribution are included as image files. Source-restricted items are excluded from direct image redistribution and are represented by source URLs and derived annotations only where the applicable terms permit. Derived annotations are released under CC BY 4.0.

[2]A demo video of Moodifier in action is available on the project site.

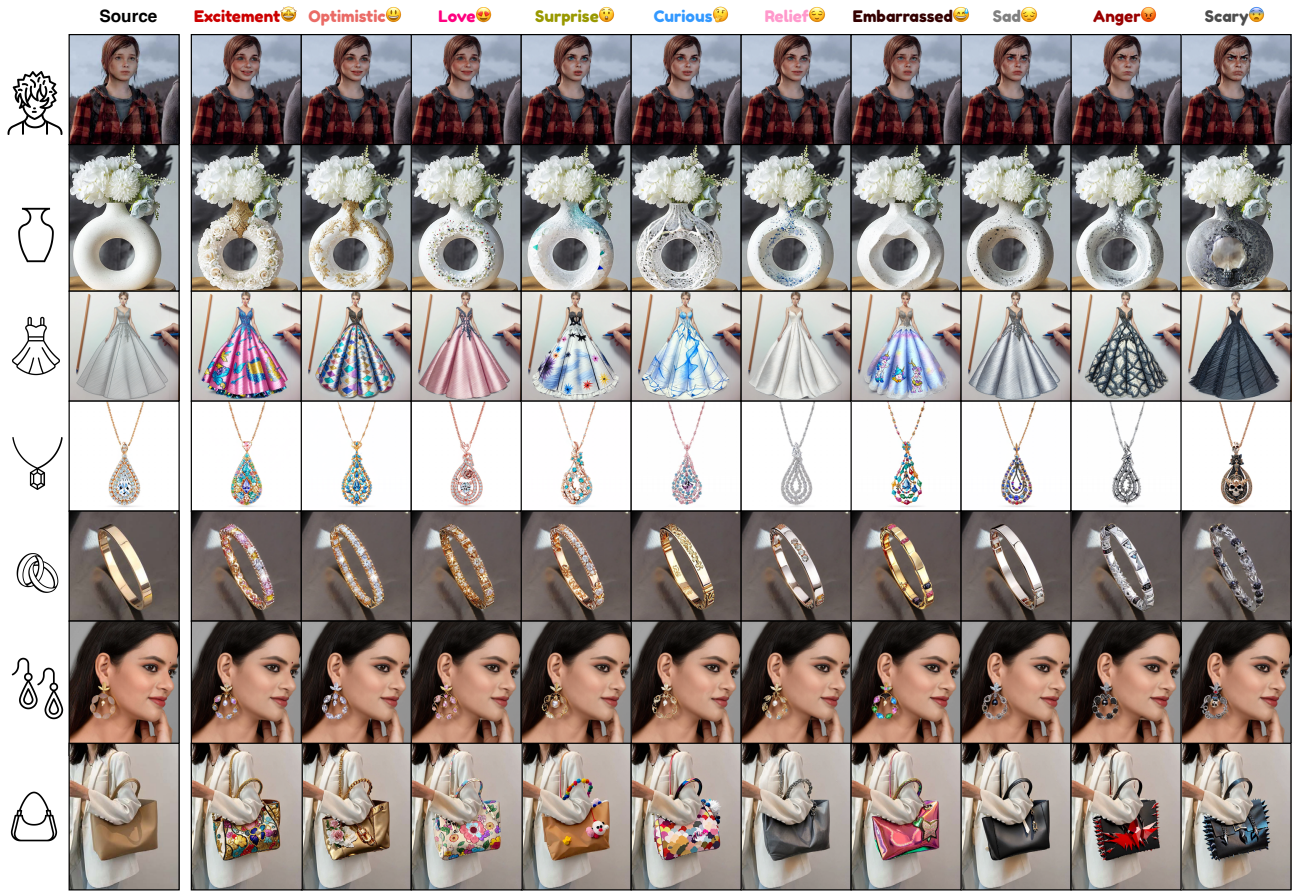


Fig. 1. Our emotion-driven Moodifier image editing system performs consistently across diverse scenarios such as product mockups, e-commerce, fashion, jewelry, and home décor. It precisely modifies target objects while preserving the overall structure and integrity of the scene. The target objects selected for editing are depicted as clip arts in the leftmost column. In non-facial contexts, emotions manifest as semantic qualities (e.g., "Excitement" as vibrant designs, "Sad" as drooping elements, "Anger" as bold styling). Zoom in for better view.

and studies visual emotion perception through curated image datasets [8]–[12]. While influential, existing datasets are typically limited in either emotion coverage, scale, or contextual diversity, which constrains learning fine-grained, context-dependent affective cues needed for generation and editing.

Contrastive Language-Image Pre-training (CLIP) and Synthetic Supervision. CLIP [13] enables strong zero-shot transfer across many vision tasks [14]–[23]. However, standard CLIP training is limited in fine-grained region-token grounding [24], [25], and alignment becomes more challenging when captions are long, compositional, or contain multiple affective triggers. Synthetic data pipelines have improved vision language pre-training by generating higher-quality pairs [26]–[29], and multimodal LLMs have further enhanced caption generation [30], [31]. Yet these efforts predominantly optimize short, descriptive captions and rarely provide structured supervision that explicitly connects emotions to their visual causes. Recent vision–language studies have further specialized CLIP-style representations for facial analysis and emotion recognition. FLIP-80M [32] introduces a large-scale face–text corpus and uses it to train FLIP, a facial language–image representation model. Exp-CLIP [33] transfers LLM-derived facial-action knowledge into a zero-shot facial expression representation space through a task-specific projection head. EmoCapCLIP [34] learns global and local facial-emotion representations from semantically rich captions. These studies demonstrate the value of language supervision for facial-affect representation. MoodifyCLIP differs in scope by learning affective representations across facial expressions, natural scenes, urban scenes, and objects, and by connecting these representations to a downstream region-controlled image-editing framework.

Image Editing with Attention Control. Text-guided editing has progressed from GAN-based methods [35], [36] to diffusion models [22], [37]–[39], where cross-attention enables controllable generation and editing [40]–[43]. Mask-based pipelines [44]–[48] offer locality but require accurate region selection, while attention-control approaches such as Prompt-to-Prompt [41], Plug-and-Play [49], and MasaCtrl [50] provide zero-shot structural preservation via attention manipulation. Instruction-tuned editors [2], [3], [51] improve usability. More recently, UltraEdit [52] introduced large-scale instruction-based editing supervision and an SD3-based editor supporting both free-form and region-based editing. Nevertheless, complex emotion-driven edits remain challenging due to (i) the abstract and context-dependent nature of emotions and (ii) the lack of explicit guidance on *where* and *how* emotion-specific changes should be applied.

III. Methodology

In this section, we present details about the three major contributions of our framework: MoodArchive, MoodifyCLIP, and Moodifier.

A. MoodArchive Curation

Existing language-image pre-training datasets provide limited supervision for affective understanding: captions are often short and factual, and emotion labels are either coarse, low-coverage, or lack context. As a result, models struggle to learn *what visual cues trigger which emotions* across diverse scenarios. To address this gap, we curate *MoodArchive*, a large-scale affective dataset with over 8 million images and structured emotion-relevant captions. Our curation pipeline is summarized in Fig. 2A and consists of four steps.

Emotion taxonomy (27 emotions) across 4 contexts. To improve both coverage and granularity compared to existing datasets [8]–[12], we adopt the GoEmotions label set [53] (27 emotions[3], see Fig. 2C). For each emotion, we collect images in four distinct contexts: *facial expressions*, *natural scenery*, *urban scenery*, and *object classes*. This context factorization is intentional: the same emotion can manifest differently across domains (e.g., facial cues vs. materials/textures vs. scene lighting), and models must learn these context-dependent triggers. As summarized in Table I, MoodArchive expands both scale and contextual coverage. Its 8,417,937 images are approximately 2.6 times the 3.3 million images in the complete EmoSet [10] collection. Moreover, each of its 27 retrieval emotions is systematically collected across four visual contexts (facial expression, natural scenery, urban scenery, and object class), forming 108 emotion–context combinations.

Importantly, the emotion associated with each retrieved image is treated as a weak retrieval label rather than an exclusive ground-truth emotional state. Because ChatGPT expands one seed emotion at a time, this label records why an image was retrieved rather than every emotion it may evoke. The retrieval emotion is used to construct search queries and organize the collected images, but is not used as a hard categorical target for training MoodifyCLIP. MoodifyCLIP is instead trained using image–caption pairs generated by LLaVA-NeXT. Because each caption is generated from the image content and is not required to repeat the retrieval label, it may describe alternative or multiple plausible interpretations. For example, a road closure may evoke annoyance, while the same situation at night may additionally evoke fear or nervousness, a thunderstorm may evoke fear, excitement, or surprise. We therefore do not assume that the 27 retrieval categories are mutually exclusive.

Emotion context expansion. Directly searching with abstract emotion words (e.g., "anger") yields noisy and generic results. We therefore use ChatGPT to decompose each emotion into concrete, context-specific descriptors and scenarios that can be used as effective retrieval queries (Fig. 2A). For example, "anger" is expanded into relatively direct facial cues such as "pursed lips" and "glaring eyes", as well as candidate scene-level queries such as "fierce wildfire". This association is motivated by prior visual-emotion studies showing that scene type, object content, color, and texture can contribute to affective perception [8], [10]. We do not treat wildfire as a validated or exclusive visual definition of anger, because the same scene may also evoke fear or excitement.

Using these refined phrases as search queries, we retrieved candidate images from nine online image platforms: Unsplash, Pexels, Pixabay, Envato Elements, Openverse, Burst by Shopify, StockSnap, Foodiesfeed, and Free Nature Stock.[4] These platforms served as image-retrieval sources. MoodArchive was not constructed by merging the existing affective benchmark datasets compared in Table I. We do not claim formal GDPR certification. Some source images may depict identifiable individuals. MoodArchive stores no names or identity labels, is not intended for identification or biometric verification, redistributes only license-permitted images, and project site provides a contact for removal requests.

Post-processing and annotation. We apply automatic filtering to remove NSFW content and corrupted files. We then generate structured captions for each image using LLaVA-NeXT [6]. It is selected because its open weights support reproducible local inference and its high-resolution image processing is suitable for detailed annotation. We used a single MLLM to maintain annotation consistency, combining multiple MLLMs at this scale would substantially increase computation and require an additional aggregation rule. The captioning stage is modular, so another MLLM following the same structured output format could be substituted without changing the downstream dataset format. Each caption follows a fixed hierarchy (Fig. 3): (1) a concise global summary, (2) three local emotional stimuli that explicitly name visual triggers, and (3) an overall emotion assessment (see Appendix for the instruction prompt). Finally, we remove the bottom 20% of images per emotion based on CLIP similarity to emotion-related phrases to reduce off-topic retrieval noise. After these filtering stages, the combined retained collection contains 8,417,937 image–caption pairs. This final retained count is the "over 8 million images" reported throughout the paper.

Human validation study. To assess comparative caption quality at scale, we conduct a human validation study on Amazon Mechanical Turk. We randomly sample 20,000 images and ask workers to compare the original web-collected alt-text with the LLaVA-generated structured caption using the interface shown in Appendix Fig. 1. A caption is considered inaccurate if any component of its structured annotation (Fig. 3) is inconsistent with the image.

Each image was independently evaluated by three workers, yielding 60,000 worker–image judgments. Participation was open to the MTurk worker pool without additional geographic, approval-rate, or previously approved-HIT qualifications, and

[3] admiration, amusement, approval, caring, desire, excitement, gratitude, joy, love, optimism, pride, relief, anger, annoyance, disappointment, disapproval, disgust, embarrassment, fear, nervousness, grief, remorse, sadness, confusion, curiosity, realization, surprise

[4] Unsplash, Pexels, Pixabay, Envato Elements, Openverse, Burst by Shopify, StockSnap, Foodiesfeed, and Free Nature Stock.

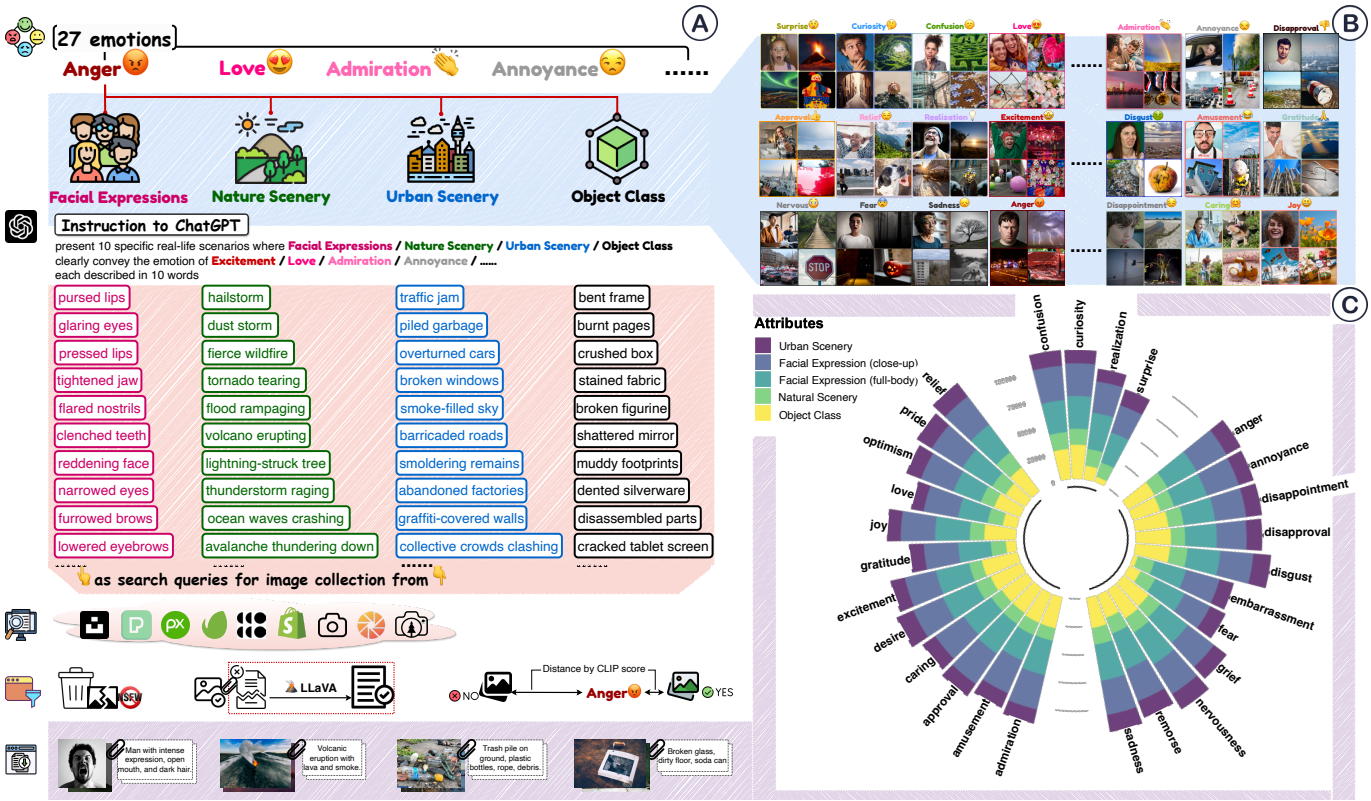


Fig. 2. **A. *MoodArchive* data collection workflow.** The dataset includes images retrieved using 27 emotion tags, each with 4 contexts: facial expression, nature scenery, urban scenery, and object class. These attributes are further detailed with fine-grained attributes to enrich the emotional context with help from ChatGPT. These scenarios were then used as search queries to find relevant images, which were filtered and saved. Each data unit in the dataset consists of an image paired with a brief text description of its content; **B. *MoodArchive* visual data examples. C. Dataset overview and basic statistics bar chart.** Distribution of image amounts across 27 emotions, and 4 contexts(Note*: facial expression consists of close-up face and full-body shots with visible face). The area of each colored region represents number of images in that group. Zoom in for better view.



Fig. 3. Each image in MoodArchive is paired with a structured caption generated by LLaVA-NeXT, including (1) a concise global summary, (2) three local emotional stimuli (i.e. visual elements evoking specific emotions), and (3) an overall emotion assessment. (See Appendix for instruction prompt.)

no per-worker HIT limit was imposed. Each completed HIT paid $0.05. The original caption was always displayed above the LLaVA-generated caption. The interface required exactly one of three mutually exclusive responses and provided explicit criteria covering physical content and emotional interpretation. No separate attention checks or post-hoc response exclusions were used. The study summary did not retain the number of unique workers, and workers' age, gender, and ethnicity were not collected.

Percentages are computed over all 60,000 mutually exclusive worker–image judgments: workers selected the LLaVA-generated caption as better in 85.6% of judgments, preferred the original alt-text in 8.6%, and selected neither caption in 5.8%. Cases judged inaccurate commonly involved inappropriate word choices, incorrect emotion interpretations, overly dramatic descriptions, or cultural mismatch. These results show that the generated captions are generally preferred to the original alt-text, while also confirming a non-negligible amount of annotation noise that may affect representation learning.

Overall, MoodArchive provides large-scale coverage of diverse emotions and contexts with structured, emotion-focused captions for affective representation learning.

B. MoodifyCLIP

At the scale of MoodArchive, occasional LLaVA-generated caption hallucinations are unavoidable, even with human validation. Our key observation is that, for emotion understanding, the benefit comes from *rich, fine-grained affective context*: long descriptions often specify *what* emotion is expressed, *where* it appears in the image, and *which* visual attributes support it, information that short captions typically miss. To reliably use such long captions, we adopt positional embedding

TABLE I

COMPARISON OF EMOTION DATASETS, HIGHLIGHTING THE NUMBER OF CATEGORIES, ATTRIBUTES, MODALITIES, AND TOTAL IMAGES ACROSS TRAINING, VALIDATION, AND TEST SETS. NOTE: 1. EMOTIC CONTAINS AMBIGUOUS LABELS, WHERE ONE IMAGE MAY CORRESPOND TO MULTIPLE LABELS. 2. EMOSET'S FULL DATASET HAS OVER 3 MILLION IMAGES, HERE, WE USE ITS SUBSET EMOSET-118K FOR THE ZERO-SHOT CLASSIFICATION TASK. 3. OUR ATTRIBUTES ARE DIVIDED INTO COARSE-GRAINED (E.G., FACIAL EXPRESSION, URBAN SCENERY) AND FINE-GRAINED (E.G., "SOFT SMILE", "GENTLE EYES" UNDER FACIAL EXPRESSION).

	Emotion		Total Images			
Dataset	Categories	Attributes	Train	Validation	Test	Modality Types
Emotion6 [9]	6	-	1,380	-	600	Images, Labels
Emotic [11]	26	-	23,268	3,316	7,203	Images, Labels*, Bbox
EmoSet(118K)* [10]	8	6	82,667	35,435	17,716	Images, Labels
Affection [12]	8	-	443,258	26,113	52,188	Images, Labels, Text
MoodArchive (ours)	27	4*		8,417,937		Images, Search queries, Text

interpolation [54]. MoodifyCLIP uses fine-grained region–text alignment and optimal-transport weighting to emphasize locally consistent image–text correspondences within long structured captions. This design may reduce, but does not eliminate, the influence of occasional mismatched descriptions. Concretely, it combines three mechanisms: (1) a *global* contrastive objective on both full captions and concise summaries, (2) a *fine-grained* region–text alignment objective, and (3) an *optimal transport* objective to downweight mismatched region–text pairs. Together, these components encourage consistent emotion-visual grounding and reduce the impact of occasional caption errors.

Global Contrastive Loss (GL). We follow CLIP and use a visual encoder E_i and a text encoder E_t to map an image and its captions into a shared embedding space. For each image I, we consider two text views: a fine-grained caption $\tilde{T}^f$ (the full five-sentence LLaVA annotation) and a summary caption $\tilde{T}^s$ (the first and last sentences, i.e., summary + emotion assessment. See Fig. 3). The corresponding embeddings are: $I^f = E_i(I)$, $T^f = E_t(\tilde{T}^f)$, $T^s = E_t(\tilde{T}^s)$. As outlined in Algorithm 1 in the Appendix, we optimize an InfoNCE contrastive loss between the image embedding I^f and the fine-grained text embedding T^f: $\mathcal{L}_f = (\mathcal{L}^f_{t2v} + \mathcal{L}^f_{v2t})/2$, with the explicit formulation:

$$\mathcal{L}^f_{t2v} = -\sum_{i=1}^{N} \log \frac{\exp(\cos(I^f_i, T^f_i)/\tau)}{\sum_{j=1}^{N} \exp(\cos(I^f_j, T^f_i)/\tau)},$$
$$\mathcal{L}^f_{v2t} = -\sum_{i=1}^{N} \log \frac{\exp(\cos(T^f_i, I^f_i)/\tau)}{\sum_{j=1}^{N} \exp(\cos(T^f_j, I^f_i)/\tau)}. \tag{1}$$

Similarly, we add an objective for the summary caption:

$$\mathcal{L}_s = (\mathcal{L}^s_{t2v} + \mathcal{L}^s_{v2t})/2. \tag{2}$$

Fine-grained Loss (FG). An emotion is not expressed uniformly across an image: it is typically conveyed by a few localized cues (e.g., facial muscles, color accents, textures). To capture this locality, we align each fine-grained affective stimulus description with the image regions that support it.

Given fine-grained text embeddings $\mathbf{T}^{fg}$ and image patch embeddings $\mathbf{I}^{fg} = \{I^{fg}_{i,m}\}_{m=1}^{M}$, we compute text-to-patch attention scores $w_{i,j,m}$ between the j-th stimulus text and the m-th image patch (details in Fig. 4 and Algorithm 1 in the Appendix). We normalize over patches to obtain attention weights:

$$\alpha_{i,j,m} = \frac{\exp(w_{i,j,m})}{\sum_{m'=1}^{M} \exp(w_{i,j,m'})}. \tag{3}$$

These weights explicitly enforce that each stimulus j focuses on its most relevant visual regions. We then form the stimulus-specific region representation by attention pooling:

$$I^{fg}_{i,j} = \sum_{m=1}^{M} \alpha_{i,j,m}\, I^{fg}_{i,m}. \tag{4}$$

We train the model with a bidirectional contrastive objective over these region representations:

$$\mathcal{L}^{fg}_{t2v} = -\sum_{i=1}^{N}\sum_{j=1}^{3} \log \frac{\exp(\cos(I^{fg}_{i,j}, T^{fg}_{i,j})/\tau)}{\sum_{m=1}^{M} \exp(\cos(I^{fg}_{i,m}, T^{fg}_{i,j})/\tau)},$$
$$\mathcal{L}^{fg}_{v2t} = -\sum_{i=1}^{N}\sum_{j=1}^{3} \log \frac{\exp(\cos(T^{fg}_{i,j}, I^{fg}_{i,j})/\tau)}{\sum_{m=1}^{M} \exp(\cos(T^{fg}_{i,j}, I^{fg}_{i,m})/\tau)}. \tag{5}$$

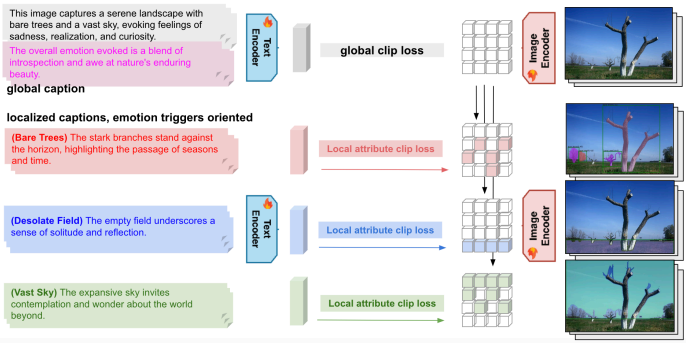


Fig. 4. Illustration of the fine-grained (FG) loss implementation in MoodifyCLIP. The model processes both global caption and region-specific emotional triggers, applying dedicated contrastive losses to align specific image regions with their corresponding emotional descriptions.

We combine them symmetrically as $\mathcal{L}_{fg} = (\mathcal{L}^{fg}_{t2v} + \mathcal{L}^{fg}_{v2t})/2$. This FG loss directly teaches the model which local visual cues correspond to which affective elements, rather than relying on global image-text alignment (see implementation details in Fig. 4 and Algorithm 1 in Appendix).

Optimal Transport Loss (OT). The fine-grained loss enforces *local* alignment between a few semantically matched image regions and text attributes ("zoom-in"). However, an emotion is typically expressed by *multiple* correlated cues (e.g., facial expression, color palette, posture, context), and the correspondence between regions and text descriptions is not one-to-one. We therefore introduce an optimal-transport objective to perform a global and mutually consistent region–text assignment ("zoom-out") [55].

Let $\mathbf{S} = \langle \mathbf{I}_{\text{fg}}, \mathbf{T}_{\text{fg}} \rangle_{\text{batch}} \in \mathbb{R}^{N \times N}$ denote the batch-wise similarity matrix between fine-grained image features and fine-grained text features, where larger values indicate higher similarity. We convert similarity into a transport cost (distance) matrix:

$$\mathbf{W} = \mathbf{1} - \mathbf{S}. \tag{6}$$

Intuitively, OT answers: *given all image cues and all text cues in the batch, what is the lowest-cost way to distribute "matching mass" so that high-similarity pairs receive more mass and mismatched pairs receive less mass?*

We compute the entropy-regularized transport plan using the Sinkhorn matrix-scaling algorithm [56]:

$$\mathbf{P}^{ot} = \text{Sinkhorn}(\mathbf{W}) \tag{7}$$

$$= \arg\min_{\mathbf{P} \in \Pi(\mathbf{a}, \mathbf{b})} \langle \mathbf{W}, \mathbf{P} \rangle - \epsilon H(\mathbf{P}), \tag{8}$$

where $\mathbf{a}$ and $\mathbf{b}$ are uniform marginals over samples (image/text) and $\Pi(\mathbf{a}, \mathbf{b})$ denotes the set of transport plans with these marginals; $H(\mathbf{P})$ is the entropy regularizer.

Specifically, we set $\mathbf{a} = \mathbf{b} = \frac{1}{N}\mathbf{1}$ and form the element-wise kernel $\mathbf{K} = \exp(-\mathbf{W}/\epsilon)$. Starting from $\mathbf{v}^{(0)} = \mathbf{1}$, Sinkhorn alternately applies

$$\mathbf{u}^{(l+1)} = \mathbf{a} \oslash (\mathbf{K}\mathbf{v}^{(l)}), \qquad \mathbf{v}^{(l+1)} = \mathbf{b} \oslash (\mathbf{K}^\top \mathbf{u}^{(l+1)}),$$

where $\oslash$ denotes element-wise division. Alternating these row- and column-scaling updates enforces the prescribed marginals, yielding the transport plan $\mathbf{P}^{ot} = \text{diag}(\mathbf{u})\mathbf{K}\,\text{diag}(\mathbf{v})$.

As a result, $\mathbf{P}^{ot}$ assigns larger mass to low-cost (high-similarity) pairs and suppresses high-cost mismatches, producing a globally consistent alignment that is especially useful when affective cues are ambiguous or overlapping.

Finally, we use $\mathbf{P}^{ot}$ to reweight similarities and apply a batch-level classification loss:

$$\mathcal{L}_{ot} = \text{CrossEntropy}\big(\mathbf{P}^{ot} \odot \mathbf{S},\ \mathbf{I}\big), \tag{9}$$

where $\mathbf{I}$ is the identity matrix indicating the positive pairs along the diagonal (implementation details are provided in Algorithm 1 in the Appendix).

Our total loss combines all components as:

$$\mathcal{L}_{\text{moodifyclip}} = \lambda_f \cdot \mathcal{L}_f + \lambda_s \cdot \mathcal{L}_s + \lambda_{fg} \cdot \mathcal{L}_{fg} + \lambda_{ot} \cdot \mathcal{L}_{ot} \tag{10}$$

C. Moodifier

With MoodifyCLIP fine-tuned on MoodArchive, we propose **Moodifier**, a training-free emotion-driven image editing system. As shown in Fig. 5 and Algorithm 1, Moodifier performs emotional editing by converting an abstract target emotion into (i) an executable text condition and (ii) spatial guidance, and then enforcing these constraints through cross-attention control in a diffusion model.

Given a source image V and a target emotion E, we first extract visual features $f_V = \text{Enc}_{\text{vis}}(V)$ and feed them to an MLLM (LLaVA-NeXT in our implementation) to obtain an emotion-specific editing prompt and a spatial attention map: $P_E, M_E = \text{MLLM}(f_V, E)$ (see Appendix for instruction prompts). The prompt P_E is then refined by MoodifyCLIP to translate the abstract emotion into concrete, context-aware visual attributes. Next, we apply a non-iterative inversion method [57] to map the source image into the diffusion latent space, yielding z_T. Finally, inspired by prompt-to-prompt attention control [41], we guide denoising by explicitly blending the cross-attention maps: we preserve the source attention outside emotion-relevant regions and inject target attention inside the MLLM-provided mask, producing an edited latent z_0^* that decodes to the final image V'.

Algorithm 1 Moodifier: MLLM-Enhanced Emotional Editing

Require: Source image V, target emotion E
Ensure: Emotionally edited image V'

$f_V \leftarrow \text{Enc}_{\text{vis}}(V)$
$P_E, M_E \leftarrow \text{MLLM}(f_V, E)$ (see Appendix for prompts)
$P_E \leftarrow \text{MoodifyCLIP}(P_E)$
$z_T \leftarrow \text{Inv}_{\text{dm}}(V)$
Initialize $z_T^* \leftarrow z_T$
for $t = T, T-1, \ldots, 1$ **do**
  $z_{t-1},\ M_t \leftarrow \text{DM}(z_t, \emptyset, t)$
  $z_{t-1}^{*,\text{raw}},\ M_t^* \leftarrow \text{DM}(z_t^*, P_E, t)$
  $\tilde{M}_t \leftarrow \text{BlendMaps}(M_t, M_t^*, M_E, t)$
  $z_{t-1}^* \leftarrow \text{DM}(z_t^*, P_E, t)\{M \leftarrow \tilde{M}_t\}$
  $z_{t-1}^* \leftarrow (1 - \mathbb{1}_{M_E>0}) \odot z_{t-1} \ + \ \mathbb{1}_{M_E>0} \odot z_{t-1}^*$
  $z_t \leftarrow z_{t-1};\quad z_t^* \leftarrow z_{t-1}^*$
end for
return z_0^* (decoded to V' by the diffusion decoder)

In a text-conditioned diffusion model, the only handle we need for controlled editing is cross-attention: it determines which text concepts affect which spatial locations at each denoising step. Concretely, at step t the model computes an attention map $M_t = \text{Softmax}\left(\frac{QK^T}{\sqrt{d}}\right)$, where $Q = \ell_Q(\phi(z_t))$ are queries from the current latent (visual features) and $K = \ell_K(\psi(P))$ are keys from the conditioning prompt (here P denotes the prompt used for conditioning, i.e., P_E in our setting). Each entry $M_{i,j}$ measures how much the j-th text token influences the i-th pixel/location.

What we want for emotion editing. We need two signals: *what* to change and *where* to change. Given the source image and target emotion, our MLLM provides (i) an emotion-specific prompt P_E describing the desired affective changes and (ii) a spatial stimulus map M_E indicating emotion-relevant regions. In Algorithm 1, DM denotes a single diffusion step that simultaneously denoises and exposes cross-attention. Specifically, $\text{DM}(z_t, P_E, t)$ performs the step $t \to t$–1 conditioned on P_E, and outputs both the denoised latent z_{t-1} and the attention map M_t.

How we enforce localized, structure-preserving edits. We compute two attention maps at each step: a source attention

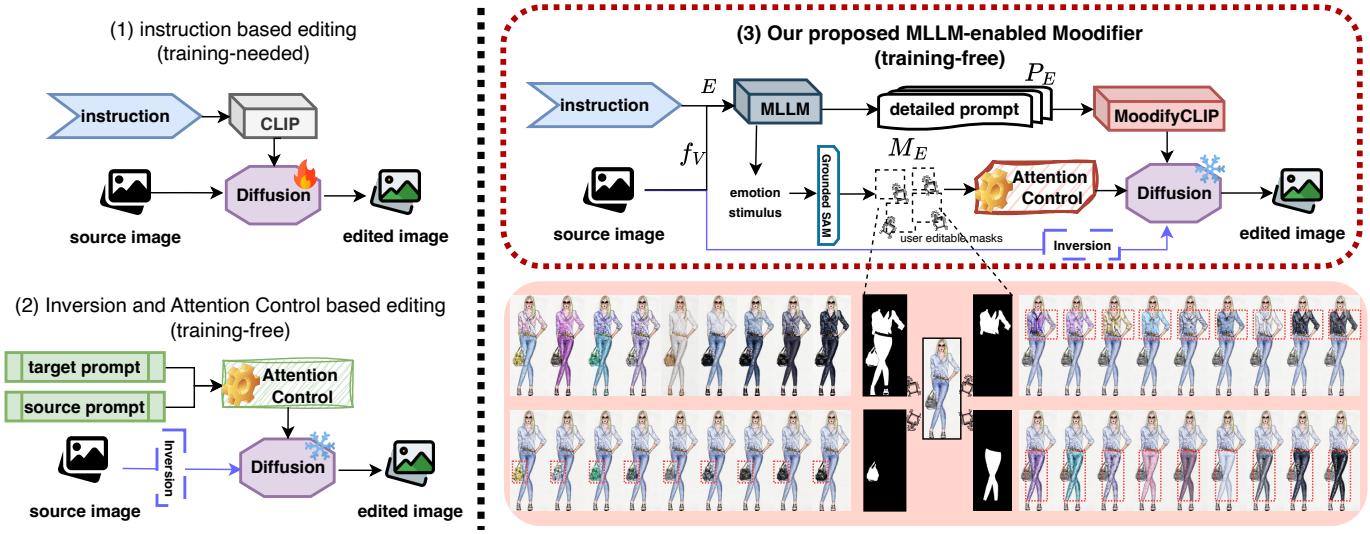


Fig. 5. Our MLLM-enhanced Moodifier (right panel (3)) integrates ideas from instruction-based and attention-control approaches, converting simple commands into rich emotional prompts for targeted editing. We use LLaVA-NeXT as the MLLM. The bottom of panel (3) demonstrates both fully automated processing and user-controlled region selection while ensuring natural transitions by modifying latent representations instead of using inpainting. The four mask configurations in the lower example block demonstrate both single-region and multi-region editing: the upper-left configuration jointly selects the upper garment, trousers, and handbag through their union mask, the upper-right, lower-right, and lower-left configurations select only the upper garment, trousers, and handbag, respectively. Thus, users may edit each region individually or combine multiple regions within one editing pass. All regions selected in the same pass share the same target emotion. The left panels (1) and (2) summarize existing editing paradigms for comparison.

map from the source branch and a target attention map from the emotion-conditioned branch, and then blend them: $\tilde{M}_t = \text{BlendMaps}(M_t, M_t^*, M_E, t)$. The goal is simple: preserve the source structure where the emotion should *not* change, and inject target semantics only where the emotion *should* appear.

$$\tilde{M}_t = \begin{cases} \text{Refine}(M^{\text{src}}, M^{\text{tgt}}, M_E) & t \geq \tau_c \\ M^{\text{tgt}} & t < \tau_c \end{cases} \tag{11}$$

$$\text{Refine}(M^{\text{src}}, M^{\text{tgt}}, M_E)_{i,j} = \begin{cases} (M^{\text{tgt}})_{i,j} & \text{if } (i,j) \in M_E \\ (M^{\text{src}})_{i,j} & \text{otherwise} \end{cases} \tag{12}$$

where τ_c is a timestep threshold that switches the blending regime, and $M_{\text{src}} \equiv M_t$ (source attention), $M_{\text{tgt}} \equiv M_t^*$ (target attention). Here the condition $(i,j) \in M_E$ indicates that location (i,j) is selected by the stimulus map (i.e., an emotion-relevant region).

Why the schedule works. Following Algorithm 1, reverse denoising proceeds from $t = T$ to $t = 1$, therefore, larger t denotes an earlier and noisier denoising step. At early steps ($t \geq \tau_c$), we use refined attention to anchor global structure and identity by mostly following the source attention outside emotion-relevant regions. At later steps ($t < \tau_c$), we rely on target attention to inject emotion-specific details. Finally, the stimulus map M_E is also used to blend source and target latents only inside emotion-relevant regions, so the rest of the image remains unchanged. Moodifier allows users to select either an individual region or multiple spatially disjoint regions for editing. Given a selected subset $\mathcal{S}$ of binary region masks $\{M^{(k)}\}$, the corresponding stimulus map is constructed as their pixelwise union, $M_E = \bigvee_{k \in \mathcal{S}} M^{(k)}$. When $|\mathcal{S}| = 1$, Moodifier performs single-region editing; when $|\mathcal{S}| > 1$, all selected regions are edited simultaneously using the same attention-control and latent-blending procedure. As illustrated in Fig. 5, users can independently select the upper garment, trousers, or handbag, or jointly select all three regions through their union mask. All regions selected within the same editing pass share the same target emotion E, assigning different target emotions to different regions within one pass is not currently supported.

IV. Experiments and Results

A. Experimental Setup

MoodArchive Dataset. *MoodArchive* is a large-scale affective image corpus designed to support emotion understanding and emotion-driven editing. It contains over 8 million images organized around 27 emotion-based retrieval categories and four visual contexts per retrieval emotion (Fig. 2C). This scale and structured coverage provide the diversity needed for models to learn not only coarse emotion labels, but also the situational cues and visual attributes that express each emotion.

Emotion and context coverage. As shown in Fig. 2C, the dataset spans 27 emotions, and each emotion is organized into four contexts: (1) *facial expression* (close-up faces and full-body shots with visible faces), (2) *natural scenery* (natural environments and vistas), (3) *urban scenery* (city and man-made environments), and (4) *object class* (emotion-eliciting objects on simple backgrounds). This context partitioning makes the same abstract emotion appear in visually distinct ways, which is essential for learning context-dependent affective cues.

Hierarchical attributes and metadata. Fig. 2A summarizes the hierarchical curation of MoodArchive. Each data unit consists of an image paired with an automatically generated caption and structured metadata: 1) a *weak retrieval label* recording the seed emotion used to construct the retrieval query, 2) a *context label* indicating one of the four contexts,

and 3) *fine-grained attributes* (key phrases or scenarios) that describe concrete visual and situational triggers of emotion. This hierarchy bridges abstract emotions and concrete visual evidence, enabling downstream models to learn both global affect recognition and localized, attribute-level grounding. The weak retrieval label indicates the emotion used for data collection and is not treated as an exclusive ground-truth annotation.

Evaluation datasets and protocols. For zero-shot classification, we evaluate on Emotion6 [9], a six-category Flickr benchmark with crowd-sourced emotion distributions; the human-annotated EmoSet-118K subset [10], which covers eight emotions and emotion-related visual attributes; and EMOTIC [11], which provides person-level multi-label emotions, continuous valence–arousal–dominance annotations, and bounding boxes in natural contexts. We report Top-1/2/3 accuracy for three CLIP variants with B/16 and L/14 backbones. We additionally evaluate FLIP-L/14 [32] using the same uncropped images and class prompts, because FLIP is pretrained on face–text pairs, this setting measures its transfer to full-image affect recognition rather than performance under a native facial-expression-recognition protocol.

For bidirectional image–text retrieval, we evaluate on SentiCap [58], Affection [12], and MoodArchive-5k, a held-out set of 5,000 human-verified image–text pairs, using R@1 and R@5. MoodArchive retrieval labels are not used as evaluation ground truth: classification follows the official benchmark annotations, retrieval uses the benchmark image–caption pairs, and editing uses the target emotion specified by the experimental protocol. All classification and retrieval results are reported using five-fold cross-validation.

MoodifyCLIP is evaluated at both ViT-B/16 and ViT-L/14 scales, while Moodifier performs emotion-driven editing without task-specific editor training. Complete model, optimization, training, inference, and hardware configurations are provided in Appendices B and C.

Moodifier evaluation protocol. Because emotion-driven editing has no unique reference output, we combine automatic content-preservation and emotion-alignment metrics with human preference judgments. The automatic evaluation uses 240 held-out MoodArchive source images, with 30 images from each of eight categories (face, bag, bracelet, clothes, earring, necklace, ring, and vase). Editing every source toward all 27 target emotions yields 6,480 outputs per method. We report structure distance, PSNR, LPIPS, MSE, SSIM, and whole-image and edited-region CLIP similarity, averaging the results over emotions, sources, and categories.

We compare Moodifier with prior editing methods [1]–[3], [49]–[51], [57], [59]–[62] and the recent UltraEdit-SD3 editor [52]. All methods use the same source images, target-emotion prompts, output resolution, and automatic evaluation protocol. UltraEdit-SD3 was added during revision and was therefore evaluated using automatic metrics only, its human-evaluation entries are marked as "–".

To complement CLIP similarity with a continuous affective measure, we use the publicly available frozen EmoNet valence and arousal predictors [63], which are independent of CLIP, MoodArchive, and MoodifyCLIP. For each target emotion E, the reference coordinates $(v_E, a_E) \in [0,1]^2$ are obtained from the NRC VAD Lexicon [64]. After mapping the EmoNet prediction $(\widehat{v}, \widehat{a})$ to the same range, we compute

$$d_{\mathrm{VA}} = \sqrt{(\widehat{v} - v_E)^2 + (\widehat{a} - a_E)^2}, \tag{13}$$

where lower values indicate closer continuous affective alignment. We apply this metric to Moodifier, Inversion-Free Editing, and UltraEdit-SD3 on the complete 240-source, 27-emotion evaluation set. Normalization and implementation details are provided in the supplementary material.

Human evaluation protocol. The original Amazon Mechanical Turk study uses the interface shown in Appendix Fig. 2 and includes 12 editing methods. It evaluates five source images from each of the eight categories under all 27 target emotions, yielding 1,080 source–emotion instances. Each instance is independently evaluated by three workers, resulting in 3,240 worker–instance judgments. For each instance, anonymized method outputs are shown in randomized order, and workers select exactly three methods under each of three criteria: *Visual Appeal*, *Emotion Accuracy*, and *Overall Best*.

Participation was open to the MTurk worker pool without additional geographic, approval-rate, previously approved-HIT, or per-worker HIT limits. Each completed HIT paid $0.05. No separate attention checks or post-hoc response exclusions were used, although the interface required complete selections. The retained study summary does not include the number of unique workers, and age, gender, and ethnicity were not collected.

For each method and criterion, an instance is counted as a hit if at least one of its three workers includes that method among the top three. The resulting instance-level hit rate measures preference coverage rather than within-instance agreement. Consequently, hit rates across methods are not required to sum to 300%.

B. Result Analysis

MoodifyCLIP evaluation. Tables II and III show that MoodifyCLIP consistently improves over the baseline CLIP variants in zero-shot emotion classification and bidirectional retrieval. The classification gain is particularly pronounced on EMOTIC, where affect recognition depends strongly on contextual and subtle visual cues. Although all models perform better on MoodArchive-5k because its captions follow the training annotation style, MoodifyCLIP also improves on SentiCap and Affection, demonstrating transfer beyond MoodArchive. The facial-domain FLIP baseline transfers most clearly to EMOTIC, where it outperforms all general-purpose L/14 CLIP variants across Top-1/2/3 accuracy. MoodifyCLIP nevertheless remains higher under all three metrics, suggesting that facial pretraining is useful for person-centered affect, while broader context-aware supervision provides additional benefits for full-image emotion recognition.

Table IV shows that fine-grained alignment improves local stimulus grounding, while optimal-transport reweighting reduces the influence of mismatched region–text pairs. Their combination achieves the strongest overall performance, indicating that the two objectives provide complementary supervision.

TABLE II
ZERO-SHOT EMOTION-CLASSIFICATION ACCURACY (%) ON EMOTION6, EMOSET, AND EMOTIC. TOP-1/2/3 ARE REPORTED. BOLD INDICATES THE BEST RESULT.

		Emotion6			EmoSet			EMOTIC		
		Top-1	Top-2	Top-3	Top-1	Top-2	Top-3	Top-1	Top-2	Top-3
B/16	$CLIP_{CommonPool}$	$41.82_{\pm 1.72}$	$68.99_{\pm 1.80}$	$83.13_{\pm 0.79}$	$40.07_{\pm 0.24}$	$62.73_{\pm 0.26}$	$74.43_{\pm 0.34}$	$27.59_{\pm 0.44}$	$39.67_{\pm 0.75}$	$48.74_{\pm 0.57}$
	$CLIP_{DataComp}$	$47.98_{\pm 1.53}$	$71.92_{\pm 1.61}$	$86.57_{\pm 1.92}$	$45.88_{\pm 0.35}$	$65.33_{\pm 0.32}$	$76.73_{\pm 0.25}$	$24.28_{\pm 0.87}$	$36.98_{\pm 0.44}$	$46.21_{\pm 0.43}$
	$CLIP_{LAION}$	$50.30_{\pm 1.05}$	$73.23_{\pm 3.00}$	$84.80_{\pm 3.11}$	**49.43**$_{\pm 0.43}$	**70.74**$_{\pm 0.40}$	**81.96**$_{\pm 0.20}$	$23.94_{\pm 0.61}$	$35.70_{\pm 0.57}$	$44.01_{\pm 0.39}$
	ModoifyCLIP (Ours)	**55.86**$_{\pm 0.84}$	**77.93**$_{\pm 1.72}$	**90.66**$_{\pm 1.50}$	$46.96_{\pm 1.31}$	$69.58_{\pm 1.93}$	$81.22_{\pm 1.32}$	**38.35**$_{\pm 0.76}$	**53.49**$_{\pm 0.84}$	**61.95**$_{\pm 0.56}$
L/14	$CLIP_{CommonPool}$	$49.95_{\pm 0.92}$	$74.75_{\pm 2.31}$	$87.02_{\pm 1.73}$	$46.62_{\pm 0.25}$	$68.71_{\pm 0.31}$	$80.49_{\pm 0.31}$	$27.55_{\pm 0.49}$	$41.55_{\pm 0.52}$	$50.34_{\pm 0.64}$
	$CLIP_{DataComp}$	$51.77_{\pm 2.48}$	$73.94_{\pm 1.52}$	$87.63_{\pm 1.13}$	$48.54_{\pm 0.38}$	$67.43_{\pm 0.28}$	$78.63_{\pm 0.27}$	$26.02_{\pm 0.75}$	$39.20_{\pm 0.83}$	$47.95_{\pm 0.67}$
	$CLIP_{LAION}$	**55.76**$_{\pm 0.29}$	$79.39_{\pm 2.27}$	$89.80_{\pm 1.72}$	$49.98_{\pm 0.40}$	$70.60_{\pm 0.40}$	$81.83_{\pm 0.30}$	$21.53_{\pm 0.26}$	$33.07_{\pm 0.20}$	$41.53_{\pm 0.32}$
	FLIP [32]	$52.40_{\pm 1.21}$	$77.60_{\pm 1.34}$	$89.20_{\pm 1.12}$	$51.10_{\pm 0.43}$	$70.10_{\pm 0.47}$	$81.50_{\pm 0.39}$	$33.70_{\pm 0.66}$	$49.80_{\pm 0.71}$	$59.20_{\pm 0.63}$
	ModoifyCLIP (Ours)	$54.83_{\pm 2.32}$	**80.51**$_{\pm 0.98}$	**91.11**$_{\pm 0.79}$	**56.33**$_{\pm 0.29}$	**72.49**$_{\pm 0.27}$	**83.56**$_{\pm 0.26}$	**38.40**$_{\pm 0.79}$	**57.09**$_{\pm 0.78}$	**67.12**$_{\pm 0.82}$

TABLE III
ZERO-SHOT BIDIRECTIONAL IMAGE–TEXT RETRIEVAL (R@1/R@5) ON SENTICAP, AFFECTION, AND MOODARCHIVE-5K.

		SentiCap				Affection				MoodArchive-5k			
		Image-to-Text		Text-to-Image		Image-to-Text		Text-to-Image		Image-to-Text		Text-to-Image	
		R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5
B/16	$CLIP_{CommonPool}$	$29.98_{\pm 0.98}$	$65.08_{\pm 1.61}$	$28.08_{\pm 1.02}$	$63.68_{\pm 1.96}$	$26.70_{\pm 0.41}$	$54.02_{\pm 0.83}$	$25.36_{\pm 1.09}$	$52.18_{\pm 0.55}$	$79.22_{\pm 1.02}$	$94.56_{\pm 0.73}$	$69.60_{\pm 0.89}$	$89.26_{\pm 0.92}$
	$CLIP_{DataComp}$	$41.08_{\pm 1.45}$	$78.46_{\pm 0.93}$	$39.74_{\pm 0.81}$	$76.22_{\pm 0.58}$	$40.30_{\pm 1.55}$	$66.98_{\pm 0.50}$	$37.74_{\pm 0.92}$	$63.96_{\pm 1.45}$	$82.90_{\pm 0.98}$	$98.42_{\pm 0.25}$	$77.44_{\pm 0.92}$	$96.18_{\pm 0.57}$
	$CLIP_{LAION}$	**42.54**$_{\pm 0.57}$	**79.24**$_{\pm 1.03}$	**41.24**$_{\pm 0.36}$	$77.80_{\pm 1.05}$	$40.78_{\pm 0.69}$	$68.20_{\pm 1.16}$	$39.66_{\pm 0.56}$	$66.84_{\pm 1.17}$	$83.22_{\pm 0.69}$	$98.24_{\pm 0.34}$	$78.40_{\pm 1.14}$	$96.62_{\pm 0.44}$
	ModoifyCLIP (Ours)	$41.10_{\pm 1.30}$	$78.46_{\pm 0.50}$	$40.40_{\pm 0.81}$	**78.54**$_{\pm 1.37}$	**42.58**$_{\pm 1.41}$	**70.60**$_{\pm 1.20}$	**43.70**$_{\pm 0.73}$	**72.58**$_{\pm 1.05}$	**84.52**$_{\pm 0.46}$	**99.02**$_{\pm 0.51}$	**81.20**$_{\pm 0.71}$	**97.96**$_{\pm 0.45}$
L/14	$CLIP_{CommonPool}$	$41.94_{\pm 0.33}$	$78.54_{\pm 0.59}$	$41.08_{\pm 1.13}$	$77.40_{\pm 0.62}$	$41.64_{\pm 1.23}$	$69.16_{\pm 1.05}$	$40.94_{\pm 0.72}$	$67.44_{\pm 1.30}$	$83.46_{\pm 1.08}$	$98.48_{\pm 0.32}$	$77.96_{\pm 1.21}$	$95.94_{\pm 0.50}$
	$CLIP_{DataComp}$	$43.72_{\pm 1.04}$	$80.08_{\pm 0.97}$	$42.46_{\pm 0.30}$	$79.18_{\pm 0.47}$	$44.14_{\pm 0.87}$	$70.96_{\pm 1.32}$	$42.84_{\pm 1.77}$	$69.10_{\pm 1.04}$	$84.04_{\pm 0.96}$	$98.62_{\pm 0.50}$	$79.92_{\pm 0.82}$	$96.84_{\pm 0.45}$
	$CLIP_{LAION}$	$43.94_{\pm 0.52}$	$80.38_{\pm 1.22}$	$42.40_{\pm 0.89}$	$79.62_{\pm 1.46}$	$43.90_{\pm 1.44}$	$70.34_{\pm 1.24}$	$44.44_{\pm 0.29}$	$69.66_{\pm 1.74}$	$85.02_{\pm 0.77}$	$98.68_{\pm 0.29}$	$81.28_{\pm 0.15}$	$97.20_{\pm 0.49}$
	ModoifyCLIP (Ours)	**44.72**$_{\pm 1.10}$	**80.46**$_{\pm 1.14}$	**44.90**$_{\pm 0.86}$	**82.18**$_{\pm 0.70}$	**46.34**$_{\pm 0.66}$	**74.02**$_{\pm 0.82}$	**48.56**$_{\pm 0.91}$	**76.56**$_{\pm 1.22}$	**86.16**$_{\pm 0.37}$	**98.84**$_{\pm 0.38}$	**85.18**$_{\pm 0.54}$	**98.96**$_{\pm 0.29}$

TABLE IV
ABLATION OF MOODIFYCLIP LOSSES USING THE VIT-L/14 BACKBONE.

CLIP Losses		Text Retrieval			Image Retrieval			Classification		
FG	OT	SentiCap R@1	Affection R@1	MoodArchive-5k R@1	SentiCap R@1	Affection R@1	MoodArchive-5k R@1	Emotion6 Acc.(%)	EmoSet Acc.(%)	Emotic Acc.(%)
×	×	$41.14_{\pm 0.84}$	$40.68_{\pm 1.30}$	$83.56_{\pm 1.13}$	$40.38_{\pm 1.20}$	$42.76_{\pm 1.59}$	$84.08_{\pm 0.48}$	$48.03_{\pm 2.32}$	$52.56_{\pm 0.13}$	$34.97_{\pm 0.31}$
✓	×	$42.46_{\pm 0.89}$	$42.64_{\pm 0.96}$	$84.36_{\pm 0.90}$	$40.46_{\pm 1.40}$	$44.00_{\pm 0.75}$	$84.58_{\pm 0.52}$	$51.82_{\pm 1.40}$	$54.10_{\pm 0.28}$	$37.85_{\pm 0.62}$
×	✓	$41.56_{\pm 0.59}$	$41.28_{\pm 0.71}$	$84.20_{\pm 0.99}$	$39.64_{\pm 0.97}$	$42.16_{\pm 1.04}$	$84.30_{\pm 0.55}$	$48.78_{\pm 1.45}$	$52.93_{\pm 0.22}$	$35.50_{\pm 0.76}$
✓	✓	**44.72**$_{\pm 1.10}$	**46.34**$_{\pm 0.66}$	**86.16**$_{\pm 0.37}$	**44.90**$_{\pm 0.86}$	**48.56**$_{\pm 0.91}$	**85.18**$_{\pm 0.54}$	**54.83**$_{\pm 2.32}$	**56.33**$_{\pm 0.29}$	**38.40**$_{\pm 0.79}$

TABLE V
AUTOMATIC AND HUMAN EVALUATION OF IMAGE-EDITING METHODS. AUTOMATIC METRICS USE 240 SOURCES AND 27 EMOTIONS (6,480 OUTPUTS PER METHOD). MTURK HIT RATES USE 40 SOURCES AND 27 EMOTIONS (1,080 INSTANCES), EACH EVALUATED BY THREE WORKERS. ULTRAEDIT-SD3 WAS NOT INCLUDED IN MTURK, "–" DENOTES UNAVAILABLE. CLIP* USES OPENAI VIT-L/14. *Whole* AND *Edited* DENOTE FULL-IMAGE AND STIMULUS-REGION SIMILARITY, RESPECTIVELY.

Method	Structure	Background Preservation				CLIP Score*		Instance-level Human Hit Rate (%)		
	Distance$_{\times 10^3}$ ↓	PSNR ↑	LPIPS$_{\times 10^3}$ ↓	MSE$_{\times 10^4}$ ↓	SSIM$_{\times 10^2}$ ↑	Whole ↑	Edited ↑	Visual Appeal ↑	Emotion Acc ↑	Overall ↑
ControlNet	21.9	22.01	102.3	104.8	80.3	12.80	10.56	27.1	23.5	25.2
Instruct-Diffusion	34.5	18.55	175.8	227.6	74.2	13.24	11.90	18.7	19.8	18.5
Instruct-Pix2Pix	22.0	23.17	112.6	100.2	81.0	13.27	11.79	36.9	34.2	32.8
EmoEdit	29.1	20.90	129.8	158.5	77.9	13.81	12.27	50.2	45.3	51.8
MasaCtrl	18.4	21.52	144.2	76.2	79.4	12.80	10.35	26.1	25.0	23.4
Plug-and-Play	**13.9**	22.02	123.5	**72.3**	80.5	12.85	10.12	34.8	32.1	33.2
Edit Friendly	15.1	21.39	130.7	104.9	78.9	13.47	11.23	27.3	25.7	26.1
StyleDiffusion	20.4	17.42	159.7	219.0	70.9	13.42	11.35	21.5	22.3	20.1
Null-text Inversion	16.3	23.04	152.1	81.9	80.0	13.76	11.66	31.2	35.8	34.5
Direct Inversion	18.5	23.34	134.2	78.6	81.1	13.76	11.47	29.6	32.1	30.2
Inversion-Free Editing	14.6	**24.11**	96.8	80.1	**82.9**	14.23	12.43	53.9	51.4	53.6
UltraEdit-SD3 (free-form)	16.1	23.46	108.2	91.7	82.6	14.53	12.38	–	–	–
Moodifier (Ours)	17.4	22.19	**94.9**	127.3	82.2	**16.13**	**14.08**	**65.1**	**71.5**	**78.3**

Moodifier Evaluation. Table V reveals a trade-off between content preservation and emotion alignment. Preservation-oriented methods, including Plug-and-Play and the inversion-based baselines, generally retain structure more effectively but provide weaker emotion alignment, whereas EmoEdit produces stronger changes at the cost of content integrity. UltraEdit-SD3 performs broadly comparably to the strongest prior baselines but remains below Moodifier in emotion-alignment scores. Moodifier achieves the best LPIPS and both CLIP scores, providing the strongest overall balance between perceptual preservation and target-emotion alignment. Among the 12 methods included in the original human study, Moodifier also obtains the highest instance-level hit rate under all three criteria.

Appendix Table I further reports Moodifier's CLIP Edited similarity and LPIPS for each of the 27 target emotions. Emotion alignment is stronger for joy, amusement, love, surprise, and excitement, whose affective meanings can be conveyed through salient visual changes, but weaker for remorse, embarrassment, and confusion, whose visual manifestations are

subtler and overlap with related emotional states.

TABLE VI
CONTINUOUS VALENCE–AROUSAL ALIGNMENT FOR THREE REPRESENTATIVE METHODS ON THE COMPLETE 240-SOURCE EVALUATION SET. THE COMPARISON INCLUDES OUR METHOD, THE STRONGEST PRIOR BASELINE ACCORDING TO THE ORIGINAL OVERALL HUMAN HIT RATE, AND THE RECENT SD3-BASED BASELINE ADDED IN THIS REVISION. LOWER VA DISTANCE INDICATES BETTER TARGET-EMOTION ALIGNMENT.

Method	EmoNet VA Distance ↓
Inversion-Free Editing	0.428
UltraEdit-SD3 (free-form)	0.406
Moodifier (Ours)	**0.369**

Among the three representative methods in Table VI, Moodifier obtains the lowest VA distance, providing affective evidence complementary to the CLIP and human evaluations.

Qualitative comparisons in Fig. 6 show that prior methods commonly exhibit identity drift, visible artifacts, or insufficiently perceptible emotional changes. In contrast, Moodifier introduces emotion-relevant attributes within localized regions while better preserving the source structure. Appendix Fig. 3 presents all 27 target-emotion edits for one fixed facial source image under identical inference settings. The grid shows clear variation for visually salient emotions while also revealing residual overlap among abstract or semantically related targets, complementing the per-emotion quantitative results.

Table VII and Fig. 7 show that detailed prompts strengthen emotion expression but may cause global drift, whereas spatial masks improve structural preservation but can produce overly subtle edits. Combining P_E and M_E provides the best trade-off, confirming that Moodifier benefits from complementary semantic and spatial guidance.

Hallucination sensitivity. MoodArchive annotation errors affect editing indirectly through MoodifyCLIP training, whereas errors in the inference-time prompt P_E directly condition diffusion. Appendix Fig. 4 isolates the latter using controlled nonexistent-object, incorrect-attribute, and incorrect-affect corruptions while fixing all other inputs and generation settings. These corruptions respectively introduce a source-unsupported bird, unnecessarily recolor the tree and its branches, and redirect an admiration edit toward dark and ominous visual cues. The experiment therefore illustrates direct inference-time prompt propagation rather than robustness to training-time annotation noise. Because the interface exposes P_E for optional user revision, these observations also motivate inspecting or correcting the prompt before generation and developing automatic prompt verification in future work.

V. CONCLUSION AND DISCUSSION

We presented an integrated emotion-driven image-editing framework comprising MoodArchive, MoodifyCLIP, and the training-free Moodifier. By grounding abstract emotions in context-aware visual attributes, the framework enables targeted edits while maintaining content structure and achieves improved emotion alignment and human preference across automatic and subjective evaluations.

Limitations and future work. First, MoodArchive annotation noise may affect representation learning, whereas hallucinated inference-time prompts P_E may directly redirect editing, controlled noise evaluation and automatic prompt verification remain future work. Second, emotional interpretation is subjective and context-dependent. The 27 retrieval categories are not mutually exclusive, and the single-seed retrieval process does not capture exhaustive multi-label, viewer-specific, or culturally varying emotion distributions. Third, emotion-driven editing lacks standardized benchmarks and reference targets. Future work will develop human-validated source–target pairs for more reproducible evaluation. Fourth, the original MTurk studies did not retain unique-worker counts or collect age, gender, or ethnicity, preventing demographic subgroup analysis. The fixed caption order in the caption-validation study may also introduce positional bias. Future studies will retain anonymized worker metadata, collect optional demographics, and randomize all presentation orders.

REFERENCES

[1] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3813–3824, 2023.

[2] T. Brooks, A. Holynski, and A. A. Efros, "Instructpix2pix: Learning to follow image editing instructions," *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18 392–18 402, 2022.

[3] Z. Geng, B. Yang, T. Hang, C. Li, S. Gu, T. Zhang, J. Bao, Z. Zhang, H. Hu, D. Chen, and B. Guo, "Instructdiffusion: A generalist modeling interface for vision tasks," in *Proc. CVPR*, 2024.

[4] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollár, and C. L. Zitnick, "Microsoft coco captions: Data collection and evaluation server," *arXiv preprint arXiv:1504.00325*, 2015.

[5] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions," *Transactions of the Association for Computational Linguistics*, vol. 2, pp. 67–78, 2014.

[6] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," *Advances in neural information processing systems*, vol. 36, 2024.

[7] P. Ekman, W. V. Friesen, M. O'sullivan, A. Chan, I. Diacoyanni-Tarlatzis, K. Heider, R. Krause, W. A. LeCompte, T. Pitcairn, P. E. Ricci-Bitti *et al.*, "Universals and cultural differences in the judgments of facial expressions of emotion." *Journal of personality and social psychology*, vol. 53, no. 4, p. 712, 1987.

[8] J. Machajdik and A. Hanbury, "Affective image classification using features inspired by psychology and art theory," in *Proceedings of the 18th ACM international conference on Multimedia*, 2010, pp. 83–92.

[9] K.-C. Peng, T. Chen, A. Sadovnik, and A. C. Gallagher, "A mixed bag of emotions: Model, predict, and transfer emotion distributions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 860–868.

[10] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 20 383–20 394.

[11] R. Kosti, J. M. Alvarez, A. Recasens, and A. Lapedriza, "Context based emotion recognition using emotic dataset," *IEEE transactions on pattern analysis and machine intelligence*, vol. 42, no. 11, pp. 2755–2766, 2019.

[12] P. Achlioptas, M. Ovsjanikov, L. Guibas, and S. Tulyakov, "Affection: Learning affective explanations for real-world visual data," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 6641–6651.

[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 *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[14] X. Gu, T.-Y. Lin, W. Kuo, and Y. Cui, "Open-vocabulary object detection via vision and language knowledge distillation," in *International Conference on Learning Representations*, 2023.

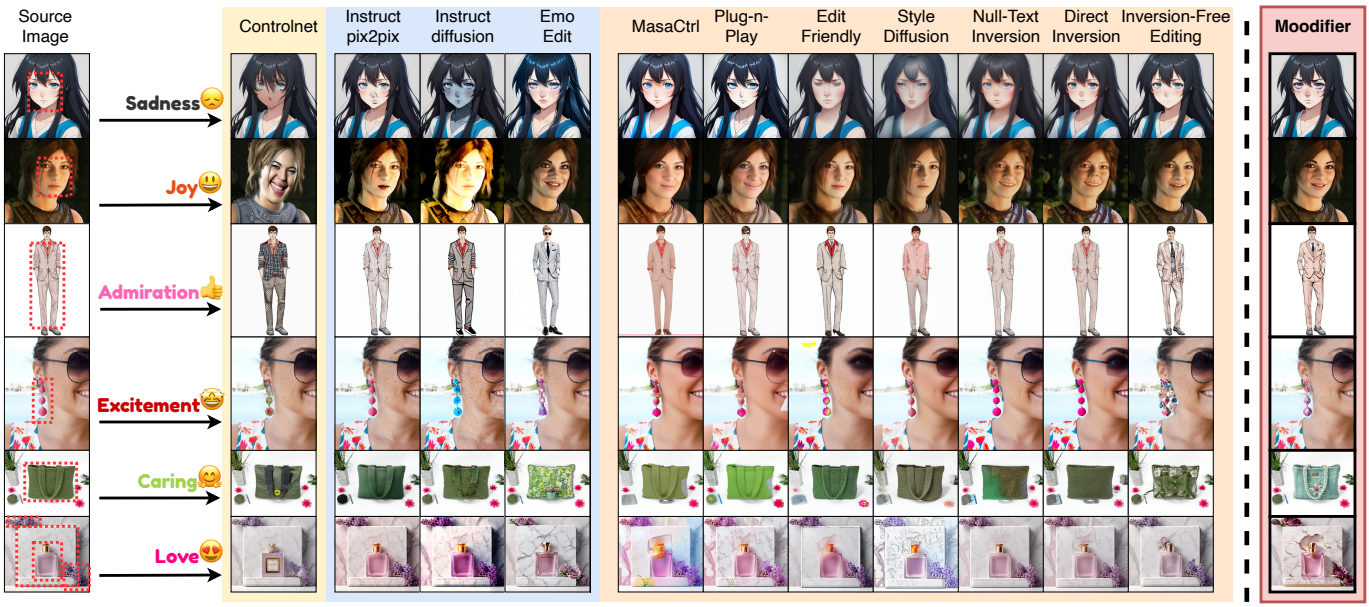


Fig. 6. Qualitative comparisons with other editing methods. Zoom in for better view.

TABLE VII
ABLATION STUDY OF MOODIFIER VARIANTS. NOTE THAT NO FACIAL IMAGES ARE INCLUDED IN THIS STUDY DUE TO THEIR FIXED EDITING REGIONS. THIS EXPLAINS DIFFERENCES IN THE LAST ROW COMPARED TO TABLE V. BOLD VALUES INDICATE BEST PERFORMANCE, † VALUES INDICATE SECOND-BEST PERFORMANCE.

CLIP Model	Edit Method		Structure	Background Preservation				CLIP Score
	Detail Prompts (P_E)	AttnMask (M_E)	Distance$_{\times 10^3}$ ↓	PSNR ↑	LPIPS$_{\times 10^3}$ ↓	MSE$_{\times 10^4}$ ↓	SSIM$_{\times 10^2}$ ↑	Similarity ↑
CLIP-ViT-L-14	×	×	23.33	18.81	150.2	155.1	61.40	15.03
	✓	×	22.30	19.01	146.2	152.3	69.32	15.80
	×	✓	18.61	20.82	120.3	134.4	81.50	15.15
	✓	✓	**18.32**	20.58	116.1	132.9	**82.23**	15.64
MoodifyCLIP-ViT-L-14	✓	×	21.90	19.91	148.7	147.6	65.61	**16.07**
	✓	✓	18.54†	**21.55**	**102.2**	**131.2**	81.63†	16.03†

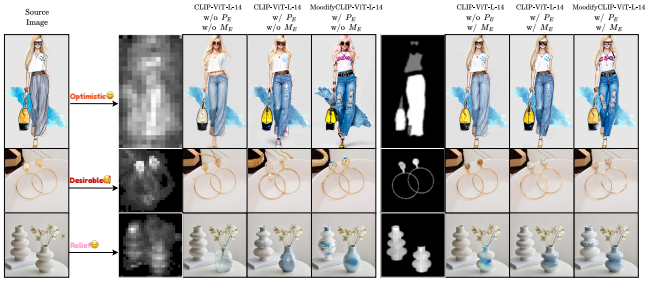


Fig. 7. A qualitative example for ablation. Ours allows faithful and rich semantic changes as well as better structural integrity preservation. (See notation reference from Fig. 5, where P_E represents the detailed MLLM-generated prompt and M_E denotes the emotion stimulus masks). Zoom in for better view.

[15] L. H. Li, P. Zhang, H. Zhang, J. Yang, C. Li, Y. Zhong, L. Wang, L. Yuan, L. Zhang, J.-N. Hwang *et al.*, "Grounded language-image pre-training," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 965–10 975.

[16] B. Li, K. Q. Weinberger, S. Belongie, V. Koltun, and R. Ranftl, "Language-driven semantic segmentation," in *International Conference on Learning Representations*, 2023.

[17] J. Xu, S. De Mello, S. Liu, W. Byeon, T. Breuel, J. Kautz, and X. Wang, "Groupvit: Semantic segmentation emerges from text supervision," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 18 134–18 144.

[18] H. Luo, L. Ji, M. Zhong, Y. Chen, W. Lei, N. Duan, and T. Li, "Clip4clip: An empirical study of clip for end to end video clip retrieval," *Neurocomputing*, vol. 508, pp. 293–304, 2021.

[19] H. Xu, G. Ghosh, P.-Y. B. Huang, D. Okhonko, A. Aghajanyan, F. Metze, L. Zettlemoyer, and C. Feichtenhofer, "Videoclip: Contrastive pre-training for zero-shot video-text understanding," in *Conference on Empirical Methods in Natural Language Processing*, 2021.

[20] K. Crowson, S. Biderman, D. Kornis, D. Stander, E. Hallahan, L. Castricato, and E. Raff, "Vqgan-clip: Open domain image generation and editing with natural language guidance," in *European Conference on Computer Vision*, 2022.

[21] K. Frans, L. Soros, and O. Witkowski, "Clipdraw: Exploring text-to-drawing synthesis through language-image encoders," *Advances in Neural Information Processing Systems*, vol. 35, pp. 5207–5218, 2022.

[22] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," *arXiv preprint arXiv:2204.06125*, 2022.

[23] Y. Vinker, E. Pajouheshgar, J. Y. Bo, R. C. Bachmann, A. H. Bermano, D. Cohen-Or, A. Zamir, and A. Shamir, "Clipasso: Semantically-aware object sketching," *ACM Transactions on Graphics (TOG)*, vol. 41, no. 4, pp. 1–11, 2022.

[24] D. Kim, A. Angelova, and W. Kuo, "Region-aware pretraining for open-vocabulary object detection with vision transformers," *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11 144–11 154, 2023.

[25] Y. Zeng, X. Zhang, and H. Li, "Multi-grained vision language pre-training: Aligning texts with visual concepts," in *International Conference on Machine Learning*. PMLR, 2022, pp. 25 994–26 009.

[26] L. Fan, D. Krishnan, P. Isola, D. Katabi, and Y. Tian, "Improving clip training with language rewrites," *Advances in Neural Information Processing Systems*, vol. 36, pp. 35 544–35 575, 2023.

[27] Y. Tian, L. Fan, P. Isola, H. Chang, and D. Krishnan, "Stablerep: Synthetic images from text-to-image models make strong visual representation learners," *Advances in Neural Information Processing Systems*, vol. 36, pp. 48 382–48 402, 2023.

[28] K. Yang, J. Deng, X. An, J. Li, Z. Feng, J. Guo, J. Yang, and T. Liu, "Alip: Adaptive language-image pre-training with synthetic caption," *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2910–2919, 2023.

[29] L. Zhao, K. Zheng, Y. Zheng, D. Zhao, and J. Zhou, "RLEG: Vision-language representation learning with diffusion-based embedding generation," in *Proceedings of the 40th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, Eds., vol. 202. PMLR, 23–29 Jul 2023, pp. 42 247–42 258.

[30] Z. Lai, H. Zhang, W. Wu, H. Bai, A. Timofeev, X. Du, Z. Gan, J. Shan, C.-N. Chuah, Y. Yang, and M. Cao, "Veclip: Improving clip training via visual-enriched captions," in *European Conference on Computer Vision*, 2023.

[31] Y. Liu, K. Wang, W. Shao, P. Luo, Y. Qiao, M. Z. Shou, K. Zhang, and Y. You, "Mllms-augmented visual-language representation learning," *CoRR*, 2023.

[32] Y. Li, X. Hou, D. Zheng, L. Shen, and Z. Zhao, "FLIP-80M: 80 million visual-linguistic pairs for facial language-image pre-training," in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 58–67.

[33] Z. Zhao, Y. Cao, S. Gong, and I. Patras, "Enhancing zero-shot facial expression recognition by LLM knowledge transfer," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2025, pp. 815–824.

[34] L. Sun, X. Jiang, H. Chen, Y. Li, Z. Lian, B. Liu, Y. Zong, W. Zheng, J. M. Leppänen, and G. Zhao, "Learning transferable facial emotion representations from large-scale semantically rich captions," *arXiv preprint arXiv:2507.21015*, 2025.

[35] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio, "Generative adversarial nets," in *Neural Information Processing Systems*, 2014.

[36] S. E. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis," in *International Conference on Machine Learning*, 2016.

[37] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[38] 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," *Advances in neural information processing systems*, vol. 35, pp. 36 479–36 494, 2022.

[39] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.

[40] 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 *International Conference on Learning Representations*, 2021.

[41] A. Hertz, R. Mokady, J. Tenenbaum, K. Aberman, Y. Pritch, and D. Cohen-Or, "Prompt-to-prompt image editing with cross attention control," *arXiv preprint arXiv:2208.01626*, 2022.

[42] B. Kawar, S. Zada, O. Lang, O. Tov, H.-T. Chang, T. Dekel, I. Mosseri, and M. Irani, "Imagic: Text-based real image editing with diffusion models," *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6007–6017, 2022.

[43] J. Gu, Y. Wang, N. Zhao, T.-J. Fu, W. Xiong, Q. Liu, Z. Zhang, H. Zhang, J. Zhang, H. Jung *et al.*, "Photoswap: Personalized subject swapping in images," *Advances in Neural Information Processing Systems*, vol. 36, pp. 35 202–35 217, 2023.

[44] A. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. McGrew, I. Sutskever, and M. Chen, "Glide: Towards photorealistic image generation and editing with text-guided diffusion models," in *International Conference on Machine Learning*, 2021.

[45] O. Avrahami, D. Lischinski, and O. Fried, "Blended diffusion for text-driven editing of natural images," *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18 187–18 197, 2021.

[46] S. Wang, C. Saharia, C. Montgomery, J. Pont-Tuset, S. Noy, S. Pellegrini, Y. Onoe, S. Laszlo, D. J. Fleet, R. Soricut, J. Baldridge, M. Norouzi, P. Anderson, and W. Chan, "Imagen editor and editbench: Advancing and evaluating text-guided image inpainting," *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18 359–18 369, 2022.

[47] O. Bar-Tal, D. Ofri-Amar, R. Fridman, Y. Kasten, and T. Dekel, "Text2live: Text-driven layered image and video editing," in *European conference on computer vision*. Springer, 2022, pp. 707–723.

[48] G. Couairon, J. Verbeek, H. Schwenk, and M. Cord, "Diffedit: Diffusion-based semantic image editing with mask guidance," in *ICLR 2023 (Eleventh International Conference on Learning Representations)*, 2023.

[49] N. Tumanyan, M. Geyer, S. Bagon, and T. Dekel, "Plug-and-play diffusion features for text-driven image-to-image translation," *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1921–1930, 2022.

[50] M. Cao, X. Wang, Z. Qi, Y. Shan, X. Qie, and Y. Zheng, "Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing," *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22 503–22 513, 2023.

[51] J. Yang, J. Feng, W. Luo, D. Lischinski, D. Cohen-Or, and H. Huang, "Emoedit: Evoking emotions through image manipulation," 2025. [Online]. Available: https://arxiv.org/abs/2405.12661

[52] H. Zhao, X. Ma, L. Chen, S. Si, R. Wu, K. An, P. Yu, M. Zhang, Q. Li, and B. Chang, "Ultraedit: Instruction-based fine-grained image editing at scale," *Advances in Neural Information Processing Systems*, vol. 37, pp. 3058–3093, 2024.

[53] D. Demszky, D. Movshovitz-Attias, J. Ko, A. Cowen, G. Nemade, and S. Ravi, "Goemotions: A dataset of fine-grained emotions," *arXiv preprint arXiv:2005.00547*, 2020.

[54] B. Zhang, P. Zhang, X. wen Dong, Y. Zang, and J. Wang, "Long-clip: Unlocking the long-text capability of clip," in *European Conference on Computer Vision*, 2024.

[55] L. Shi, J. Fan, and J. Yan, "Ot-clip: Understanding and generalizing clip via optimal transport," in *Forty-first International Conference on Machine Learning*, 2024.

[56] M. Cuturi, "Sinkhorn distances: Lightspeed computation of optimal transport," in *Advances in Neural Information Processing Systems*, vol. 26, 2013, pp. 2292–2300.

[57] S. Xu, Y. Huang, J. Pan, Z. Ma, and J. Chai, "Inversion-free image editing with natural language," in *Conference on Computer Vision and Pattern Recognition 2024*, 2024.

[58] A. Mathews, L. Xie, and X. He, "Senticap: Generating image descriptions with sentiments," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 30, 2016.

[59] I. Huberman-Spiegelglas, V. B. Kulikov, and T. Michaeli, "An edit friendly ddpm noise space: Inversion and manipulations," *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12 469–12 478, 2023.

[60] Z. Wang, L. Zhao, and W. Xing, "Stylediffusion: Controllable disentangled style transfer via diffusion models," *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7643–7655, 2023.

[61] R. Mokady, A. Hertz, K. Aberman, Y. Pritch, and D. Cohen-Or, "Null-text inversion for editing real images using guided diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 6038–6047.

[62] X. Ju, A. Zeng, Y. Bian, S. Liu, and Q. Xu, "Pnp inversion: Boosting diffusion-based editing with 3 lines of code," in *The Twelfth International Conference on Learning Representations*, 2024.

[63] P. A. Kragel, M. C. Reddan, K. S. LaBar, and T. D. Wager, "Emotion schemas are embedded in the human visual system," *Science Advances*, vol. 5, no. 7, p. eaaw4358, 2019.

[64] S. M. Mohammad, "Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*. Melbourne, Australia: Association for Computational Linguistics, 2018, pp. 174–184.

Jiarong Ye received the B.S. degree in data sciences from The Pennsylvania State University and is pursuing the Ph.D. degree in informatics there. Her research interests include generative AI, computer vision, affective computing, and biomedical data analysis.

Sharon X. Huang received the Ph.D. degree in computer science from Rutgers University in 2006. She is the David Reese Professor of Information Sciences and Technology and Head of the Department of Informatics and Intelligent Systems at The Pennsylvania State University. Her research interests include artificial intelligence, computer vision, biomedical image computing, and machine learning.