Supplementary Material for AIGIEmo

Anonymous Author(s)

1 Extended Background and Related Work

1.1 Affective Image Datasets

Affective image datasets have been widely used in psychology, human-computer interaction (HCI), computer vision (CV), and affective computing to study emotion recognition in images.

In psychology and HCI, Lang et al. introduced the International Affective Picture System (IAPS) [13, 14], a widely used dataset for assessing emotional responses to visual stimuli. To refine its applicability, Mikels et al. [24] developed IAPSa, a subset of IAPS categorized into discrete emotions such as happiness, sadness, and fear. Other datasets have since expanded the range and quality of affective stimuli. The Geneva Affective Picture Database (GAPED) [6] focused on threat-related images, while the Nencki Affective Picture System (NAPS) [22] introduced high-resolution images with validated emotional ratings. The Maastricht Affective Picture System (MAPS) [18] further enhanced stimulus diversity, improving ecological validity. More recently, the Open Affective Standardized Image Set (OASIS) [12] provided an open-access alternative, and the DIsgust-RelaTed-Images (DIRTI) database [8] specifically targeted disgust-eliciting images.

In CV and affective computing, various datasets have been developed for affective image analysis. Machajdik and Hanbury [21] introduced ArtPhoto and AbstractPhoto, focusing on artistic and synthetic images labeled with discrete emotions. Borth et al. [5] proposed VSO, linking images with sentiment concepts, while You et al. [42, 43] introduced Flickr I, Twitter I, and Twitter II for large-scale social media sentiment analysis. Other datasets, such as Emotion6 [27], FI [44], T4SA [32], and WEBEmo [26], further expanded emotion recognition in online images. In the news domain, Event [17] and EMOd [11] focused on emotion in journalistic content, while ArtEmis [1] explored emotions in paintings. EmoSet [41] provided large-scale annotated images for visual emotion analysis. More recently, FindingEmo [23] introduced 25K images for emotion recognition in complex social scenes, expanding beyond face-centered datasets with valence, arousal, and categorical emotion labels.

Recent work has also begun to examine affect in AI-generated content. Representative examples include EmoConveyance [20], AIGI-VC [31], LAI-GAI [3], and EmoArt [46]. These studies further highlight the need for dedicated affective resources tailored to AI-generated images.

In summary, psychology-driven VEA datasets are typically small, focusing on user studies with physiological data, categorical emotion labels, and valence-arousal-dominance scores. Meanwhile, VEA datasets in CV and affective computing are much larger, often relying on a combination of machine-generated weak labels and human annotations to classify emotions in images. Our work extends VEA to AIGI images, introducing a large-scale dataset with rich annotations that provide new insights into AI-generated content.

1.2 Datasets of AI-Generated Images

With the rapid progress of text-to-image models, several datasets have been developed to support research on AI-generated content. DiffusionDB [33] provides images generated by Stable Diffusion with corresponding prompts. TWIGMA [2] collects AIGI from Twitter, including metadata such as timestamps and engagement metrics. JourneyDB [30] offers images generated by Midjourney, paired with textual prompts. GenImage [51] includes AI-generated and real image pairs, serving as a benchmark for generative model detection. Artifact [29] contains both real and AIGI for evaluating AI-generated content. AIGIQA-20K [16] focuses on subjective quality assessment, featuring AIGI with human ratings. ImageReward [35] supports preference modeling and reward learning for generated images. Additionally, CIFAKE [4] and WildFake [9] provide benchmarks for synthetic-image detection in more realistic settings.

These datasets support downstream tasks such as image generation quality assessment, AI-generated content detection, prompt optimization, and related applications. However, they are not specifically designed for visual emotion analysis. In contrast, AIGIEmo is built for affective understanding and provides multi-level annotations to study emotional responses to AI-generated images.

1.3 Visual Emotion Analysis

Deep learning has significantly advanced VEA by enabling automatic extraction of emotional cues from images. Early models primarily relied on Convolutional Neural Networks (CNNs) to learn emotion representations from visual data [28, 45]. Later approaches incorporated spatial and contextual information, such as combining CNNs with Recurrent Neural Networks (RNNs) to capture sequential dependencies [52], and integrating attention mechanisms to refine feature extraction [25, 38]. More recent work has explored multi-level attention networks to better understand complex emotional expressions in images [36]. These models have improved recognition accuracy by leveraging contextual elements such as facial expressions, body posture, and scene composition.

Recent advances in Vision-Language Models (VLMs) and Multimodal Large Language Models (MLLMs) have introduced new possibilities for VEA by integrating textual information alongside visual inputs. Large-scale VLMs have been shown to enhance emotion recognition by leveraging descriptions and contextual cues [7, 15]. Visual prompting techniques further refine emotion predictions by aligning textual and image-based representations [48]. Additionally, specialized models, such as an emotion-aware vision-language framework for art analysis [47] and EmoVIT, a visual instruction-tuned model designed to improve emotion insights in images [34], demonstrate the potential of multimodal approaches. By incorporating both visual and textual features, these models provide a more comprehensive understanding of emotions in images, enabling not only recognition but also emotion attribution and contextual analysis.

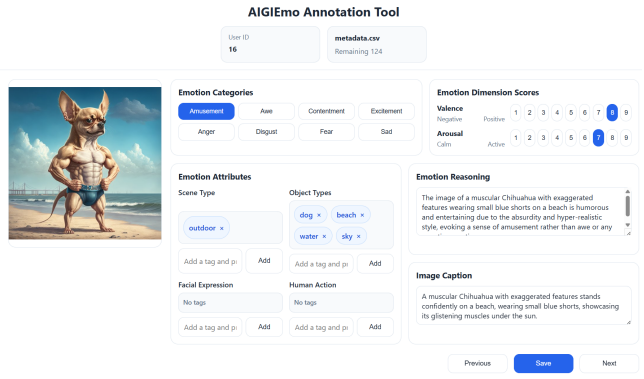

Figure 1: Web-based annotation tool for AIGIEmo.

2 AIGIEmo Dataset Details

2.1 Licensing and Access

AIGIEmo is released under the Creative Commons Attribution 4.0 International License (CC BY 4.0). The dataset is publicly available on Hugging Face and can be downloaded directly without any access request or approval: https://huggingface.co/datasets/dongSHE/AIGIEmo.

2.2 Data Governance, Privacy, and Redistribution

AIGIEmo is collected from public text-to-image channels with explicit prompt–image correspondence. The release follows source-platform and content-governance constraints in public research-oriented data sharing. We exclude unsafe content and non-text-only workflows during curation, and the released dataset is designed to reduce privacy risks by focusing on public content and research-oriented data records. Users are encouraged to adhere to ethical guidelines when using AIGIEmo, especially when analyzing sensitive content or generating derivative works. Redistribution of the dataset should maintain the original licensing terms and provide appropriate attribution to the creators.

2.3 Annotation Tool

Figure 1 shows the web-based annotation interface used for AIGIEmo. The tool supports categorical emotion selection, valence–arousal scoring with the Self-Assessment Manikin (SAM), auxiliary label editing, and sequential revision of caption and reasoning fields.

2.4 Additional Data Analysis

Beyond the statistics shown in the main paper, AIGIEmo also supports additional analysis of affective structure, style and generator bias, and the differences between the filtered pool and the final balanced subset. These analyses help further characterize the representativeness and limitations of the dataset.

2.4.1 Affective Structure in V-A Space. The eight emotion categories exhibit a clear and well-organized affective structure in the valence–arousal space. As shown in Figure 2, *Anger*, *Fear*, and *Disgust* concentrate in the low-valence, relatively high-arousal region,

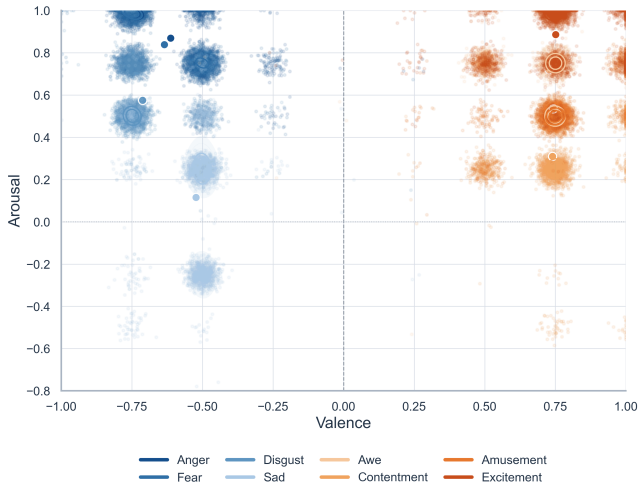

Figure 2: KDE-enhanced scatter plot of the eight emotion categories in the valence–arousal space.

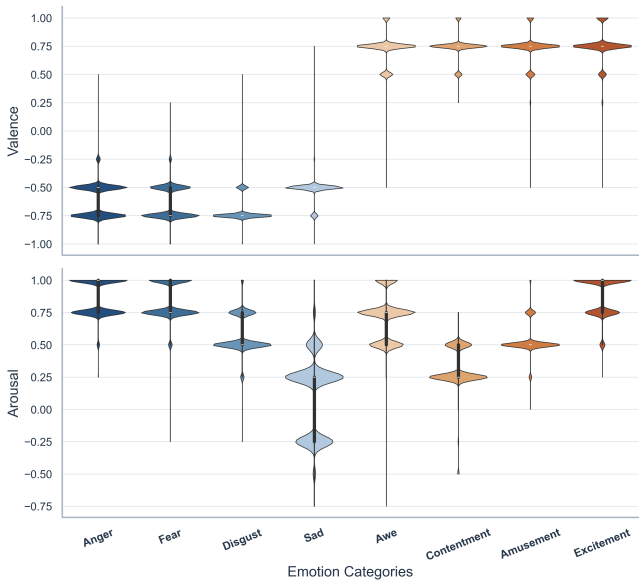

Figure 3: Marginal violin plots of valence and arousal for each emotion category.

while *Sadness* remains low in both valence and arousal. In contrast, *Excitement* occupies the high-valence, high-arousal region, and *Contentment* shifts toward high valence but lower arousal. *Amusement* and *Awe* lie between these extremes, with positive valence and moderate-to-high arousal.

Most overlap occurs between affectively adjacent categories rather than across opposite affective regions. As shown in Figure 3, the positive emotions share relatively high valence but differ in arousal, while the negative emotions are more consistently separated by valence polarity and activation level. This pattern further

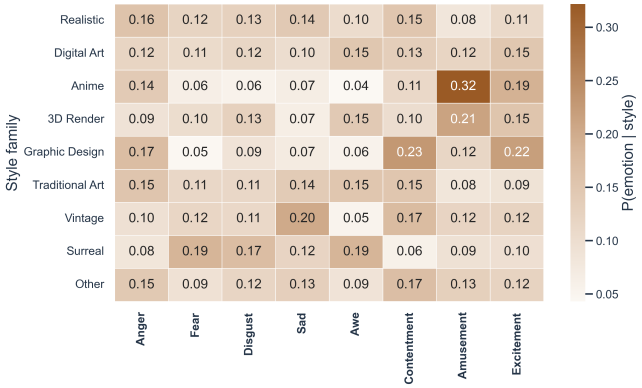

Figure 4: Conditional emotion distribution within each style family.

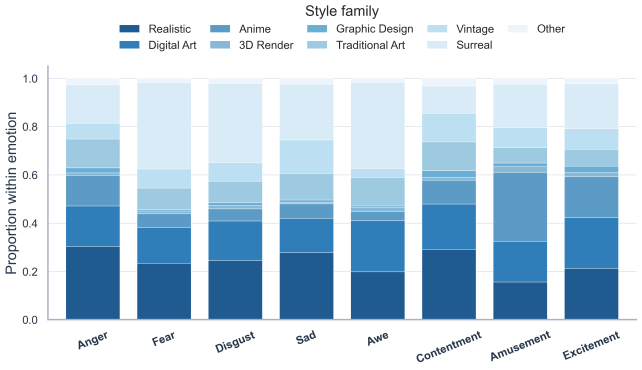

Figure 5: Style-family composition within each emotion category.

confirms the consistency between the categorical labels and the dimensional annotations.

2.4.2 Style Family and Emotion Distribution. Style-family annotations reveal meaningful visual diversity across emotion categories. As shown in Figure 4, *Anime* is strongly concentrated on *Amusement* and *Excitement*, *Graphic Design* is relatively more common in *Contentment* and *Excitement*, and *Vintage* has a noticeable concentration on *Sadness*. In contrast, *Realistic*, *Traditional Art*, and *Other* exhibit more balanced emotion distributions.

The style composition of each emotion category remains broad and diverse, while still showing meaningful variation across emotions. As shown in Figure 5, *Realistic* and *Digital Art* remain dominant in most categories, while the proportions of *Anime*, *Surreal*, and *Graphic Design* change noticeably, especially for *Amusement*, *Fear*, and *Contentment*. This result shows that AIGIEmo covers multiple visual styles while preserving informative style–emotion associations.

2.4.3 Color and Brightness Distributions. The low-level attribute analysis in the main paper uses effect-size-based association strength.

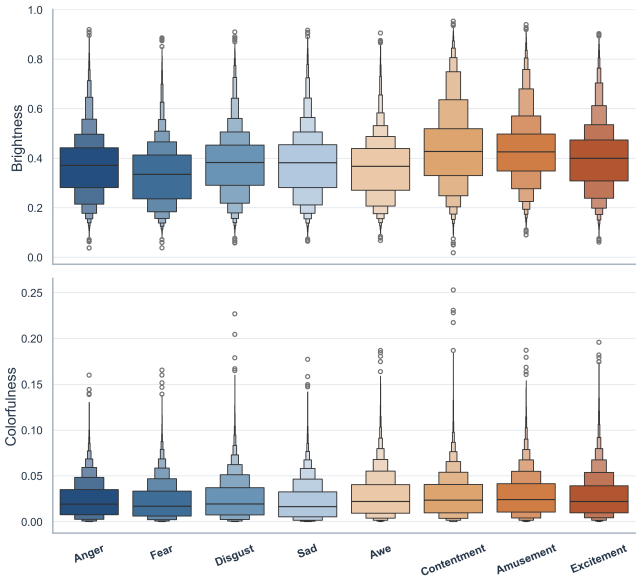

Figure 6: Brightness and colorfulness distributions across emotion categories.

For continuous attributes such as brightness, contrast, and colorfulness, we report the Kruskal–Wallis effect size

$$\epsilon^2 = \frac{H - k + 1}{n - k}, \tag{1}$$

where H is the Kruskal–Wallis statistic, k is the number of emotion categories, and n is the number of samples. For categorical attributes, we use Cramér's V,

$$V = \sqrt{\frac{\chi^2}{n \cdot \min(r-1, c-1)}}, \tag{2}$$

where χ^2 is the chi-square statistic and r, c are the numbers of rows and columns in the contingency table.

Low-level color cues show weaker association than semantic attributes, but they still exhibit stable group-level patterns. As shown in Figure 6, positive emotions such as *Contentment*, *Amusement*, and *Excitement* tend to have slightly higher brightness and colorfulness, whereas negative emotions are relatively darker and less colorful overall. At the same time, the substantial overlap across categories is also consistent with the main-paper result that low-level visual features alone provide limited affective discrimination.

The dominant hue composition remains diverse across all categories while showing interpretable differences at the group level. As shown in Figure 7, warm hues, especially *Orange*, occupy the largest proportion in most emotions, while *Blue* and *Cyan* provide the main complementary tones. The relative shares of cool and warm hues vary across emotions, which further suggests that color contributes useful but secondary cues compared with higher-level semantic attributes.

2.4.4 Semantic Attribute Enrichment. High-level semantic attributes show clear and interpretable enrichment patterns across emotion categories. We quantify semantic enrichment using lift, defined

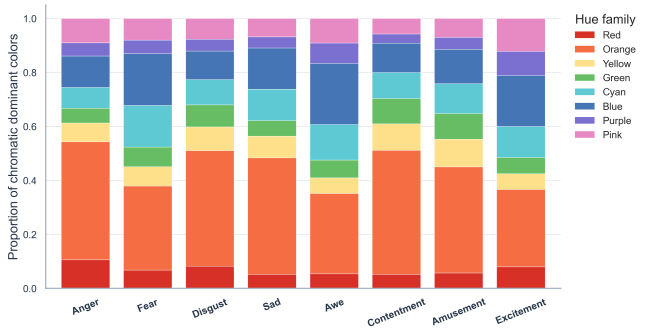

Figure 7: Hue-family composition across emotion categories.

Figure 8: Lift-score heatmap for facial expressions across emotion categories.

as the ratio between the attribute frequency within an emotion category and its global frequency in the dataset:

$$\text{Lift}(a, e) = \frac{P(a \mid e)}{P(a)} = \frac{n(a, e)/N(e)}{n(a)/N}, \tag{3}$$

where a denotes a semantic attribute and e denotes an emotion category. A lift score larger than 1 indicates that the attribute is more prevalent in that emotion than expected from its global frequency.

Facial expressions exhibit the strongest and most concentrated semantic cues. As shown in Figure 8, *Fear* is strongly enriched by fearful expressions, *Sadness* by sad expressions, *Awe* and *Surprise* are closely aligned, and positive emotions such as *Contentment*, *Amusement*, and *Excitement* are associated with happy expressions. This pattern highlights the quality and interpretability of the facial-expression annotations in AIGIEmo.

Human actions also show meaningful emotion-dependent enrichment. As shown in Figure 9, *fighting* is most enriched for *Anger*, *running* for *Fear*, *hugging* for *Contentment*, and dynamic activities such as *dancing* and *playing* are more prominent for *Excitement* and *Amusement*. These patterns indicate that action annotations provide useful behavioral cues for affective understanding.

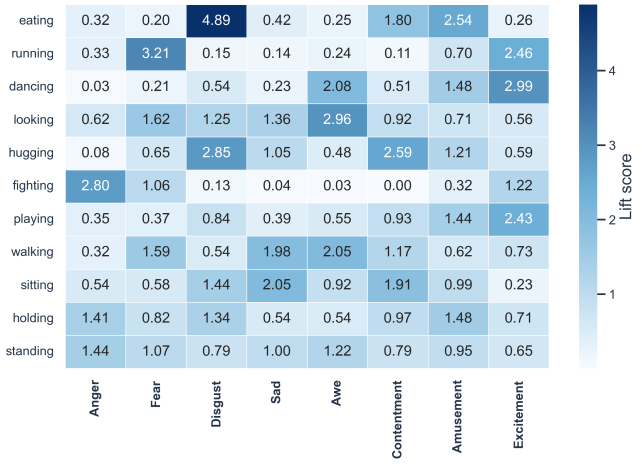

Figure 9: Lift-score heatmap for human actions across emotion categories.

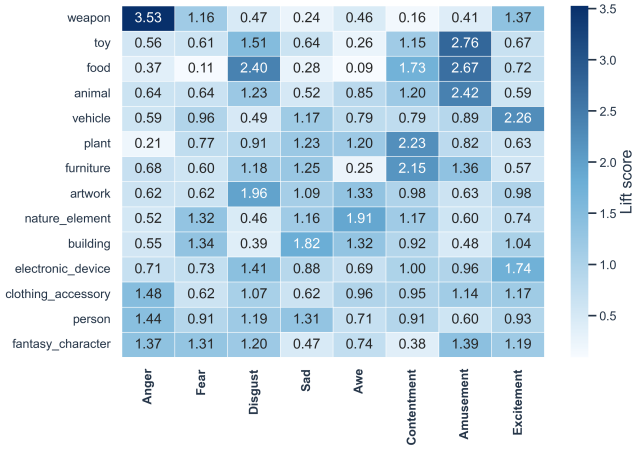

Figure 10: Lift-score heatmap for object categories across emotion categories.

Object categories further reveal diverse semantic grounding for different emotions. As shown in Figure 10, *weapon* is most enriched for *Anger*, *food* for *Disgust*, *toy* and *animal* for *Amusement*, and *vehicle* for *Excitement*. At the same time, *plant* and *furniture* are relatively more common in *Contentment*, reflecting calmer and more pleasant contexts.

Scene-level semantics remain broader than facial or action cues, but still show coherent enrichment trends. As shown in Figure 11, *fantasy_virtual* is strongly enriched for *Awe*, *studio* for *Anger*, and *outdoor_urban* for *Excitement*. These results complement the main-paper finding that scene cues are weaker than human-centered attributes, while still contributing useful contextual information.

Figure 11: Lift-score heatmap for scene categories across emotion categories.

3 Experimental Details and Additional Experiments

3.1 Experiment Setup

All baseline benchmark experiments are conducted on the balanced 160K subset of AIGIEmo, which is split into 128K/16K/16K images for training, validation, and testing. Exact duplicate prompts and images are removed before splitting, and all in-domain experiments use the same split unless otherwise noted. The visual benchmark covers emotion classification and valence–arousal prediction, while the VLM benchmark further includes caption generation and reason-for-emotion generation.

For visual models, implementations are based on PyTorch and trained on NVIDIA A800-SXM4-80GB GPUs. Model selection is performed on the validation split with early stopping, and final results are reported on the test split. For emotion classification, we report Accuracy and Macro-F1; for valence–arousal regression, we report MAE and Pearson correlation; and for text generation, we report BLEU, CIDEr, SPICE, and BERTScore. For VLM fine-tuning, LoRA-based experiments are run on a single NVIDIA A800-SXM4-80GB GPU with the same train/validation/test partition, so the full benchmark remains comparable across visual and vision–language settings.

3.2 Visual Model Implementation Details

The proposed baseline visual model follows the architecture shown in Figure 3 of the main paper. It contains four main components: a visual backbone, a shared neck, two task-specific branches for emotion classification and valence–arousal regression, and an auxiliary Attribute Module. The visual backbone extracts a global image representation from the input image, and the shared neck projects it into a common feature space used by all downstream branches. We instantiate this framework with six representative backbones, including VGG16, ResNet50, DenseNet121, ConvNeXt-T, Swin-T, and ViT-B/16, so that the benchmark covers both CNN-based and Transformer-based visual encoders.

Let x denote an input image. The backbone produces a visual representation $z = B(x)$, which is then mapped by the shared neck to a common feature $h = N(z)$. Two lightweight task adapters are applied to h to obtain task-specific features for emotion classification and V-A regression:

$$f_{\text{emo}} = A_{\text{emo}}(h), \qquad f_{\text{va}} = A_{\text{va}}(h). \tag{4}$$

The emotion branch uses a linear classifier on top of f_{emo} to predict the 8-way emotion label, while the V-A branch uses a regression head to predict valence and arousal scores. In parallel, the Attribute Module learns an auxiliary representation

$$f_{\text{att}} = E_{\text{att}}(h), \tag{5}$$

which is supervised by aggregated attribute annotations. Following the association analysis in the main paper, we use four auxiliary attribute groups with relatively strong emotion associations: facial expression, human action, scene, and object.

The attribute feature is then injected into the two main branches through two linear gating modules. This design follows the general idea of sigmoid-based feature reweighting and task-specific gating used in prior architectures [10, 19]. Specifically, the attribute feature is first projected into the corresponding task space, and a sigmoid gate controls how much auxiliary information is fused into each branch:

$$g_1 = \sigma(W_1 f_{\text{att}} + b_1), \qquad \tilde{f}_{\text{emo}} = f_{\text{emo}} + g_1 \odot P_1(f_{\text{att}}), \tag{6}$$

$$g_2 = \sigma(W_2 f_{\text{att}} + b_2), \qquad \tilde{f}_{\text{va}} = f_{\text{va}} + g_2 \odot P_2(f_{\text{att}}), \tag{7}$$

where $P_1(\cdot)$ and $P_2(\cdot)$ denote linear projections and $\odot$ is element-wise multiplication. The final emotion classifier and V-A regressor operate on $\tilde{f}_{\text{emo}}$ and $\tilde{f}_{\text{va}}$, respectively. This design keeps the model lightweight while allowing the main tasks to selectively use auxiliary attribute cues.

All branches are trained end-to-end with joint supervision from emotion labels, V-A scores, and attribute annotations. During inference, attribute cues are predicted from the image by the Attribute Module itself rather than injected from human annotations, so the full model remains a purely image-driven predictor at test time. In addition to the proposed baseline with different backbones, we also report several prior VEA-specific models, including WSCNet [40], StyleNet [49], PDANet [50], Stimuli-Aware [39], and MDAN [37], to provide a broader and practically useful benchmark suite for AIGI emotion understanding.

3.3 Vision–Language Model Prompts and Settings

For vision–language model evaluation, we use fixed task prompts for both zero-shot and fine-tuned settings. Emotion labels are decoded by exact matching, and valence–arousal outputs are parsed as constrained numeric responses in $[-1, 1]$. For caption and reason-for-emotion generation, we use task-specific prompts aligned with the corresponding benchmark definitions.

Fine-tuned evaluation is implemented with LoRA on the training split and reported on the test split. Our first-stage multimodal fine-tuning is conducted with LLaMA-Factory, using a unified JSON generation prompt that jointly supervises emotion, valence, arousal, caption, and reason generation. The system prompt is summarized below.

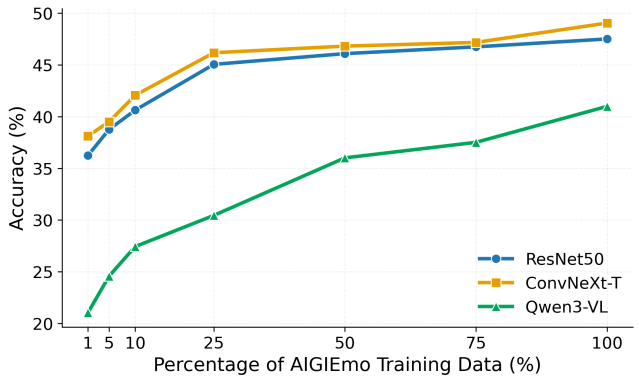

Figure 12: Data-mixing results evaluated on the AIGIEmo test set.

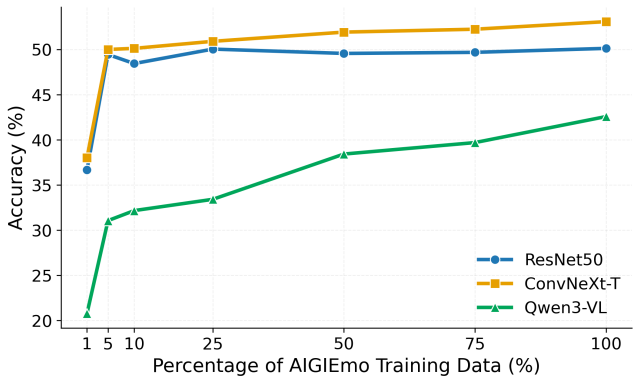

Figure 13: Data-mixing results evaluated on AIGI-VC [31].

System prompt. You are a careful multimodal annotation assistant. Analyze the image and return only valid JSON.

Requirements.

(1) Choose exactly one emotion label from {Amusement, Anger, Awe, Contentment, Disgust, Excitement, Fear, Sad}.
(2) Predict valence as a float in $[-1, 1]$.
(3) Predict arousal as a float in $[-1, 1]$.
(4) Generate a concise English caption with at most 20 words.
(5) Generate a short reason with at most 50 words.

Output schema.

```
{"emotion": "string", "valence": 0.0, "arousal":
0.0, "caption": "string", "reason": "string"}
```

At evaluation time, the generated JSON is parsed field by field. The emotion output is normalized to the benchmark label space, including mapping `Sad` to `Sadness`, while valence and arousal are clipped to the valid range. Caption and reason outputs are evaluated directly against the corresponding reference annotations.

3.4 Extended Data-Mixing Evaluation for AIGI Emotion Recognition

We further examine the contribution of AIGIEmo under a controlled data-mixing setting. Specifically, we use EmoSet [41] as the default base training set and progressively add different proportions of the AIGIEmo training split, while keeping the model architecture and optimization protocol unchanged. This setting evaluates whether AIGIEmo provides task-specific supervision beyond a strong natural-image emotion dataset.

We report two complementary test settings. The first uses the AIGIEmo test split, which measures in-domain improvement on our benchmark. The second uses AIGI-VC [31], a recent AIGI dataset with emotion labels, which measures whether the benefit of AIGIEmo also transfers to another AI-generated image benchmark.

AIGIEmo consistently improves in-domain emotion recognition as more AIGIEmo data is added to training. As shown in Figure 12, all three models, ResNet50, ConvNeXt-T, and Qwen3-VL, exhibit steady accuracy gains when the AIGIEmo proportion increases from 1% to 100%. The improvements are especially clear for Qwen3-VL, indicating that AIGIEmo provides effective supervision not only for conventional visual backbones but also for multimodal models.

The same trend remains visible when testing on an external AIGI dataset. As shown in Figure 13, increasing the amount of AIGIEmo training data also improves performance on AIGI-VC, although the gains are more moderate than those on the AIGIEmo test split. This result is consistent with the expected dataset gap across different AIGI datasets, while still showing that AIGIEmo contributes useful and transferable supervision for emotion recognition on AI-generated images.

Table 1: Cross-dataset transfer between AIGIEmo and AIGI-VC.

Train \ Test	AIGIEmo	AIGI-VC
AIGIEmo	0.581	**0.620**
AIGI-VC	**0.495**	0.592

Table 2: Cross-dataset transfer between AIGIEmo and EmoSet.

Train \ Test	AIGIEmo	EmoSet
AIGIEmo	0.581	**0.426**
EmoSet	**0.412**	0.782

3.5 Cross-Dataset Transfer

We further evaluate cross-dataset transfer to measure how well emotion supervision learned from AIGIEmo generalizes across different benchmarks. Table 1 summarizes transfer results between AIGIEmo and AIGI-VC [31], while Table 2 reports the corresponding comparison with EmoSet [41], all under the aligned 8-way setting.

AIGIEmo provides strong and transferable supervision for AIGI emotion recognition. As shown in Table 1, training on AIGIEmo

transfers effectively to AIGI-VC and even slightly exceeds the AIGI-VC in-domain result, indicating that AIGIEmo captures robust AIGI-specific affective patterns. In contrast, training on AIGI-VC transfers less effectively to AIGIEmo, which highlights the stronger benchmark coverage of AIGIEmo for AI-generated image emotion understanding.

The comparison with EmoSet in Table 2 further shows that AIGIEmo remains more targeted to the AIGI setting while preserving reasonable cross-domain generalization. Overall, these results support the role of AIGIEmo as both a strong in-domain benchmark and a practically useful source of transferable supervision for emotion recognition on AI-generated images.

4 Potential Applications

4.1 AIGI Emotion Recognition

AIGIEmo supports supervised learning and benchmark evaluation for categorical emotion classification and dimensional affect prediction on AI-generated images. It can be used to study how current visual models and multimodal systems understand affect in prompt-driven visual generation.

4.2 Affective Captioning and Reasoning

Because AIGIEmo provides human-finalized captions and emotion reasoning, it can support research on affective caption generation, emotion explanation, and language-grounded affective understanding for AI-generated images.

4.3 Multimodal AIGI Understanding

The combination of prompts, images, auxiliary attributes, captions, and reasoning makes AIGIEmo suitable for multimodal understanding tasks. These include prompt-aware emotion prediction, visual–textual affect alignment, and human-centered interpretation of generated content.

4.4 Data-Centric Benchmarking and Alignment Research

AIGIEmo can also support research on dataset bias, human–AI affective agreement, prompt-driven emotion control, and evaluation of AIGI systems from a human-centered perspective. As AI-generated content becomes more widely used, such applications may become increasingly important for affective computing and multimedia research.

References

[1] Panos Achlioptas, Maks Ovsjanikov, Leonidas Guibas, and Niloy J. Mitra. 2021. ArtEmis: Affective language for visual art. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 21461–21471.

[2] Anonymous. 2023. TWIGMA: A Large-Scale Dataset of AI-Generated Images from Twitter. In *NeurIPS 2023 Datasets and Benchmarks Track*.

[3] M. Behnke, M. Kłoskowski, M. Klichowski, et al. 2026. Using Artificial Intelligence to Generate Affective Images: Methodology and Initial Library. *Advances in Methods and Practices in Psychological Science* 9, 1 (2026). doi:10.1177/25152459251415336

[4] Jordan J. Bird. 2023. *CIFAKE: Real and AI-Generated Synthetic Images*. https://www.kaggle.com/datasets/birdy654/cifake-real-and-ai-generated-synthetic-images

[5] Damian Borth, Rongrong Ji, Tao Chen, Thomas Breuel, and Shih-Fu Chang. 2013. Large-scale visual sentiment ontology and detectors using adjective noun pairs. In *Proceedings of the 21st ACM international conference on Multimedia*. 223–232.

[6] Stefan Dan-Glauser and Klaus R. Scherer. 2011. The Geneva affective picture database (GAPED): A new 730-picture database focusing on valence and normative significance. *Behavior Research Methods* 43, 2 (2011), 468–477.

[7] Yasaman Etesam, Özge Nilay Yalçın, Chuxuan Zhang, and Angelica Lim. 2024. Contextual Emotion Recognition Using Large Vision Language Models. *arXiv preprint arXiv:2405.08992* (2024).

[8] Anke Haberkamp, Julia A. Glombiewski, Florian Schmidt, and Antonia Barke. 2017. The DIsgust-RelaTed-Images (DIRTI) database: Validation of a novel standardized picture set of disgust eliciting images. *Behavior Research Methods* 49, 2 (2017), 2061–2072.

[9] Yan Hong and Jianfu Zhang. 2024. WildFake: A Large-scale Challenging Dataset for AI-Generated Images Detection. *arXiv preprint arXiv:2402.11843* (2024).

[10] Jie Hu, Li Shen, and Gang Sun. 2018. Squeeze-and-Excitation Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7132–7141.

[11] Ron Kosti, Jose M Alvarez, Adria Recasens, and Agata Lapedriza. 2019. EMOd: A dataset for emotion analysis in images. *IEEE Transactions on Affective Computing* 10, 1 (2019), 85–99.

[12] Benedek Kurdi, Sergio Lozano, and Mahzarin R. Banaji. 2017. Introducing the Open Affective Standardized Image Set (OASIS). *Behavior Research Methods* 49, 2 (2017), 457–470.

[13] Peter J. Lang, Margaret M. Bradley, and Bruce N. Cuthbert. 1997. *International Affective Picture System (IAPS): Technical Manual and Affective Ratings*. Technical Report. NIMH Center for the Study of Emotion and Attention, University of Florida.

[14] Peter J Lang, Margaret M Bradley, Bruce N Cuthbert, et al. 2005. *International affective picture system (IAPS): Affective ratings of pictures and instruction manual*. NIMH, Center for the Study of Emotion & Attention Gainesville, FL.

[15] Yuxuan Lei, Dingkang Yang, Zhaoyu Chen, Jiawei Chen, Peng Zhai, and Lihua Zhang. 2024. Large Vision-Language Models as Emotion Recognizers in Context Awareness. In *Proceedings of the Asian Conference on Machine Learning (ACML)*.

[16] Chunyi Li, Tengchuan Kou, Yixuan Gao, Yuqin Cao, Wei Sun, Zicheng Zhang, Yingjie Zhou, Zhichao Zhang, Weixia Zhang, Haoning Wu, et al. 2024. AIGIQA-20K: A Large Database for AI-Generated Image Quality Assessment. *arXiv preprint arXiv:2404.03407* (2024).

[17] Jianshu Lu, Zhe Lin, Hailin Jin, Jianchao Yang, and James Z Wang. 2016. Visual emotion analysis for event-centric affective understanding. *IEEE Transactions on Affective Computing* 7, 4 (2016), 453–466.

[18] Benedikt Lugrin, Tim Hildebrandt, and Sabrina C. Eimler. 2016. The Maastricht Affective Picture System (MAPS): A new standardized stimulus set for research on emotion and affect. *Journal of Psychophysiology* 30, 2 (2016), 77–90.

[19] Jiaqi Ma, Zhe Zhao, Xinyang Yi, Jilin Chen, Lichan Hong, and Ed H. Chi. 2018. Modeling Task Relationships in Multi-task Learning with Multi-gate Mixture-of-Experts. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 1930–1939.

[20] Lin Ma, Dengkai Chen, Yuan Feng, Xinggang Hou, and Jing Chen. 2024. Emotional Conveyance Analysis of Artificial Intelligence Painting. In *Proceedings of the Eleventh International Symposium of Chinese CHI* (Denpasar, Bali, Indonesia) *(CHCHI '23)*. Association for Computing Machinery, New York, NY, USA, 44–54. doi:10.1145/3629606.3629612

[21] Jana Machajdik and Allan Hanbury. 2010. Affective image classification using features inspired by psychology and art theory. In *Proceedings of the 18th ACM international conference on Multimedia*. 83–92.

[22] Artur Marchewka, Łukasz Żurawski, Katarzyna Jednoróg, and Anna Grabowska. 2014. The Nencki Affective Picture System (NAPS): Introduction to a novel, standardized, wide-range, high-quality, realistic picture database. *Behavior Research Methods* 46, 2 (2014), 596–610.

[23] Laurent Mertens, Elahe Yargholi, Hans Op de Beeck, Jan Van den Stock, and Joost Vennekens. 2024. FindingEmo: An Image Dataset for Emotion Recognition in the Wild. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*. https://openreview.net/forum?id=1q3b2Z95ec

[24] Joseph A. Mikels, Barbara L. Fredrickson, Gregory R. Larkin, Casey M. Lindberg, Sam J. Maglio, and Patricia A. Reuter-Lorenz. 2005. Emotional category data on images from the International Affective Picture System. *Behavior Research Methods* 37, 4 (2005), 626–630.

[25] Trisha Mittal, Pooja Guhan, Uttaran Bhattacharya, Rohan Chandra, Aniket Bera, and Dinesh Manocha. 2020. EmotiCon: Context-Aware Multimodal Emotion Recognition Using Frege's Principle. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[26] Rameswar Panda, Jianming Zhang, Haoxiang Li, Joon-Young Lee, Xin Lu, and Amit K Roy-Chowdhury. 2018. Contemplating Visual Emotions: Understanding and Overcoming Dataset Bias. In *European Conference on Computer Vision*.

[27] Kuan-Chuan Peng, Tsuhan Chen, Amir Sadovnik, and Andrew C. Gallagher. 2015. A Mixed Bag of Emotions: Model, Predict, and Transfer Emotion Distributions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[28] Kuan-Chuan Peng, Tsuhan Chen, Amir Sadovnik, and Andrew C. Gallagher. 2015. A Mixed Bag of Emotions: Model, Predict, and Transfer Emotion Distributions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 860–868.

[29] Md Awsafur Rahman, Bishmoy Paul, Najibul Haque Sarker, Zaber Ibn Abdul Hakim, and Shaikh Anowarul Fattah. 2023. Artifact: A Large-Scale Dataset With Artificial And Factual Images For Generalizable And Robust Synthetic Image Detection. In *2023 IEEE International Conference on Image Processing (ICIP)*. 2200–2204. doi:10.1109/ICIP49359.2023.10222083

[30] Keqiang Sun, Jialiang Zhang, Shiguang Wang, and Changsheng Xu. 2023. JourneyDB: A Benchmark for Generative Image Understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1–10.

[31] Yu Tian, Yixuan Li, Baoliang Chen, Hanwei Zhu, Shiqi Wang, and Sam Kwong. 2025. AI-generated image quality assessment in visual communication. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence (AAAI'25/IAAI'25/EAAI'25)*. AAAI Press, Article 822, 9 pages. doi:10.1609/aaai.v39i7.32795

[32] Lucia Vadicamo, Andrea Cimino, Stefano Cresci, Felice Dell'Orletta, and Maurizio Tesconi. 2017. T4SA: A dataset for topic-based sentiment analysis in the wild. In *Proceedings of the 25th ACM international conference on Multimedia*. 1795–1803.

[33] Zijie J. Wang, Evan Montoya, David Munechika, Haoyang Yang, Benjamin Hoover, and Duen Horng Chau. 2023. DiffusionDB: A Large-scale Prompt Gallery Dataset for Text-to-Image Generative Models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 893–911. doi:10.18653/v1/2023.acl-long.51

[34] Hongxia Xie, Chu-Jun Peng, Yu-Wen Tseng, Hung-Jen Chen, Chan-Feng Hsu, Hong-Han Shuai, and Wen-Huang Cheng. 2024. EmoVIT: Revolutionizing Emotion Insights with Visual Instruction Tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[35] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. 2023. ImageReward: learning and evaluating human preferences for text-to-image generation. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*. 15903–15935.

[36] Liwen Xu, Zhengtao Wang, Bin Wu, and Simon Lui. 2022. MDAN: Multi-Level Dependent Attention Network for Visual Emotion Analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 9479–9488.

[37] Liwen Xu, Zhengtao Wang, Bin Wu, and Simon Lui. 2022. MDAN: Multi-level Dependent Attention Network for Visual Emotion Analysis. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 9469–9478. doi:10.1109/CVPR52688.2022.00926

[38] Jingyuan Yang, Jie Li, Xiumei Wang, Yuxuan Ding, and Xinbo Gao. 2021. Stimuli-Aware Visual Emotion Analysis. *IEEE Transactions on Image Processing* 30 (2021), 7432–7445.

[39] Jingyuan Yang, Jie Li, Xiumei Wang, Yuxuan Ding, and Xinbo Gao. 2021. Stimuli-Aware Visual Emotion Analysis. *Trans. Img. Proc.* 30 (Jan. 2021), 7432–7445. doi:10.1109/TIP.2021.3106813

[40] Jufeng Yang, Dongyu She, Yu-Kun Lai, Paul L. Rosin, and Ming-Hsuan Yang. 2018. Weakly Supervised Coupled Networks for Visual Sentiment Analysis. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7584–7592. doi:10.1109/CVPR.2018.00791

[41] Jingyuan Yang, Jialiang Zhang, Shiguang Wang, and Changsheng Xu. 2023. EmoSet: A Large-Scale Visual Emotion Dataset with Rich Attributes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

[42] Quanzeng You, Hailin Jin, and Jiebo Luo. 2016. Building a large scale dataset for image emotion recognition: The fine print and the benchmark. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence*.

[43] Quanzeng You, Jiebo Luo, Hailin Jin, and Jianchao Yang. 2015. Robust image sentiment analysis using progressively trained and domain transferred deep networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 29.

[44] Quanzeng You, Jiebo Luo, Hailin Jin, and Jianchao Yang. 2016. Building a large scale dataset for image emotion recognition: the fine print and the benchmark. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence* (Phoenix, Arizona) *(AAAI'16)*. AAAI Press, 308–314.

[45] Quanzeng You, Jiebo Luo, Hailin Jin, and Jianchao Yang. 2016. Building a Large Scale Dataset for Image Emotion Recognition: The Fine Print and The Benchmark. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI)*. 308–314.

[46] Cheng Zhang, Hongxia Xie, Bin Wen, Songhan Zuo, Ruoxuan Zhang, and Wen-Huang Cheng. 2025. EmoArt: A Multidimensional Dataset for Emotion-Aware Artistic Generation. In *Proceedings of the 33rd ACM International Conference on Multimedia* (Dublin, Ireland) *(MM '25)*. Association for Computing Machinery, New York, NY, USA, 12644–12650. doi:10.1145/3746027.3758201

[47] Jing Zhang, Liang Zheng, Meng Wang, and Dan Guo. 2024. Training a Small Emotional Vision Language Model for Visual Art Comprehension. In *Computer Vision – ECCV 2024 (18th European Conference on Computer Vision) (Lecture Notes in Computer Science)*. Springer.

[48] Qixuan Zhang, Zhifeng Wang, Dylan Zhang, Wenjia Niu, Sabrina Caldwell, Tom Gedeon, Yang Liu, and Zhenyue Qin. 2024. Visual Prompting in LLMs for Enhancing Emotion Recognition. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 4484–4499.

[49] Wei Zhang, Xuanyu He, and Weizhi Lu. 2020. Exploring Discriminative Representations for Image Emotion Recognition With CNNs. *IEEE Transactions on Multimedia* 22, 2 (2020), 515–523. doi:10.1109/TMM.2019.2928998

[50] Sicheng Zhao, Zizhou Jia, Hui Chen, Leida Li, Guiguang Ding, and Kurt Keutzer. 2019. PDANet: Polarity-consistent Deep Attention Network for Fine-grained Visual Emotion Regression. In *Proceedings of the 27th ACM International Conference on Multimedia* (Nice, France) *(MM '19)*. Association for Computing Machinery, New York, NY, USA, 192–201. doi:10.1145/3343031.3351062

[51] Mingjian Zhu, Hanting Chen, Qiangyu Yan, Xudong Huang, Guanyu Lin, Wei Li, Zhijun Tu, Hailin Hu, Jie Hu, and Yunhe Wang. 2023. GenImage: A Million-Scale Benchmark for Detecting AI-Generated Image. *arXiv preprint arXiv:2306.08571* (2023).

[52] Xinge Zhu, Liang Li, Weigang Zhang, Tianrong Rao, Min Xu, Qingming Huang, and Dong Xu. 2017. Dependency Exploitation: A Unified CNN-RNN Approach for Visual Emotion Recognition. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI)*. 3595–3601.