# EmoEdit: Evoking Emotions through Image Manipulation

Jingyuan Yang

2025.05.28

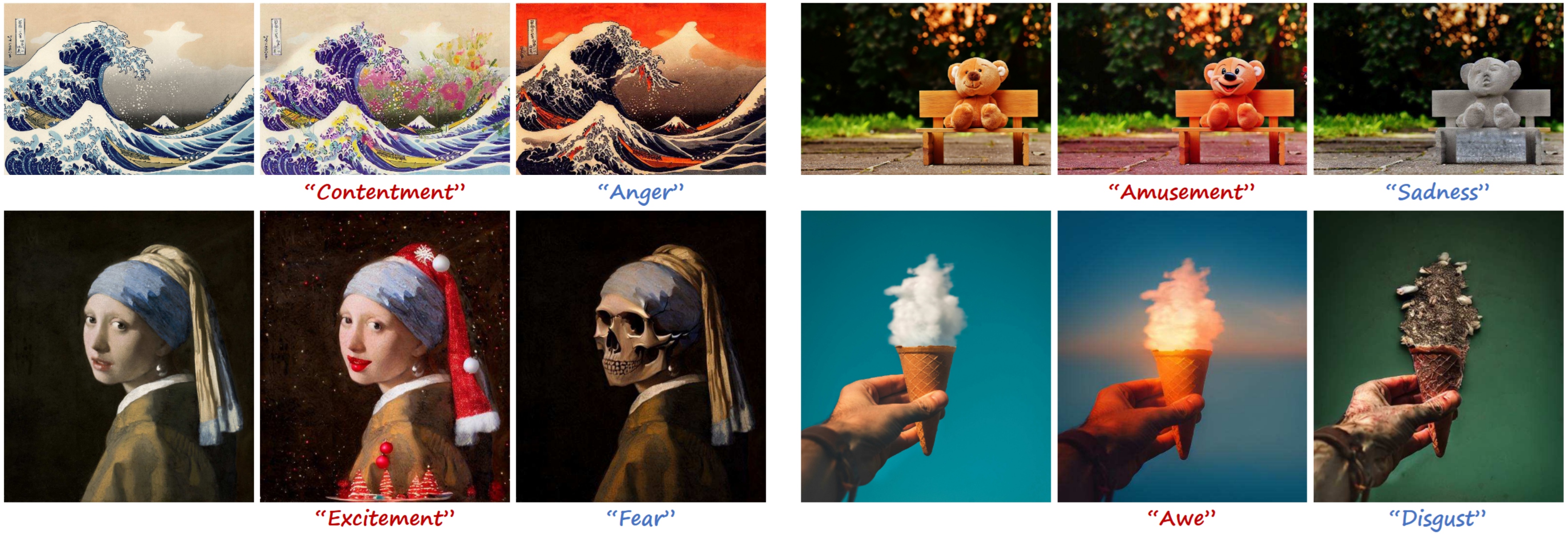
# Overview

- ❑ EmoSet
- ❑ Problem definition
- ❑ Evaluation metrics
- ❑ Possible solutions
  - ❑ SDEdit
  - ❑ Inversion
  - ❑ Textual inversion
  - ❑ Conceptor
  - ❑ Concept decomposition
- ❑ Experiments

# Introduction

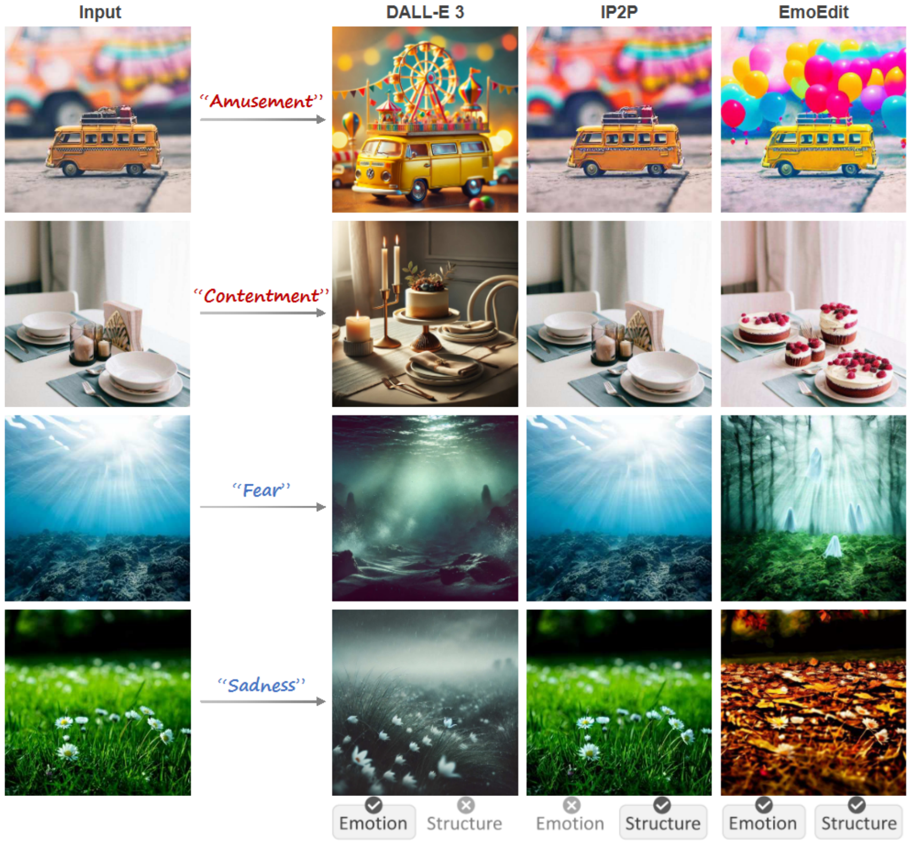
"The emotion expressed by wordless simplicity is the most abundant."

– William Shakespeare



# Introduction

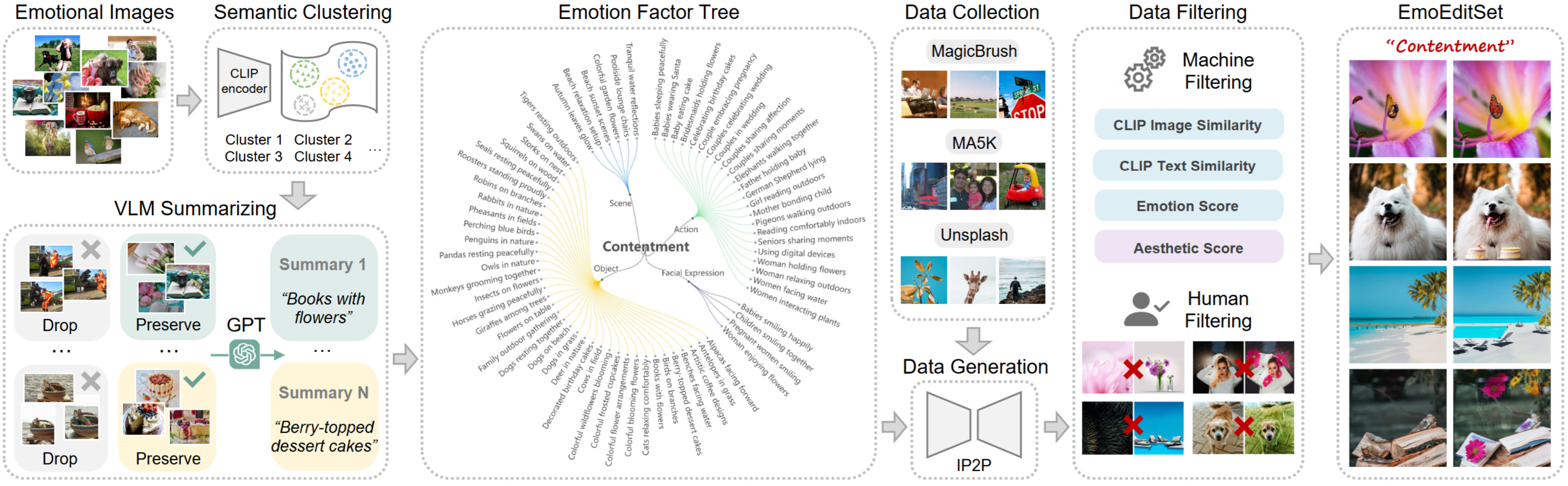
- **Observations**
  - While DALL-E 3 conveys emotions well, IP2P remains faithful to original structure, neither approach satisfies both aspects.
  - EmoEdit fills this gap by creating images with both emotion fidelity and structure preservation.



# Introduction

- **Contributions**
  - **EmoEdit**, a content-aware AIM framework capable of generating emotion-evoking, contextually fitting, and structurally faithful variant of a user-provided image, requiring only emotion words as prompts.
  - **EmoEditSet**, the first large-scale AIM dataset, featuring40,120 image pairs labeled with emotion directions and content instructions, establishing a high-quality, semantically diverse benchmark.
  - **Emotion adapter**, trained with diffusion loss and theproposed instruction loss, functions as a plug-and-play module that enhances generative models with emotion awareness once trained.

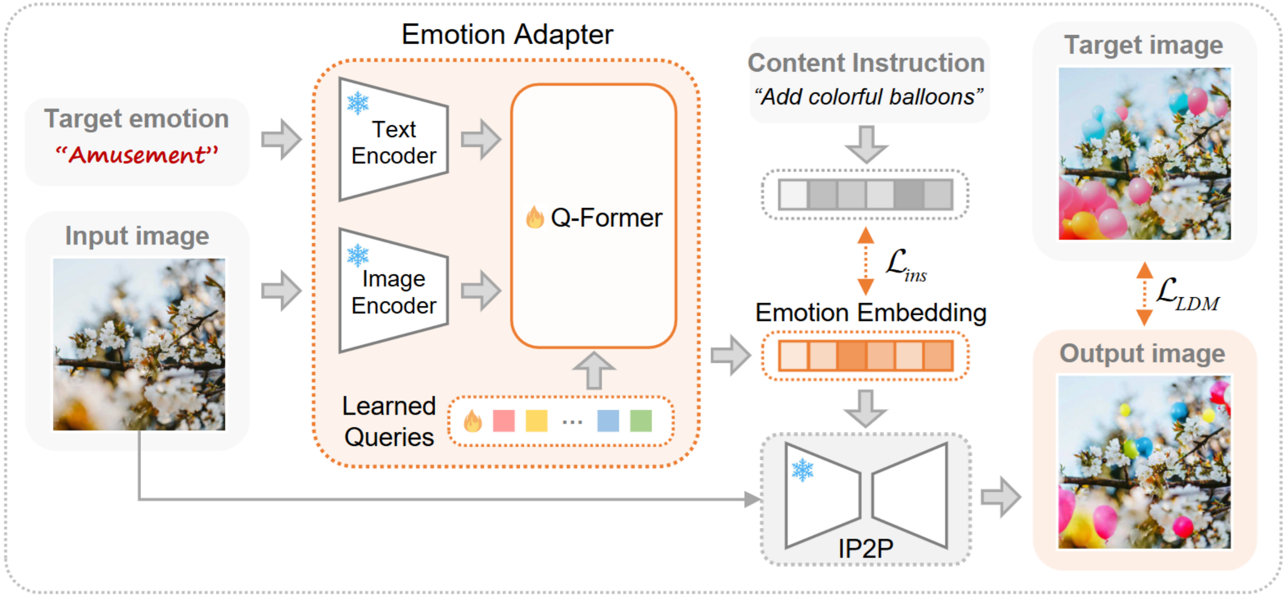
# Overview of EmoEditSet



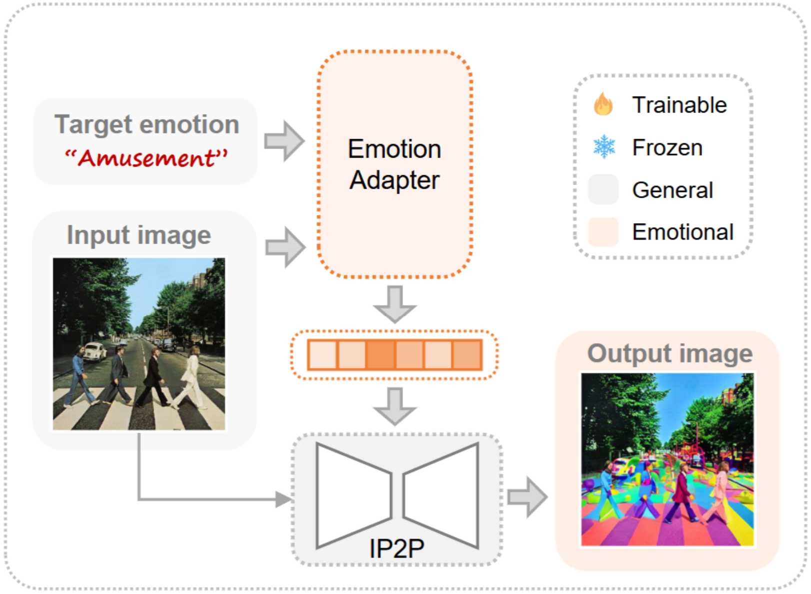
(a) Emotion Attribution

(b) Data Construction

# Overview of EmoEdit



(a) Training Process of EmoEdit



(b) Inference Process of EmoEdit

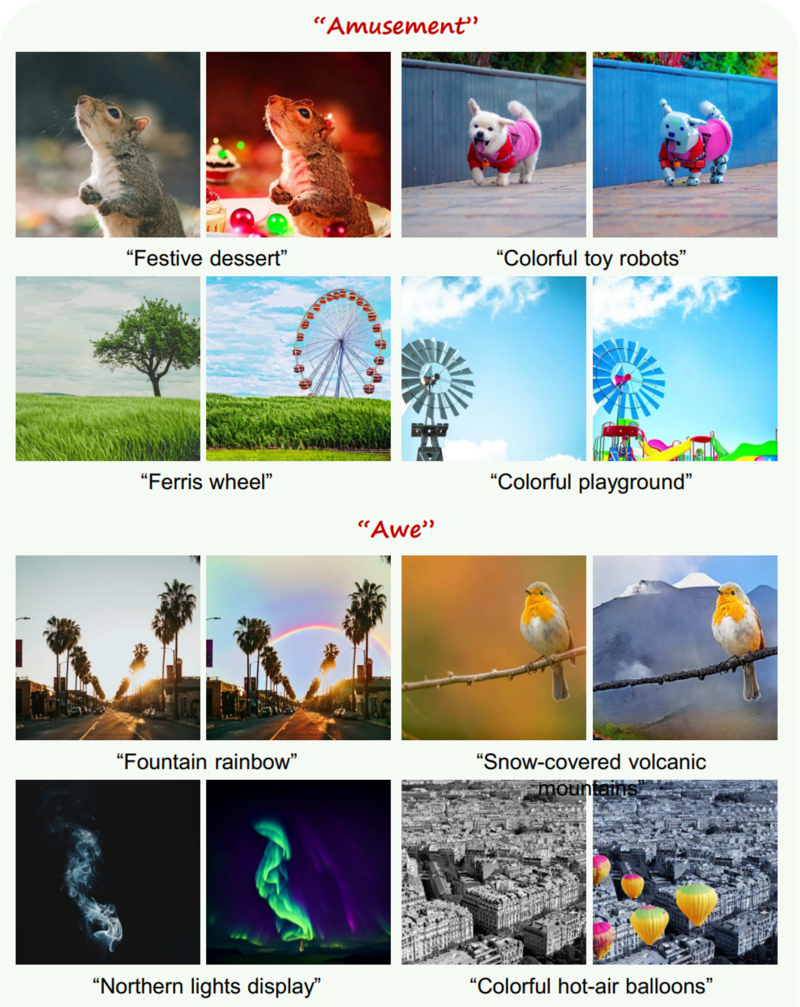
$$A_s = softmax(\frac{[q;e_t]W_q^s([q;e_t]W_k^s)^T}{\sqrt{d_k}})[q;e_t]W_v^s$$

$$A_c = softmax(\frac{A_s W_q^c(e_i W_k^c)^T}{\sqrt{d_k}})e_i W_v^c$$

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

$$\mathcal{L}_{ins} = \frac{1}{M}\|c_e - \mathcal{E}_{txt}\left(t_{ins}\right)\|_2^2$$

# Images in EmoEditSet



“Amusement”

“Festive dessert”

“Colorful toy robots”

“Ferris wheel”

“Colorful playground”

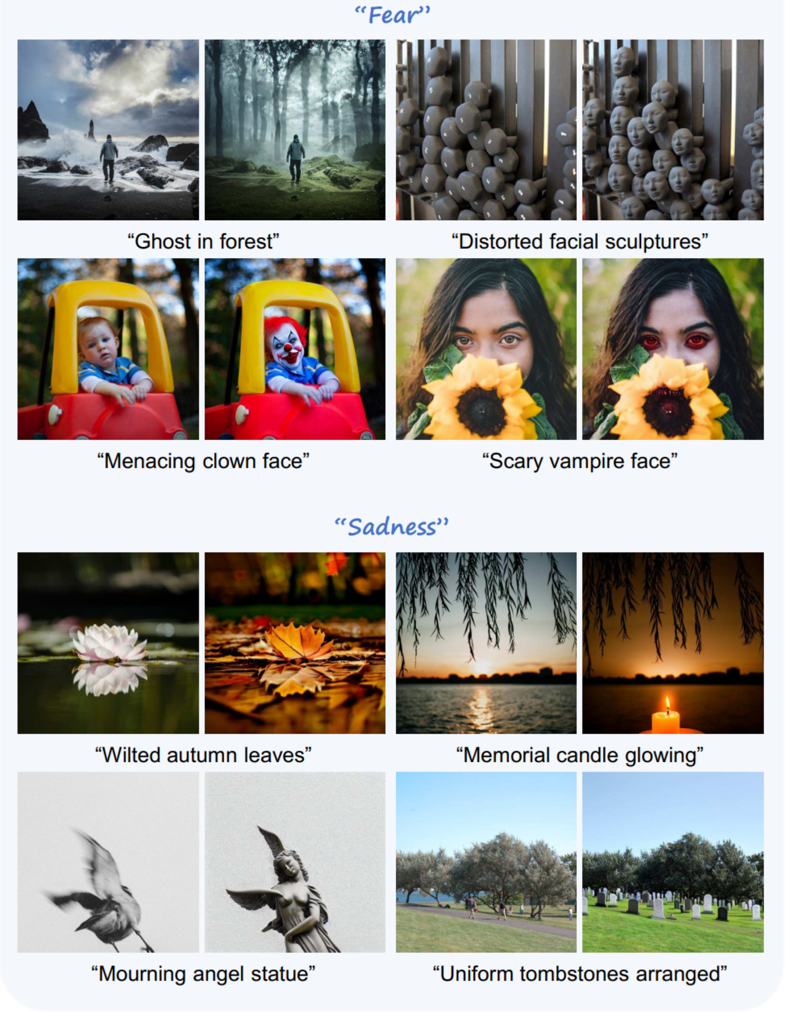
“Awe”

“Fountain rainbow”

“Snow-covered volcanic mountains”

“Northern lights display”

“Colorful hot-air balloons”



“Fear”

“Ghost in forest”

“Distorted facial sculptures”

“Menacing clown face”

“Scary vampire face”

“Sadness”

“Wilted autumn leaves”

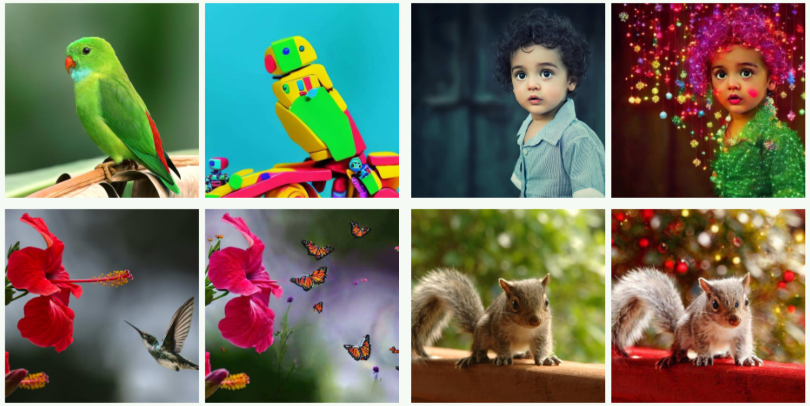
“Memorial candle glowing”

“Mourning angel statue”

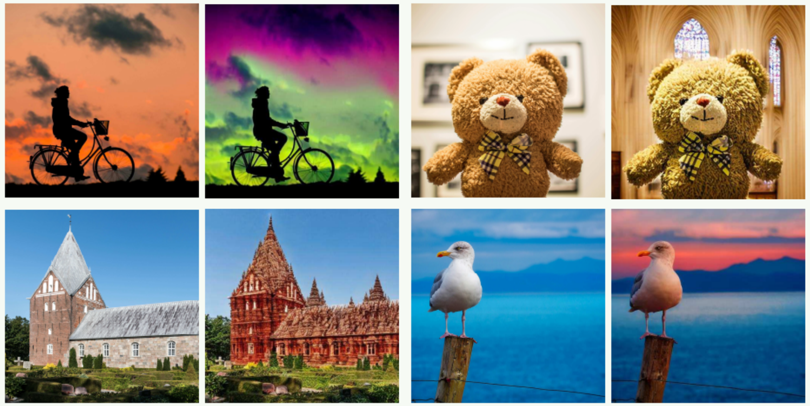
“Uniform tombstones arranged”

# Images generated by EmoEdit

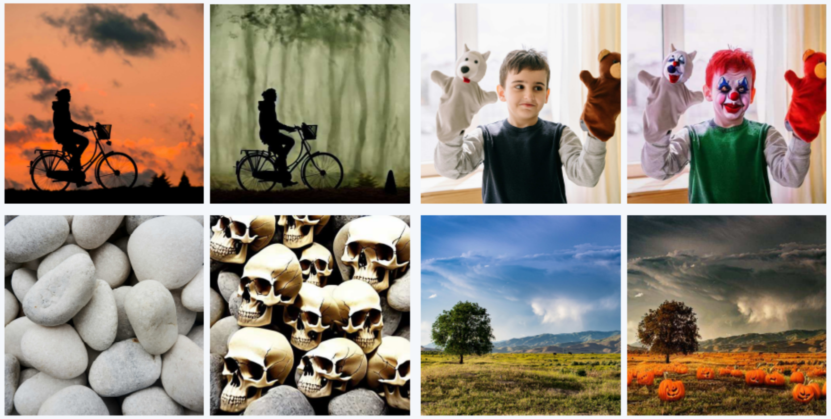
"Amusement"



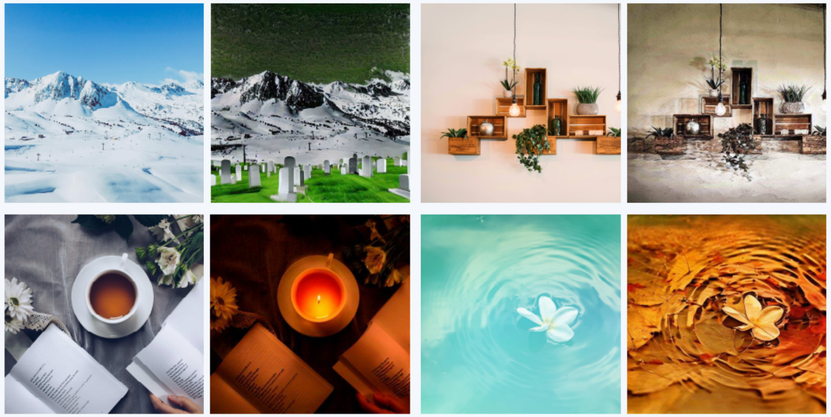
"Awe"



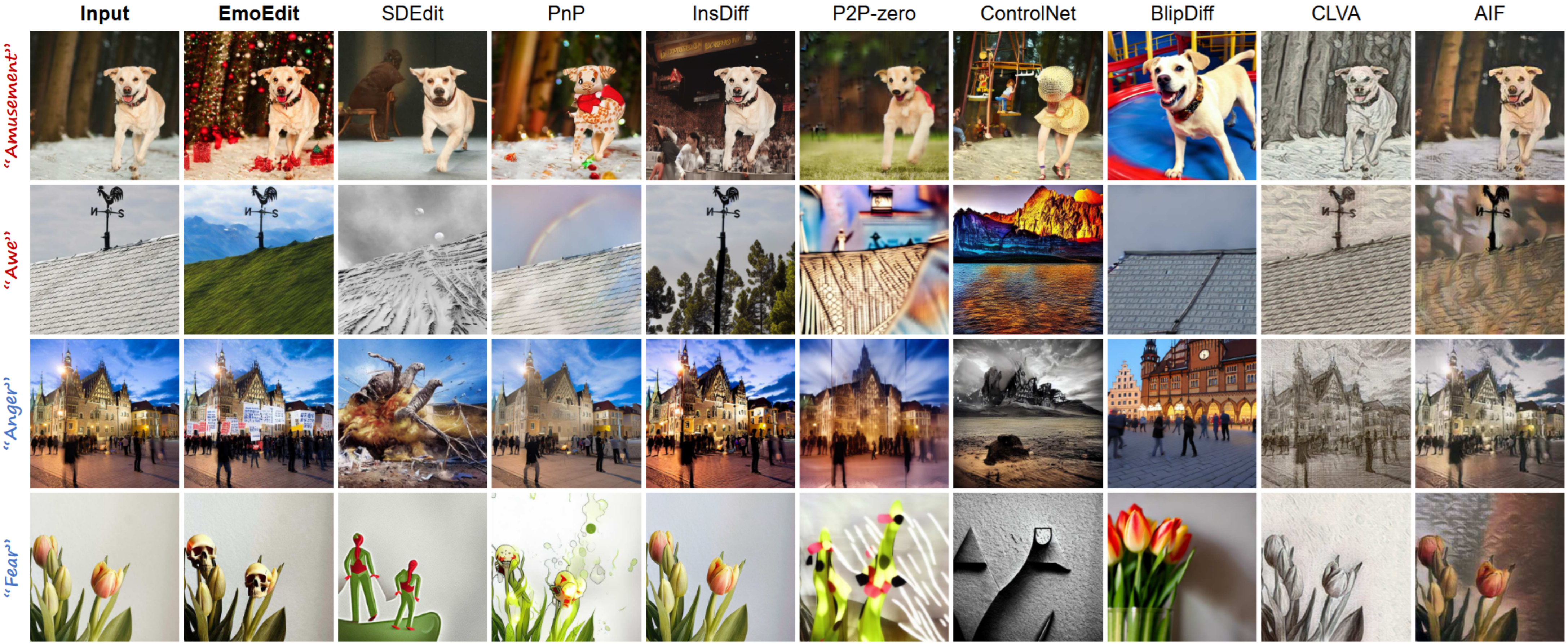
"Fear"



"Sadness"



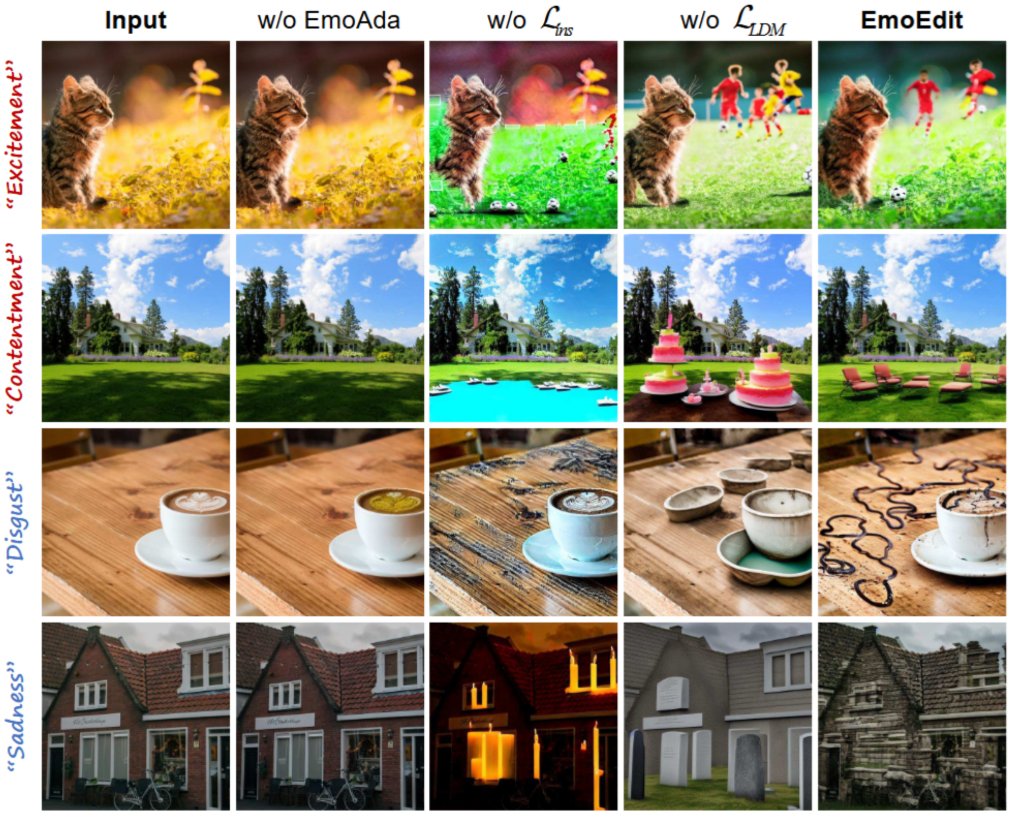
# Results


Input
EmoEdit
SDEdit
PnP
InsDiff
P2P-zero
ControlNet
BlipDiff
CLVA
AIF
"Amusement"
"Awe"
"Anger"
"Fear"

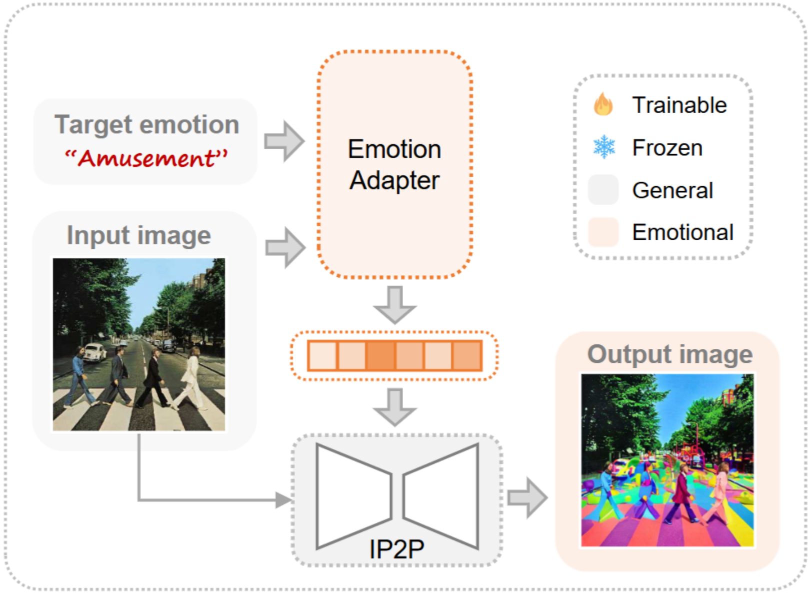
# Results

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

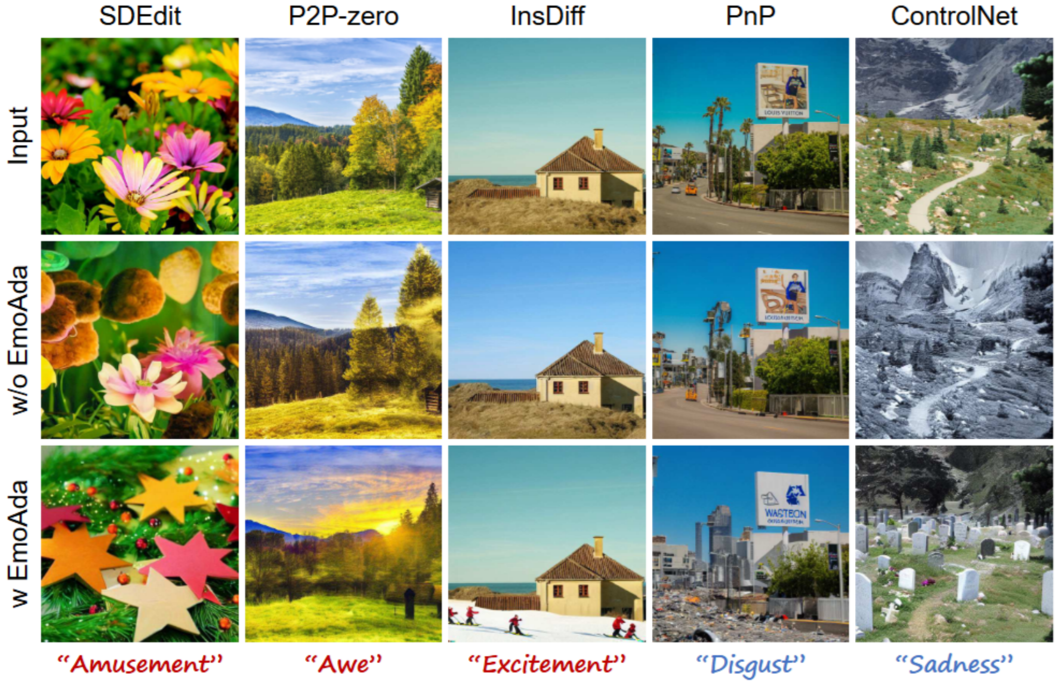
| Method | PSNR ↑ | SSIM ↑ | LPIPS ↓ | CLIP-I ↑ | Emo-A ↑ | Emo-S ↑ |
|---|---|---|---|---|---|---|
| SDEdit [20] | 15.43 | 0.415 | 0.459 | 0.638 | 38.21% | 0.221 |
| PnP [37] | 14.41 | 0.436 | 0.381 | **0.851** | 23.83% | 0.095 |
| InsDiff [8] | 10.75 | 0.318 | 0.505 | 0.796 | 19.22% | 0.060 |
| P2P-Zero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



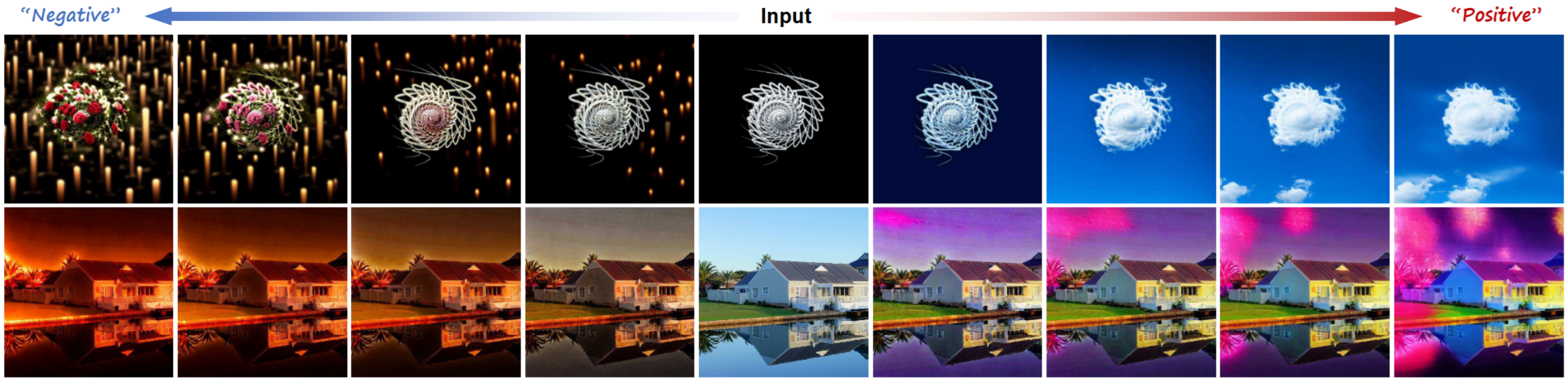
# Results

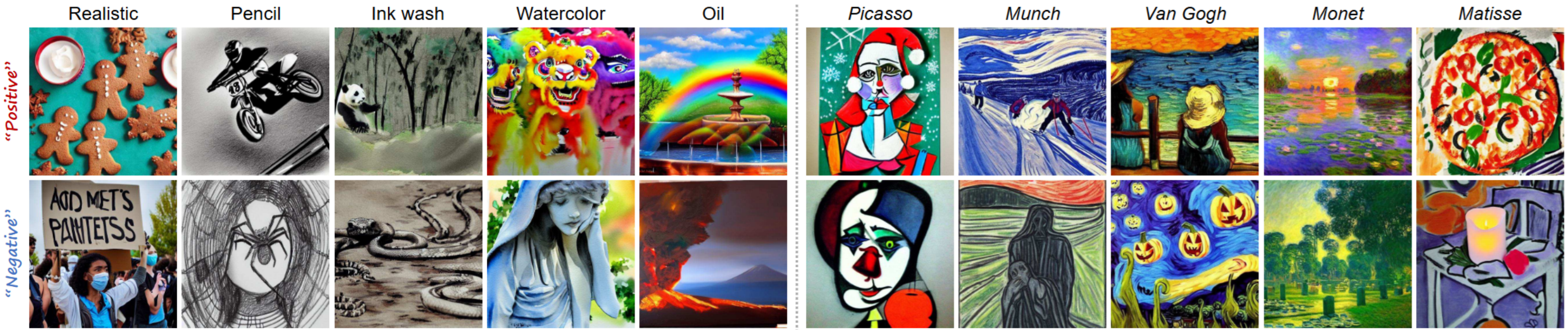


(b) Inference Process of EmoEdit



# Results





(a) Art Mediums

(b) Representative Artists

# Thank you!

Hui Huang

Jiawei Feng

Weibin Luo

Dani Lischinski

Daniel Cohen-Or