Signal Processing: Image Communication 148 (2026) 117658



Contents lists available at ScienceDirect

Signal Processing: Image Communication

journal homepage: www.elsevier.com/locate/image





Beyond illusions of competence: Revisiting zero-shot learning in emotion recognition with FEAr dataset

Zhong Ken Hew [a], Lai-Kuan Wong [b], Chee Seng Chan [a,*]

[a] Faculty of Computer Science and Information Technology, Universiti Malaya, Kuala Lumpur, Malaysia
[b] Multimedia University, Cyberjaya, Malaysia

ARTICLE INFO

Keywords:
Zero-shot learning
Image emotion recognition
LVLMs

ABSTRACT

This paper investigates the efficacy of Zero-Shot Learning (ZSL) in the domain of emotion recognition, uncovering three significant shortcomings: (i) deceptive performance, (ii) a lack of standardized metrics and (iii) a deficiency in robust emotion semantic attributes. Our analysis shows that current ZSL methods suffer from misleading performance evaluations due to inconsistent dataset splits, which inflate results and obscure the true capabilities of ZSL models in emotion recognition. Moreover, the absence of standardized evaluation metrics exacerbates this issue, making it challenging to compare and validate the effectiveness of different ZSL models accurately. Moreover, the lack of robust semantic attributes for emotions hampers ZSL performance in emotion recognition, making it difficult to align these attributes with visual features. To address these challenges, we propose a new framework, namely ***Fair Evaluation for Affective Recognition*** (***FEAr***) alongside the FEAr-accuracy Harmonization (FaH) evaluation metric as innovative solutions to these problems. The ***FEAr*** framework establishes standardized, valence-balanced split configurations for benchmark emotion datasets, enabling consistent and reproducible evaluation in zero-shot learning. Meanwhile, the FaH metric offers a robust framework for evaluating ZSL models, enabling reliable cross-study comparisons. Furthermore, the CLIP-Affective Attributes Generator (CLIP-AAG) is proposed to generate emotional semantic attributes that are more effectively aligned with the corresponding emotional visual information. This study aims to foster a paradigm shift towards transparent, comparable, and replicable research practices in the field of emotion recognition. The dataset and code associated with this work are available at https://github.com/HewZK2000/CLIP-AAG.git

1. Introduction

In the digital era, social networks have become the principal arena for individuals to express their emotions, often through sharing images and videos that capture moments from their lives. This trend has captured the attention of researchers in computational visual emotion recognition, owing to its diverse applications across domains such as mental health diagnostics, sentiment analysis, and precision marketing. Numerous methodologies [1–3] were developed and trained to discern fundamental emotions. The term "fundamental emotions" refers to a set of basic emotional states namely, happiness, sadness, anger, fear, surprise, and disgust that are universally acknowledged and exhibited by humans, transcending cultural barriers. This notion, deeply ingrained in the psychological frameworks of scholars such as Paul Ekman, posits that these emotional states are universally expressed and perceived. Consequently, this understanding has been instrumental in establishing foundational computational models that detect and categorize emotional expressions in digital media.

As the field of psychology continues to expand the lexicon of emotions, the traditional computational models [1–3] have increasingly encountered limitations. These models often require retraining to incorporate recognition of newly identified emotional states, a process that is both time-intensive and resource-intensive. In response to these challenges, researchers have pivoted towards Zero-Shot Learning (ZSL) as a promising alternative. ZSL is designed to identify classes it has not encountered during its training phase [4], offering a potential solution to the rigidity of traditional models. For instance, studies [5] have explored the utilization of Large Visual Language Models (LVLMs) such as CLIP [6] and CoOP [7], which have shown markedly superior performance in recognizing a broader spectrum of emotions without the need for retraining. However, the transition to ZSL has unveiled pivotal challenges: notably, (i) **deceptive performance**, (ii) **the absence of standardized metrics**, and (iii) **deficiency in robust emotion semantic attributes**.

* Corresponding author.
E-mail address: cs.chan@um.edu.my (C.S. Chan).

https://doi.org/10.1016/j.image.2026.117658
Received 12 December 2024; Received in revised form 12 February 2026; Accepted 13 July 2026
Available online 25 July 2026
0923-5965/© 2026 Elsevier B.V. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

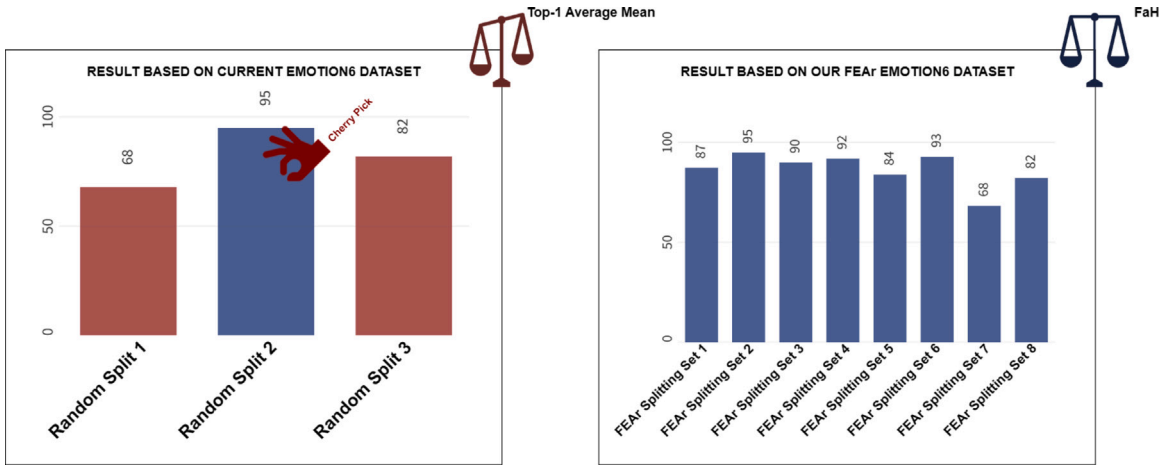


Fig. 1. Left: A selectively reported, non-comprehensive comparison of prior results on the Emotion6 dataset. Right: Emotion6 performance under FEAr settings, providing a more comprehensive evaluation of zero-shot learning models.

First, the emergence of *deceptive performance* within the literature highlights a concerning trend: the portrayal of a model's capabilities is often narrowly tailored to specific configurations, painting an incomplete picture of its true efficacy. As illustrated on the left side of Fig. 1, significant variability arises when employing diverse random seen-unseen splits during model training. Presently, many studies tend to favor the split yielding the highest performance for their reported outcomes. However, this practice disregards a notable 27% decline in performance evident with alternative split configurations. This inconsistency in splits underscores a significant challenge and raises concerns about the selective reporting of results to artificially boost performance metrics.

Second, this issue is further exacerbated by the reliance on evaluation metrics that fail to capture the models' holistic performance. Current evaluation metrics, such as Average per-class top-1 accuracy, harmonic mean, and average accuracy, are commonly used to evaluate ZSL model performance. However, as depicted in Fig. 1, our results reveal significant performance variations across different seen-unseen splits. This suggests that these metrics may not adequately capture the model's true robustness, as they reflect its effectiveness only within specific seen-unseen configurations. These findings indicate the need for more comprehensive metrics to evaluate model performance consistently across varying conditions. As a result, this scenario presents significant obstacles to researchers benchmarking their innovations against these precedents. The lack of standardized split settings across studies undermines the ability to conduct meaningful comparisons, as the performance outcomes are inherently tied to the chosen configurations. Without a unified evaluation framework, the purported success of existing methods remains questionable, casting doubt on their reported effectiveness and impeding progress in the field. As a whole, these issues highlight the pressing need for continued innovation and the establishment of standardized practices within the domain of emotion recognition using ZSL approaches.

Third, the semantic attributes for emotions are not effectively constructed. ZSL relies on semantic attributes to recognize unseen classes by aligning image embeddings with the embeddings of these attributes. However, the existing semantic attributes, generated either from NLP models such as Word2Vec [8] or from CLIP's text encoder [6], lack the quality needed for effective alignment with the visual features of emotions. Word2Vec, for instance, learns embeddings from word co-occurrence, producing vectors that overlook the attributes needed to capture the visual details of emotions. Furthermore, using simple phrases such as "image of sadness" limits CLIP's ability to generate high-quality word embeddings. This highlights that both approaches fail to produce semantic attributes of sufficient quality for emotional representation.

To dismantle the barrier of **Deceptive Performance** and foster a reproducible research environment, we introduce the **Fair Evaluation for Affective Recognition** (***FEAr***). Unlike static dataset releases, ***FEAr*** is established as a dataset-agnostic benchmarking standard designed for universal application across emotion ZSL tasks. It standardizes the experimental input via two distinct sparsity tiers: **Ratio Configuration-I (Minor Split)**, which selects two unseen classes to test minimal generalization, and **Ratio Configuration-II (Major Split)**, which utilizes a rigorous 50% seen-unseen distribution. Crucially, ***FEAr*** imposes a **Valence-Balanced Constraint**, mandating that unseen sets must contain both positive and negative valence categories. This ensures a comprehensive assessment of a model's discriminative power across the emotional landscape, preventing performance inflation arising from polarity bias. We validate this benchmark by applying it to three foundational datasets: FI [9], Emotion6 [10], and ArtPhoto [11].[1]

Besides, the introduction of the **FEAr-accuracy Harmonization (FaH)** evaluation metrics marks a significant advancement in the assessment of model performance. By calculating the harmonic mean of accuracy across the diverse conditions of the ***FEAr*** dataset, the FaH metrics provide a holistic view of a model's efficacy, ensuring a comprehensive understanding of its generalizability and robustness.

Moreover, a novel method, the CLIP-Affective Attribute Generator (AAG), is introduced to *generate semantic attributes enriched with emotional visual information*. This approach aims to produce attributes that better capture emotional visual details, facilitating more effective alignment between image and word embeddings.

The contributions of our work are:

1. We demonstrate that current implementations of ZSL in emotion recognition are beset by deceptive performance and a notable absence of standardized metrics. This critical oversight by researchers could hinder progress in the field, potentially undermining key applications such as mental health diagnostics.
2. We promote reproducible experimentation by introducing a refined dataset framework, termed ***FEAr***, designed to enhance research clarity, consistency, and comparability, as well as pioneering a new evaluation metric, that promise a more accurate depiction of model performance and stability in the image emotion ZSL field.
3. We establish a set of baseline models, encompassing both standard approaches (e.g., CLIP, CoOP, SAE, DeVise, and EMO-L2P) and proposed CLIP-AAG. By integrating CLIP-AAG, which generates semantic attributes enriched with emotional visual details,

[1] This benchmarking suite will be publicly released.

we set a new benchmark for aligning image and word embeddings in emotion ZSL. This foundational benchmark enables researchers to assess the relative effectiveness of new methodologies against well-defined standards, thereby driving the field towards greater scientific rigor and innovation.

2. Related works

2.1. Image emotion recognition

In the early stages, emotion classification was primarily approached by analyzing handcrafted features within a limited number of predefined categories [12–15]. However, the evolution of deep learning has significantly shifted the focus towards leveraging deep features. Notably, Guo et al. [16] introduced the LD-MAN network. This multimodal framework recognizes emotions by extracting and integrating features from images, text, and layout through an attention mechanism. Similarly, Zhang et al. [17] developed the GOSR-CNN, which combines graph-based semantic refinement with convolutional neural networks to enhance visual emotion recognition. Another approach by [1] explored convolutional outputs as a style representation for emotion classification, highlighting the importance of feature correlation.

The recognition of an expanding spectrum of emotions has posed a challenge for traditional models, necessitating frequent and resource-intensive retraining. To address this, ZSL has emerged as a viable solution. ZSL offers a solution by enabling predictions for unseen classes without labeled data. Recent advancements in ZSL [18,19] have shown promising results in bridging the gap between seen and unseen classes, making ZSL a scalable choice for dynamic tasks like emotion recognition.

For ZSL in image emotion recognition, Affective Structural Embedding (ASE) [20] tackled the affective gap by introducing a latent space bridged with DeepSentibank concepts [21]. In contrast, ASE+ [22] enhanced this model by incorporating Tweet2Vec-derived tweet features [23] as auxiliary data. In a novel approach, [5] applied prompt learning to the text encoder of CLIP [6], utilizing GPT-2 generated prompts [24] combined with emotion words for contrastive learning, marking a significant advancement in emotion recognition through ZSL.

2.2. Emotional datasets

Datasets in this domain are classified into two categories, based on emotion models: Categorical Emotion State (CES) and Dimensional Emotion Space (DES).

CES organizes emotions into discrete categories like anger, fear, sadness, etc. This includes Mikel's eight emotions [25] and Ekman's six emotions [26], as well as CES-based datasets such as FI [9], Artphoto [11], and others. DES, conversely, represents emotions in a multi-dimensional space, typically valence-arousal-dominance (VAD), providing a nuanced emotional spectrum. DES datasets include IAPS [27], whereas certain datasets such as Emotion6 [10] are shared across both models.

In zero-shot image emotion recognition, dataset splits often adopt one of two ratios: (i) a minimal unseen class scenario to mirror potential real-world applications, or (ii) an equal division between seen and unseen classes to test model robustness. However, the specificity of these split settings namely, the identification of seen vs. unseen classes remains poorly documented. This omission introduces a significant challenge. The perceived performance of models depends on these undisclosed configurations, leading to misleading portrayals of their capabilities. Our investigations reveal model **performance discrepancies** when classes are assigned as seen or unseen, underscoring the *deceptive performance* issue. The lack of transparent split information hampers accurate model comparison and evaluation, suggesting an urgent need for standardized documentation to ensure genuine progress in the field.

2.3. Evaluation metrics

In zero-shot image emotion recognition, the landscape of evaluation metrics varies, with some studies adopting Average per class top-1 accuracy and harmonic mean (H-mean), while others opt for average accuracy. Specifically, Average per class top-1 accuracy assesses model performance based on the mean accuracy across unseen classes. H-mean evaluates performance via the harmonic mean of average accuracies for seen and unseen classes. Conversely, works such as [5,22] rely on average evaluation accuracy.

However, our analyses reveal a significant flaw: model performance exhibits marked inconsistencies across different configurations of seen and unseen emotion classes. This inconsistency underscores the limitations of current evaluation metrics, which only reflect model efficacy within particular seen-unseen class splits. The situation is compounded by the absence of standardized split settings in existing literature, allowing authors to selectively report outcomes that showcase their model in the most favorable light. Such practices lead to inflated performance metrics, distorting the model's actual effectiveness.

The core issue lies in the non-standardized dataset splits and the resultant over-reliance on metrics like Average per class top-1 accuracy and H-mean, which are inadequate for a holistic assessment. These metrics fail to accurately portray a model's generalizability across diverse emotional spectrums without a universal benchmark for splits. Hence, there is an urgent need for the adoption of standardized metrics and dataset splits to prevent selective reporting and ensure the genuine advancement of the field.

3. Methodology

3.1. Preliminaries

Let $D = \{E_1, E_2, E_3, \ldots, E_N\}$ represent the set of N distinct emotion classes. The emotion dataset $X = \{C_1, C_2, C_3, \ldots, C_N\}$, where the emotional label for subset C_i corresponds to the emotion class E_i. Each subset C_i consists of the images' features extracted for emotions, I_i, and the attribute labels, A_i, related to its respective emotion classes, and can be expressed as $C_i = \{I_i, A_i\}$. In ZSL, the dataset is split into seen and unseen classes, denoted as $Z = X_s \cup X_u$, where Z represents the complete ZSL dataset, X_s signifies the seen classes with $X_s = \{(I_i^s, A_i^s)\}$, and X_u indicates the unseen classes with $X_u = \{(I_i^u, A_i^u)\}$. It is critical that the seen and unseen classes are disjoint, i.e., $X_s \cap X_u = \emptyset$.

The aim of ZSL is to minimize the distance between I_i^s and A_i^s of seen classes by learning two mapping functions $f : I_i^s \rightarrow Y$ and $g : A_i^s \rightarrow Y$, where Y represents a shared embedding space. The training loss, L_{train}, is defined as:

$$L_{train} = L(f(I_i^s; w_1), g(A_i^s; w_2)) \tag{1}$$

Here, L_{train} represents the training loss, and L denotes the loss function. The function $f(\cdot; w_1)$ is a mapping function parameterized by weights w_1, aimed at transforming the image features of seen emotion classes into the embedding space Y. Conversely, $g(\cdot; w_2)$ is another mapping function parameterized by weights w_2, tasked with mapping the attributes of seen emotion classes into the same embedding space Y.

During the testing phase, given an unseen image feature vector (I^u) and a set of candidate unseen classes' attributes (A_i^u), the goal of ZSL in emotion recognition is to identify the emotion class that minimizes the distance between the unseen image feature vector and the attributes in the embedding space. This is formulated as:

$$\arg\min_i \left(f(I^u; w_1), g(A_i^u; w_2) \right) \tag{2}$$

where $f(\cdot; w_1)$ and $g(\cdot; w_2)$ are the mapping functions for unseen image features and attributes, aimed at finding the closest matching unseen emotion class in the embedding space Y.

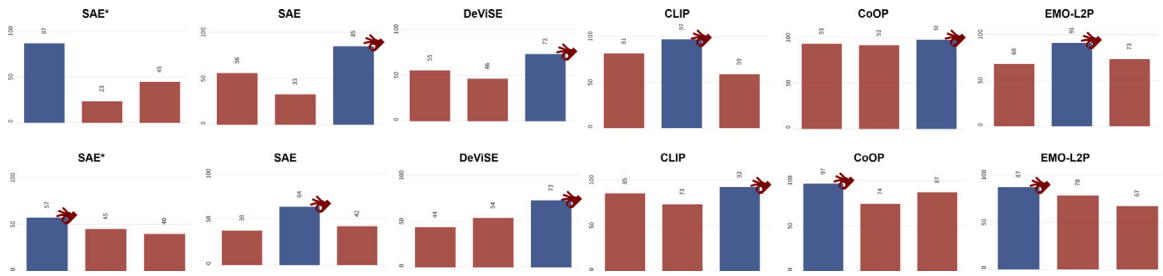


Fig. 2. Model performance varies inconsistently across architectures with different zero-shot learning dataset split configurations.

3.2. Deceptive performance & lack of effective evaluation method

Establishing standardized dataset splitting configurations is crucial in zero-shot image emotion recognition. However, existing studies commonly overlook the importance of explicitly specifying the classes designated as unseen, leading to a significant problem termed as **deceptive performance**.

Problem 1. Existing methodologies in zero-shot image emotion recognition research typically delineate datasets into seen and unseen classes, employing standardized ratio configurations. Despite this, a notable oversight persists. The specific emotion classes designated as 'unseen' often remain undisclosed. This gap in methodology documentation creates a condition known as *performance inconsistency*, in which variations in the selection of unseen classes significantly influence the model's effectiveness. Fig. 2 illustrates that both traditional ZSL models and LVLMs suffer from significant performance inconsistencies, depending on which emotion classes are designated as 'unseen'. It is typical for existing studies to selectively report split settings that showcase optimal results, overlooking substantial performance declines under alternative configurations within the same emotional datasets. This prevalent issue affects both types of models. It spans multiple datasets, highlighting a serious concern. The current approaches may present a misleadingly optimistic view of model capabilities by manipulating seen-unseen splits. This practice undermines the reliability and generalizability of reported performance metrics in emotional recognition tasks.

For instance, in the FI dataset [9] shown at the top of Fig. 2, both traditional ZSL and LVLMs exhibit inconsistent performance. With traditional ZSL, the SAE* [28] model's performance varies from 87% to 23%, resulting in a 64% discrepancy when 'unseen' classes are changed. Similarly, LVLMs show comparable trends; the CLIP [6] model's performance ranges from 97% to 59%, with a 38% discrepancy. This issue is not isolated to a single emotional dataset but is prevalent across all emotional datasets. For example, models trained on the ArtPhoto dataset [11] also encounter this problem, as illustrated in the bottom part of Fig. 2. Selecting different ArtPhoto [11] 'unseen' classes can lead to significant performance variations within a model. Specifically, the DeViSE model's [29] performance ranges from 73% to 44% across different 'unseen' classes. Likewise, LVLMs trained with ArtPhoto [11] face the same issue. For instance, CoOP [7] achieves a high performance of 97% and a low performance of 74% when using different 'unseen' classes.

Consider an emotional dataset X, from which ZSL datasets Z are derived. In one configuration, Z_1, a pair of emotions is designated as 'unseen', formulated as:

$$Z_1 = X_s^{(Z_1)} \cup X_u^{(Z_1)} \tag{3}$$

where $X_s^{(Z_1)}$ encompasses the majority, save for the two 'unseen' classes $X_u^{(Z_1)} = C_{N-1}, C_N$. Another configuration, Z_2, assigns half of the emotion classes as 'unseen':

$$Z_2 = X_s^{(Z_2)} \cup X_u^{(Z_2)}, \tag{4}$$

with $X_s^{(Z_2)}$ and $X_u^{(Z_2)}$ partitioning the dataset equally.

Though Z_1 and Z_2 are intended as controlled configurations, variations arise, denoted as Z_1' and Z_2', where similar ratios are preserved, albeit with a different assortment of unseen classes. This scenario reveals that model performance, denoted as $M(Z)$, deviates from $M(Z')$, underscoring the influence of specific unseen class selections on model efficacy. This inconsistency highlights the limitations of current evaluation metrics, which assess model performance under narrowly defined split settings. Thus, it fails to capture the model's full capabilities.

Problem 2. The prevalent practice in ZSL research of showcasing model performance within narrowly defined split settings (Z) introduces a myopic understanding of a model's true capabilities. This approach, while convenient, sidesteps the broader performance landscape that alternate split configurations (Z') may unveil. The crux of the challenge lies in the existing evaluation metrics, which are calibrated to specific split settings. While effective within their scoped contexts, these metrics fall short of capturing the multifaceted performance of ZSL models across varied unseen class configurations. This myopia muddies the waters for comparative analysis and stifles a holistic understanding of a model's versatility and robustness.

Moreover, the lack of standardized 'unseen' class configurations exacerbates the reproducibility crisis, impeding the ability to benchmark new methods against existing ones reliably. The resulting variability in model performance, attributed to different split settings (Z'), casts doubt on the reported efficacy, underscoring a pressing need for a more detailed and comprehensive evaluation framework in ZSL research. This refined exposition endeavors to spotlight the critical aspects of evaluating ZSL models within the domain of emotion recognition, advocating for a paradigm shift towards more transparent, consistent, and comprehensive methodologies.

Existing studies tend to showcase model performance solely within specific split settings (Z), failing to comprehensively assess its capabilities. Consequently, the evaluation metrics used in this domain may not adequately represent the overall model performance, as they overlook alternative split settings (Z'). This inconsistency poses a challenge for future researchers aiming to compare their results with existing works. The absence of standardized split settings further compounds this issue, hindering the reproducibility and reliability of findings in the field of ZSL.

These problems underscore the critical need for standardized dataset splitting ratio configurations and comprehensive evaluation metrics to ensure the accurate assessment and comparison of ZSL models in emotion recognition.

3.3. Fair and effective evaluation

To mitigate the challenge of performance inconsistency in ZSL for emotion recognition, arising from non-standardized split settings, we introduced ***FEAr*** framework and FaH evaluation. This initiative aims to foster reproducibility by establishing standardized split settings across emotion datasets and introducing effective evaluation metrics for a comprehensive assessment of model capabilities.

3.3.1. Addressing variability through balanced unseen class selection

The essence of our approach lies in crafting split settings that balance emotional polarity in unseen classes, thereby enabling a more holistic evaluation of ZSL models. This balance ensures that models are equally challenged across both positive and negative emotional spectrums, addressing the variability in performance that stems from arbitrary unseen class selections.

Solution 1: Balanced 'Unseen' Class Configuration. We first ensure that the emotion label space admits both positive and negative *valence* categories. Let $D = \{E_1, E_2, \dots, E_N\}$ denote the set of N distinct emotion classes, and let $\phi(\cdot)$ be a valence assignment function such that $\phi(E_i) \in \{p, n\}$. Based on this assignment, the emotion classes are partitioned into valence-specific subsets:

$$D_p = \{E_i \in D : \phi(E_i) = p\}, \qquad D_n = \{E_i \in D : \phi(E_i) = n\}, \tag{5}$$

with the requirement that both valence categories are present:

$$|D_p| > 0 \quad \wedge \quad |D_n| > 0. \tag{6}$$

Valence-balanced 'unseen' class selection. Under a given ***FEAr*** framework, we select m 'unseen' classes, i.e., $|X_u| = m$, where m is determined by the split configuration. To preserve the valence distribution of the complete emotion label space within the unseen set, we first compute the ideal (fractional) number of 'unseen' classes per valence category:

$$\tilde{m}_\rho = m \cdot \frac{|D_\rho|}{|D|}, \qquad \rho \in \{p, n\}, \tag{7}$$

which enforces the principle that the proportion of positive or negative valence classes in the 'unseen' set should approximate that of the full label space.

Since 'unseen' classes must be selected as integers, the final allocation (m_p, m_n) is obtained by selecting integers that sum to m and minimize the deviation from the ideal valence proportions:

$$(m_p, m_n) = \arg\min_{a+b=m} \sum_{\rho \in \{p,n\}} \left| a_\rho - \tilde{m}_\rho \right|. \tag{8}$$

*Instantiation under **FEAr**'s ratio configuration.* We next illustrate how Eq. (7)-Eq. (8) determine the 'unseen' set valence composition under the two ***FEAr*** evaluation ratio configurations.

Ratio configuration-I (minor split). ratio configuration-I selects $m = 2$ 'unseen' classes. The ideal (fractional) allocation is

$$\tilde{m}_p = 2 \cdot \frac{|D_p|}{|D|}, \qquad \tilde{m}_n = 2 \cdot \frac{|D_n|}{|D|}. \tag{9}$$

The integer allocation (m_p, m_n) is then obtained by Eq. (8) with the constraint $m_p + m_n = 2$. When both valence categories are present (Eq. (6)), the ratio configuration enforces inclusion of both valences whenever feasible, yielding the common outcome:

$$(m_p, m_n) = (1, 1), \qquad X_u^{(Z_1)} = \{C^{(p)}_{\text{unseen1}}, C^{(n)}_{\text{unseen2}}\}. \tag{10}$$

Ratio configuration-II (major split). ratio configuration-II selects $m = \lfloor N/2 \rfloor$ 'unseen' classes. The ideal (fractional) allocation is

$$\tilde{m}_p = \lfloor N/2 \rfloor \cdot \frac{|D_p|}{|D|}, \qquad \tilde{m}_n = \lfloor N/2 \rfloor \cdot \frac{|D_n|}{|D|}. \tag{11}$$

Applying Eq. (8) with $m_p + m_n = \lfloor N/2 \rfloor$ yields the integer-valued unseen allocation (m_p, m_n). The resulting 'unseen' set is constructed by selecting m_p positive-valence classes and m_n negative-valence classes:

$$X_u^{(Z_2)} = X_u^{(p)} \cup X_u^{(n)}, \qquad |X_u^{(p)}| = m_p, \quad |X_u^{(n)}| = m_n, \tag{12}$$

which preserves the valence distribution of the full label space up to integer rounding.

Solution 2: FaH Computation. The FaH metric is calculated as the harmonic mean of the accuracy scores obtained across multiple split settings, with a small constant, $\rho = 1 \times 10^{-1}$, added to each accuracy score to prevent division by zero. It is defined as:

$$FaH = \begin{cases} \dfrac{S}{\sum_{i=1}^{S} \dfrac{1}{Acc_i + \rho}}, & \text{if at least one } Acc_i \neq 0 \\ 0, & \text{if all } Acc_i = 0 \end{cases} \tag{13}$$

where S represents the total number of splits, Acc_i denotes the accuracy score from the ith split, and ρ is a small constant ($\rho = 1 \times 10^{-1}$) added to each accuracy score. The inclusion of the small constant ρ prevents infinite values in the calculation when certain accuracies are zero. Additionally, if all accuracy scores are zero, the resulting FaH value will also be zero, providing a consistent representation of model performance under such conditions. This metric inversely weights the accuracy scores, penalizing lower scores more significantly than higher ones. Such an approach ensures that the model's weaknesses are effectively highlighted and addressed, particularly in challenging split configurations.

Our proposition for standardized split settings and the FaH evaluation metric offers a pioneering approach to assessing ZSL models in emotion recognition. By ensuring a balanced representation of emotional polarity in unseen classes and introducing a harmonized evaluation metric, we aim to establish a new standard that addresses the challenges of performance inconsistency and the lack of comprehensive evaluation metrics. This framework promotes a more nuanced understanding of model performance across diverse emotional landscapes and sets a precedent for future research in the field.

4. Fair evaluation for affective recognition

To operationalize the proposed ***FEAr*** framework and enable fair, reproducible evaluation, we instantiate it on three widely used emotion recognition benchmarks: FI [9], Emotion6 [10], and ArtPhoto [11]. Rather than introducing a new dataset, ***FEAr*** defines a standardized ratio configuration for generating valence-balanced seen–unseen splits on existing emotion datasets. This ensures that performance comparisons across zero-shot learning models are not confounded by arbitrary or undisclosed split configurations.

Specifically, for each dataset, we first assign emotion classes to positive or negative valence categories, following established affective psychology conventions. We then apply the two ***FEAr*** ratio configurations, which are ratio configuration-I (Minor Split) and ratio configuration-II (Major Split), to generate multiple valence-balanced 'seen-unseen' partitions. These standardized splits serve as the foundation for all subsequent evaluations and PaSt reporting.

FI [9]: FI is a large-scale affective image dataset collected from Flickr and Instagram. It contains images annotated via Amazon Mechanical Turk according to Mikels' eight emotion categories [25]. While the original dataset reports 23,308 images, the version obtained from the official repository contains 22,700 images, which we use consistently across all experiments.

ArtPhoto [11]: ArtPhoto is a small-scale dataset collected from an art-sharing platform, comprising 806 images labeled by their uploaders using Mikels' eight emotion categories [25]. Due to its limited size and strong stylistic bias, ArtPhoto presents a particularly challenging testbed for zero-shot emotion recognition under varying unseen-class configurations.

Emotion6 [10]: Emotion6 consists of 1980 images collected from Flickr and annotated with Ekman's six basic emotions [26]. Unlike FI and ArtPhoto, which follow Mikels' taxonomy, Emotion6 provides a complementary evaluation setting with a smaller, coarser-grained emotion label space.

Table 1
Complete Split Settings of emotion dataset under ***FEAr* framework** . Abbreviations: AW: Awe, CT: Contentment, AM: Amusement, SD: Sad, DG: Disgust, AG: Anger, EX: Excitement, FR: Fear, JY: Joy, SP: Surprise.

	FI & ArtPhoto				Emotion 6			
	6:2 Split		4:4 Split		4:2 Split		3:3 Split	
ID	Seen	Unseen	Seen	Unseen	Seen	Unseen	Seen	Unseen
1	AW,CT,AM,SD,DG,AG	EX,FR	CT,AM,DG,AG	EX,AW,FR,SD	SP,DG,FR,SD	JY,AG	FR,SD,SP	JY,AG,DG
2	AW,CT,AM,FR,DG,AG	EX,SD	CT,AM,SD,AG	EX,AW,FR,DG	AG,SP,FR,SD	JY,DG	DG,SD,SP	JY,AG,FR
3	AW,CT,AM,FR,SD,AG	EX,DG	CT,AM,SD,DG	EX,AW,FR,AG	AG,DG,SP,SD	JY,FR	DG,FR,SP	JY,AG,SD
4	AW,CT,AM,FR,SD,DG	EX,AG	CT,AM,FR,AG	EX,AW,SD,DG	AG,DG,FR,SP	JY,SD	AG,SD,SP	JY,DG,FR
5	EX,CT,AM,SD,DG,AG	AW,FR	CT,AM,FR,DG	EX,AW,SD,AG	JY,DG,FR,SD	SP,AG	AG,FR,SP	JY,DG,SD
6	EX,CT,AM,FR,DG,AG	AW,SD	CT,AM,FR,SD	EX,AW,DG,AG	AG,JY,FR,SD	SP,DG	AG,DG,SP	JY,FR,SD
7	EX,CT,AM,FR,SD,AG	AW,DG	AW,AM,DG,AG	EX,CT,FR,SD	AG,DG,JY,SD	SP,FR	FR,SD,JY	SP,AG,DG
8	EX,CT,AM,FR,SD,DG	AW,AG	AW,AM,SD,AG	EX,CT,FR,DG	AG,DG,FR,JY	SP,SD	DG,SD,JY	SP,AG,FR
9	EX,AW,AM,SD,DG,AG	CT,FR	AW,AM,SD,DG	EX,CT,FR,AG	–	–	DG,FR,JY	SP,AG,SD
10	EX,AW,AM,FR,DG,AG	CT,SD	AW,AM,FR,AG	EX,CT,SD,DG	–	–	AG,SD,JY	SP,DG,FR
11	EX,AW,AM,FR,SD,AG	CT,DG	AW,AM,FR,DG	EX,CT,SD,AG	–	–	AG,FR,JY	SP,DG,SD
12	EX,AW,AM,FR,SD,DG	CT,AG	AW,AM,FR,SD	EX,CT,DG,AG	–	–	AG,DG,JY	SP,FR,SD
13	EX,AW,CT,SD,DG,AG	AM,FR	AW,CT,DG,AG	EX,AM,FR,SD	–	–	DG,FR,SD	JY,SP,AG
14	EX,AW,CT,FR,DG,AG	AM,SD	AW,CT,SD,AG	EX,AM,FR,DG	–	–	AG,FR,SD	JY,SP,DG
15	EX,AW,CT,FR,SD,AG	AM,DG	AW,CT,SD,DG	EX,AM,FR,AG	–	–	AG,DG,SD	JY,SP,FR
16	EX,AW,CT,FR,SD,DG	AM,AG	AW,CT,FR,AG	EX,AM,SD,DG	–	–	AG,DG,FR	JY,SP,SD
17	–	–	AW,CT,FR,DG	EX,AM,SD,AG	–	–	–	–
18	–	–	AW,CT,FR,SD	EX,AM,DG,AG	–	–	–	–
19	–	–	EX,AM,DG,AG	AW,CT,FR,SD	–	–	–	–
20	–	–	EX,AM,SD,AG	AW,CT,FR,DG	–	–	–	–
21	–	–	EX,AM,SD,DG	AW,CT,FR,AG	–	–	–	–
22	–	–	EX,AM,FR,AG	AW,CT,SD,DG	–	–	–	–
23	–	–	EX,AM,FR,DG	AW,CT,SD,AG	–	–	–	–
24	–	–	EX,AM,FR,SD	AW,CT,DG,AG	–	–	–	–
25	–	–	EX,CT,DG,AG	AW,AM,FR,SD	–	–	–	–
26	–	–	EX,CT,SD,AG	AW,AM,FR,DG	–	–	–	–
27	–	–	EX,CT,SD,DG	AW,AM,FR,AG	–	–	–	–
28	–	–	EX,CT,FR,AG	AW,AM,SD,DG	–	–	–	–
29	–	–	EX,CT,FR,DG	AW,AM,SD,AG	–	–	–	–
30	–	–	EX,CT,FR,SD	AW,AM,DG,AG	–	–	–	–
31	–	–	EX,AW,DG,AG	CT,AM,FR,SD	–	–	–	–
32	–	–	EX,AW,SD,AG	CT,AM,FR,DG	–	–	–	–
33	–	–	EX,AW,SD,DG	CT,AM,FR,AG	–	–	–	–
34	–	–	EX,AW,FR,AG	CT,AM,SD,DG	–	–	–	–
35	–	–	EX,AW,FR,DG	CT,AM,SD,AG	–	–	–	–
36	–	–	EX,AW,FR,SD	CT,AM,DG,AG	–	–	–	–

4.1. Standardized split settings

To eliminate ambiguity in 'seen-unseen' class selection, we instantiate the proposed ***FEAr*** framework on three widely used emotion recognition benchmarks: FI [9], Emotion6 [10], and ArtPhoto [11]. Rather than introducing a new dataset, ***FEAr*** provides a standardized procedure for generating valence-balanced split settings, thereby addressing the issue of inconsistent and often unspecified configurations that lead to deceptive performance reporting in zero-shot emotion recognition.

Table 1 summarizes all split configurations under two ratio settings corresponding to the two ***FEAr*** ratio configurations. Following Eq. (9), the first configuration (ratio configuration-I) designates two emotion classes as unseen, with the constraint that both positive and negative valence categories are represented. For FI and ArtPhoto, which each contain eight emotion classes evenly distributed across valence, this results in $4 \times 4 = 16$ valid unseen-class combinations per dataset. In contrast, Emotion6 contains only two positive and four negative emotion classes, yielding $2 \times 4 = 8$ valid valence-balanced splits. Across all three datasets, ratio configuration-I produces a total of 40 standardized split settings.

The second configuration (ratio configuration-II), assigns approximately half of the emotion classes as unseen while preserving valence balance. For FI and ArtPhoto, this corresponds to selecting four unseen classes composed of two positive and two negative emotions. Enumerating all such combinations yields 36 split settings for each dataset.

Emotion6 constitutes a boundary case due to its asymmetric valence distribution. With six total emotion classes, ratio configuration-II yields $m = \lfloor N/2 \rfloor = 3$ unseen classes. Applying Eqs. (7) and (8) naturally results in two complementary unseen-class compositions: either two positive and one negative emotion, or two negative and one positive emotion. These outcomes follow directly from the ***FEAr*** formulation and do not require any modification to the splitting rules. Enumerating all valid combinations under both cases yields 16 split settings for Emotion6.

Overall, ratio configuration-II produces 88 split settings across the three datasets. Combined with ratio configuration-I, the ***FEAr*** framework generates a total of 128 standardized seen–unseen split configurations. These carefully controlled settings form the basis for FaH evaluation and enable fair, reproducible, and comprehensive benchmarking of zero-shot emotion recognition models.

Note. ***FEAr*** standardizes seen–unseen split choices without enforcing exhaustive usage; any single split from this table may be used for evaluation if explicitly specified.

4.2. FEAr-accuracy harmonization

In our exploration of ZSL models for emotion recognition, we found that their performance is sensitive to split settings. Traditional evaluations, which often focus on results from models trained on specific datasets with particular split configurations, fail to offer a comprehensive view of a model's general performance. This limitation has led to selectively superior results, potentially overshadowing a model's weaknesses or inconsistencies across scenarios.

To counteract this issue and promote a more transparent, fair evaluation of ZSL models in emotion recognition, we introduce the FaH

Table 2
Empirical comparison of aggregation strategies on a representative subset (FI Dataset, Ratio Configuration - I, SAE*).

	Per-split score (%)								Aggregation (%)				
Model	1	2	3	4	5	6	7	8	Mean	H-Mean	Median	TrimMean	Min
SAE* (FIRST RATIO)	**23.39**	86.60	84.88	79.99	73.91	71.01	69.65	65.28	69.34	58.88	72.46	74.12	23.39

H-Mean: harmonic mean. TrimMean: trimmed mean removing the best and worst split.

evaluation metric. This metric, developed with our ***FEAr*** framework, provides a holistic measure of a model's performance across multiple split configurations. The FaH metric calculates the harmonic mean of the accuracy scores obtained for each split setting within the dataset. For instance, in the case of the FI dataset, which includes 16 sets for the ratio configuration-I and 36 sets for the ratio configuration-II, the FaH metric would be applied separately to these two groups, with $S = 16$ and $S = 36$ respectively, according to Eq. (13).

To motivate our aggregation choice, we analyze a representative subset of results on the FI dataset under Ratio Configuration-I using SAE*. We compare several common aggregation strategies, including the arithmetic mean, harmonic mean, median, trimmed mean, and minimum. As shown in Table 2, while most splits achieve consistently high performance, one split shows a pronounced degradation. The arithmetic mean remains relatively high, as it is dominated by strong splits and therefore under-represents worst case behavior. Robust statistics such as the median and trimmed mean further attenuate the impact of difficult splits, effectively masking failure modes. In contrast, the minimum score reflects the worst case but is overly conservative, as it ignores performance on the remaining splits. The harmonic mean occupies an intermediate regime. It substantially penalizes low-performing splits while still accounting for overall performance across splits, making it more sensitive to instability than the arithmetic mean without collapsing to the pessimism of the minimum.

In contrast, the harmonic mean incorporates all split results while assigning greater influence to low-performing splits, leading to a principled reduction in aggregated performance when failures occur. This behavior aligns with the FaH objective: overall performance should reflect robustness to adverse split conditions rather than peak performance alone.

5. Baselines & analysis

We introduced a baseline method, namely the CLIP-Affective Attribute Generator (CLIP-AAG), aimed at pushing the boundaries of current approaches by leveraging the strengths of Large Visual-Language Models (LVLMs) like CLIP [6] in recognizing unseen emotion classes. This section details the innovative CLIP-AAG method, which aims to enhance CLIP text features with effective features.

5.1. CLIP-AAG: Enhancing semantic attributes through a generative approach

Semantic attributes play a crucial role in ZSL, guiding the model to connect semantic cues with visual features for class recognition. Since manually annotating all attributes across domains is impractical, Word2Vec embeddings are often used as semantic attributes in early-stage image emotion ZSL. As explained in [30], Word2Vec learns embeddings from word co-occurrence, focusing on abstract emotional concepts such as "grief" and "melancholy".

While some general actions or events (e.g., crying, rain) may appear in these embeddings, they primarily reflect abstract descriptions rather than fine-grained visual details. Emotions like "sadness" are visually represented through specific objects and actions (e.g., slumped posture, tears) and are often not captured by word embeddings, highlighting the challenge of aligning semantic attributes with visual features in ZSL. Therefore, researchers have shifted to using models like CLIP, which better align text and image embeddings by learning from both modalities, addressing this gap in ZSL tasks.

The CLIP model [6] bridges the visual and textual domains by effectively linking visual content to textual descriptions, providing a comprehensive understanding of how text interprets images. Recent studies demonstrate that high-quality captions significantly enhance the performance of CLIP models [31], underscoring the importance of caption quality. Current research often utilizes a single sentence to represent emotion semantic attributes in the context of zero-shot image emotion recognition. However, this approach is insufficient, as emotions exhibit high intra-class variance. For instance, "sadness" can be visually represented in diverse ways, such as crying, a funeral, or a teardrop.

Building on the findings of [31], improving the captions for each emotion by incorporating specific elements, such as events and objects associated with those emotions, could lead to more accurate prompt representations, thereby enhancing the performance of CLIP [6]. Unfortunately, validating these emotional attributes requires domain expertise—an unexplored process.

To overcome this limitation, we propose generating affectively rich semantic attributes using CLIP's text encoder as the backbone, and extensively integrating them with the visual representations of emotions processed by CLIP's visual encoder. Our approach utilizes a generative method to produce these attributes, ensuring that no visual information is required to generate unseen attributes for inferences. By incorporating these affectively rich semantic attributes, we aim to strengthen the link between emotion attributes and their visual representations, thereby improving the effectiveness of CLIP in zero-shot image emotion recognition.

5.1.1. Integrating affective attribute generator with CLIP

Our proposed CLIP-AAG innovatively leverages a Wasserstein Generative Adversarial Network (WGAN) [32] as the generator to create semantic attributes enriched with affective features. By employing this generator, we can produce unseen attributes that are rich in affective content. As illustrated in Fig. 3, our training process comprises two stages.

In Stage 1, the pretrained CLIP model [6] is fine-tuned using our seen datasets through contrastive learning, following CLIP's original methodology. Let V_e and T_e denote the CLIP visual and text encoders, respectively. For an image–text pair (I_i, A_i), features are encoded as $\delta_i = V_e(I_i)$ and $\epsilon_i = T_e(A_i)$. For a batch of N pairs, the similarity scores are scaled by a temperature parameter τ. The image-to-text loss for the ith sample is calculated by contrasting the positive pair against $N-1$ negative samples via the cross-entropy function, resulting in:

$$\mathrm{L}^{(i)}_{img2txt} = -\log \frac{\exp(\delta_i \cdot \epsilon_i/\tau)}{\sum_{j=1}^{N} \exp(\delta_i \cdot \epsilon_j/\tau)} \tag{14}$$

where the text-to-image loss, $\mathrm{L}^{(i)}_{txt2img}$, is computed symmetrically by swapping the visual and text features in the denominator. Finally, the CLIP loss, L_{CLIP} is calculated by taking the mean of both losses across the batch, as represented by:

$$\mathrm{L}_{\mathrm{CLIP}} = \frac{1}{2N} \sum_{i=1}^{N} \left(\mathrm{L}^{(i)}_{img2txt} + \mathrm{L}^{(i)}_{txt2img} \right) \tag{15}$$

This combined loss function optimizes the mutual alignment of image and text representations within a shared latent space.

In Stage 2, we train an Affective Attribute Generator (AAG) module to produce semantic attributes enriched with affective content. Our approach treats visual features as affective information and trains the

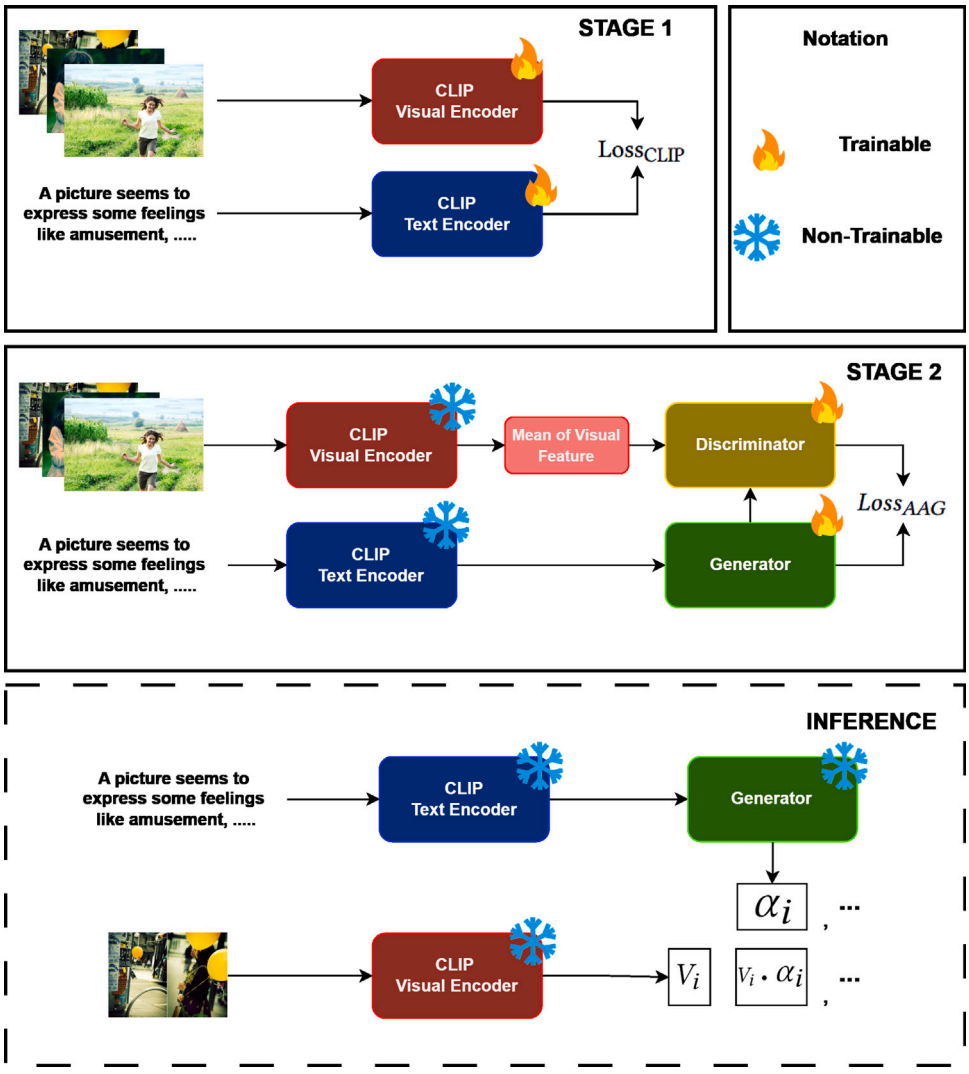


Fig. 3. The training and inference pipelines of proposed CLIP-AAG.

WGAN [32] to generate new attributes that encapsulate them. The goal is to create semantic attributes closely aligned with various visual emotional features. However, the high intra-class variance in the visual representation of emotions, where events, objects, and actions can each convey the same emotion presents a significant challenge. This variability complicates the generative model's task of comprehensively creating semantic attributes that represent the full spectrum of emotional visual features.

To address this challenge, we draw inspiration from [33] and use the mean of visual representations. The mean of visual representations serves as a central point, capturing the general trend in the visual representation of emotions. This approach aims to create a more stable and representative set of semantic attributes that can effectively encapsulate the affective content inherent in the visual features. Additionally, we designed our generator as a UNET [34], as shown in Fig. 4, to preserve the original text attribute information in the newly generated semantic attributes. Fig. 4 also presents the generator's details on the right side.

Let us assume generator and critic in AAG are represented by G and D, respectively. In this stage, we freeze CLIP and use it to obtain both δ_i and ϵ_i. Then, we feed our G with ϵ_i, which represented by $\alpha_i = G(\epsilon_i)$, where α_i represent output of G and the loss of generator is represented by:

$$L_G = -D(\alpha_i) \tag{16}$$

where L_G represent generator loss. Once the G is trained, we will begin to train the D, represented by:

$$L_D = -D(\delta_i) + D(\alpha_i) \tag{17}$$

where L_D represent critic loss. The total loss for AAG, L_{AAG} can be represented as:

$$L_{AAG} = \{L_G, L_D\} \tag{18}$$

where each loss optimized independently. Finally, the CLIP-AAG loss, $L_{CLIP-AAG}$ can be represented as:

$$L_{CLIP-AAG} = L_{CLIP} + L_{AAG} \tag{19}$$

Implementation details of AAG training. The AAG is implemented as a WGAN with weight clipping. Both the generator and critic operate in the 512-dimensional CLIP embedding space. The generator follows an encoder–decoder MLP architecture with skip connections (UNet-style), consisting of linear layers 512 → 256 → 128 → 256 → 512 with LeakyReLU activations. Residual connections are applied at both the intermediate and output levels to preserve the original text-attribute signal. The critic is a lightweight two-layer MLP 512 → 256 → 1

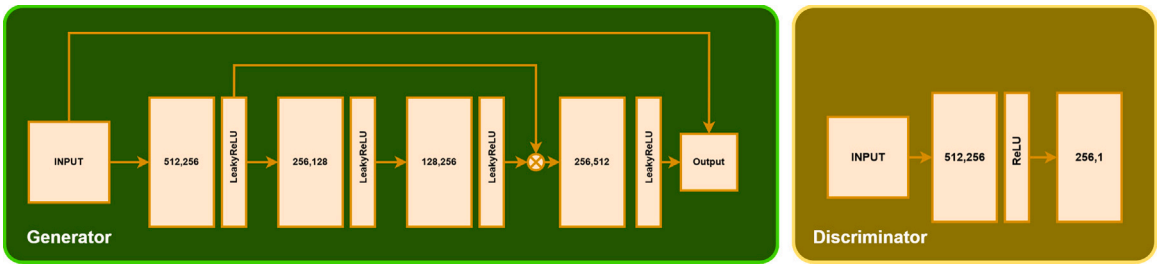


Fig. 4. The architecture of the Affective Attribute Generator.

Table 3
Baseline performance on ***FEAr*** FI dataset. The notations [1], [2], and [3] represent the Top 1, 2, and 3 performances, respectively. Bottom row indicates H-Mean/Std Dev.

Baselines performance on *FEAr* FI dataset														
Sets	Ratio configuration - I							Ratio configuration - II						
	SAE*	SAE	DeViSE	CoOP	CLIP	EMO-L2P	CLIP-AAG	SAE*	SAE	DeViSE	CoOP	CLIP	EMO-L2P	CLIP-AAG
1	73.91	66.88	69.14	93.90	58.75	73.15	60.92	29.16	32.45	42.82	79.40	58.91	55.09	61.06
2	50.11	52.84	59.20	97.30	95.30	83.85	95.57	19.02	39.28	34.05	79.70	59.07	56.77	61.12
3	36.37	48.53	52.52	94.60	91.81	85.96	91.78	35.13	30.82	36.26	76.20	60.68	57.38	65.83
4	69.65	73.54	69.86	91.90	91.14	82.70	91.41	15.67	31.31	31.02	82.30	78.31	58.48	77.67
5	79.99	70.98	64.71	93.50	81.40	68.37	81.47	31.76	42.67	30.99	70.50	79.78	59.10	79.72
6	52.04	73.03	61.09	95.20	91.77	79.42	93.03	18.50	29.52	23.50	77.40	78.13	61.76	78.17
7	34.60	84.63	73.46	92.60	94.05	83.58	94.75	23.69	28.39	29.38	75.10	64.51	59.60	68.62
8	71.01	72.36	65.07	95.20	91.82	78.47	91.14	21.79	21.01	32.96	79.00	64.37	62.18	64.05
9	61.56	34.39	56.79	93.30	85.23	78.87	89.89	27.39	25.53	37.24	79.20	70.51	64.34	74.11
10	65.28	55.58	46.86	91.70	87.31	75.71	87.57	17.22	19.58	21.17	82.20	76.21	60.99	76.92
11	23.39	33.43	54.12	94.20	89.52	83.70	90.16	24.68	35.97	29.02	81.20	79.20	60.05	79.34
12	44.64	75.03	54.86	96.30	91.85	82.52	91.56	18.75	24.51	31.30	72.10	78.26	66.50	78.24
13	86.60	44.41	56.43	97.40	91.65	68.08	91.52	24.58	20.35	26.07	76.10	63.12	50.86	64.39
14	63.33	64.60	44.65	97.70	96.47	85.81	97.66	15.70	15.79	26.75	70.50	58.62	50.94	58.90
15	24.95	69.92	60.04	97.30	93.90	90.97	94.80	29.03	30.98	32.76	69.80	62.60	51.58	62.51
16	84.88	82.70	75.55	97.60	96.85	88.80	96.90	13.33	16.89	22.81	66.10	69.97	56.68	69.77
17								24.62	27.26	26.54	74.40	74.69	55.06	75.13
18								15.33	24.20	30.69	67.40	68.96	56.60	69.33
19								40.02	27.80	35.94	62.80	58.87	52.00	62.98
20								16.30	29.72	39.48	67.00	60.39	55.48	59.54
21								28.76	24.20	29.17	77.20	66.40	56.52	66.34
22								13.60	29.67	30.83	71.30	66.22	54.26	65.98
23								24.58	30.39	27.93	64.30	65.00	53.74	64.56
24								16.65	19.33	29.25	72.20	67.93	57.18	66.37
25								41.84	33.11	34.38	86.50	58.07	55.03	59.61
26	Not Available							15.38	20.77	31.14	80.60	51.12	57.13	51.28
27								47.44	19.92	24.87	84.10	57.57	57.30	58.19
28								13.14	20.96	36.24	86.60	82.85	65.23	82.81
29								11.25	26.87	29.05	78.60	85.07	63.24	85.03
30								15.12	20.44	28.48	78.50	81.34	67.87	82.22
31								36.74	32.39	23.75	73.70	77.03	59.25	77.17
32								12.66	32.82	42.15	79.00	76.14	61.16	76.20
33								36.13	42.05	37.58	85.40	81.54	62.27	79.77
34								11.08	36.32	33.08	84.00	78.65	65.23	79.27
35								11.73	47.49	31.16	86.00	82.10	64.80	82.46
36								12.43	48.30	39.96	84.20	82.01	71.27	81.74
FaH	**49.21**	**58.04**	**59.10**	**95.04[1]**	**88.22[3]**	**80.18**	**89.06[2]**	**19.19**	**26.82**	**30.59**	**76.22[1]**	**68.96[3]**	**58.67**	**69.69[2]**

with ReLU activation. We optimize both networks using RMSProp with a learning rate of 1×10^{-5} and train for 500 epochs. The critic and generator are updated with a 1:1 update ratio ($n_{\text{critic}} = 1$), and critic weights are clipped to $[-0.01, 0.01]$ after each update, following the original WGAN formulation. During training, we monitor zero-shot accuracy on the unseen split using generated attributes and select the final model based on the best unseen ZSL performance.

5.1.2. Inference with CLIP-AAG

At inference, the CLIP-AGG model utilizes affectively rich semantic attributes to predict emotions for unseen images. The process begins with the CLIP model [6] extracting features from both the candidate image and the textual attributes. These textual features are then input into the AAG module to obtain the affectively rich semantic attributes. Subsequently, the model calculates the dot product between the image embedding and each candidate's affective semantic attribute. The attribute with the highest dot product determines the predicted emotion, demonstrating the model's enhanced zero-shot emotion recognition capability.

5.2. Baselines comparison

In our experiment, we evaluate six methods on the 3 emotion datasets with ***FEAr*** framework: traditional ZSL models SAE [28] and DeViSE [29], and LVLM-based models CLIP [6], CoOP [7], EMO-L2P [5] and proposed CLIP-AAG. Using FaH metrics, we aim to accurately assess their performance and establish baselines for future research. Additionally, we aim to highlight the performance inconsistencies in traditional ZSL and LVLM methods.Table 3, 4, 5 represent

Table 4
Performance of baselines for ***FEAr*** Emotion6 dataset. The notations [1], [2], and [3] represent the Top 1, 2, and 3 performances, respectively. Bottom row indicates FaH.

Baselines performance on *FEAr* EMOTION6 dataset														
Sets	Ratio configuration - I							Ratio configuration - II						
	SAE*	SAE	DeViSE	CoOP	CLIP	EMO-L2P	CLIP-AAG	SAE*	SAE	DeViSE	CoOP	CLIP	EMO-L2P	CLIP-AAG
1	53.33	49.85	57.88	87.60	87.42	81.06	87.42	33.13	30.51	48.69	82.40	83.43	73.64	82.83
2	47.73	60.30	56.06	93.60	95.00	83.94	94.70	27.78	32.83	29.60	59.40	68.59	65.25	71.01
3	54.39	52.27	63.03	94.50	90.00	84.09	90.61	27.27	29.49	32.83	62.90	80.20	70.10	80.00
4	51.52	50.15	42.88	89.80	91.97	81.82	92.12	32.83	36.06	42.83	79.10	81.41	74.14	82.93
5	52.58	50.76	55.00	81.40	83.94	70.91	85.00	36.97	34.55	28.99	77.80	86.97	76.87	84.75
6	48.79	52.88	51.06	83.90	92.88	84.09	93.79	31.52	28.89	36.97	79.40	77.47	66.77	77.78
7	53.18	44.85	48.64	85.20	68.33	64.09	68.33	28.38	31.82	45.35	79.00	80.30	66.67	81.31
8	47.12	44.70	51.36	77.70	82.27	77.58	88.18	34.34	35.66	29.39	56.00	57.88	54.24	61.92
9								34.55	32.63	38.18	62.20	73.94	65.05	76.67
10								30.81	30.00	35.66	73.50	70.00	62.53	68.79
11								27.27	32.02	27.88	77.20	77.47	71.11	84.95
12	Not Available							38.59	29.39	36.36	63.70	63.54	59.39	66.46
13								25.56	24.24	34.85	40.10	66.06	55.86	68.18
14								31.62	36.36	29.49	63.50	71.01	62.53	70.30
15								37.88	38.28	29.80	64.90	56.87	58.69	59.60
16								31.31	31.52	29.60	66.50	61.01	59.60	56.77
FaH	**51.04**	**50.42**	**52.70**	**86.47[2]**	**85.74[3]**	**77.88**	**86.78[1]**	**31.50**	**31.86**	**33.91**	**65.96[3]**	**71.18[2]**	**64.60**	**72.34[1]**

Table 5
Baseline performance on ***FEAr*** ArtPhoto dataset. The notations [1], [2], and [3] represent the Top 1, 2, and 3 performances, respectively. Bottom row indicates FaH.

Baselines performance on *FEAr* ArtPhoto dataset														
Sets	Ratio configuration - I							Ratio configuration - II						
	SAE*	SAE	DeViSE	CoOP	CLIP	EMO-L2P	CLIP-AAG	SAE*	SAE	DeViSE	CoOP	CLIP	EMO-L2P	CLIP-AAG
1	41.82	49.55	76.79	89.10	86.36	72.73	92.27	23.36	24.39	31.79	55.50	59.84	45.49	60.25
2	47.23	52.03	49.59	87.50	89.30	73.06	90.77	25.00	26.02	32.67	59.90	61.99	51.53	63.27
3	57.14	45.71	52.86	78.90	86.86	78.29	88.00	23.81	20.80	30.95	65.20	57.39	44.36	59.40
4	42.31	51.65	67.27	76.40	84.62	69.23	86.81	23.93	31.83	34.24	60.90	65.01	48.98	67.72
5	47.47	54.38	53.03	84.30	80.65	70.05	83.41	29.11	25.56	31.25	58.00	62.89	50.22	64.67
6	49.25	61.57	57.54	74.30	73.13	73.51	73.51	27.12	24.29	32.90	54.20	61.30	44.35	61.30
7	43.60	47.67	64.01	89.50	88.37	79.07	89.53	23.03	32.89	27.73	61.80	61.62	47.37	63.38
8	40.78	42.46	62.68	78.80	79.33	67.04	81.01	24.72	21.39	34.61	56.10	63.61	48.33	66.11
9	43.78	41.08	50.96	87.00	84.86	78.38	88.11	32.70	20.44	31.51	65.70	61.04	47.14	63.76
10	46.19	37.71	60.38	76.30	73.73	69.49	80.08	24.33	26.76	28.22	66.70	65.94	46.47	68.86
11	45.00	44.29	62.14	91.40	92.14	87.14	95.00	24.16	29.19	27.93	53.60	64.35	49.28	67.22
12	40.14	52.38	67.66	85.00	87.07	75.51	86.39	25.78	25.16	33.76	58.40	63.04	47.83	59.32
13	39.81	40.74	56.67	85.20	85.19	76.85	86.11	31.83	24.23	23.17	63.90	61.40	48.46	64.27
14	53.93	63.67	47.19	86.50	89.89	70.04	90.26	22.76	30.18	31.16	55.20	57.29	46.55	58.57
15	46.78	46.20	62.11	96.50	80.70	78.36	74.27	27.14	26.13	34.23	48.70	54.77	44.72	57.54
16	42.13	43.82	69.38	80.30	84.27	79.21	85.39	23.98	27.83	27.89	60.90	66.74	51.58	67.19
17								24.72	19.38	28.03	61.90	65.03	50.33	65.70
18								18.98	21.25	32.13	46.50	55.81	47.31	55.24
19								25.17	25.17	30.24	48.60	52.98	43.71	53.64
20								21.01	21.57	29.07	58.80	56.58	44.82	56.30
21								26.10	25.00	26.76	51.60	56.04	47.80	57.97
22								18.63	20.10	28.88	58.80	59.07	50.00	61.03
23								32.29	38.55	31.63	53.00	56.63	43.86	53.98
24								27.59	19.44	31.44	46.40	56.43	43.57	55.17
25								26.45	18.39	28.54	61.80	60.54	50.00	63.43
26	Not Available							27.32	30.41	30.34	52.10	62.89	52.32	65.98
27								26.08	18.99	36.05	50.40	61.77	51.90	62.28
28								26.42	31.44	31.40	58.80	66.29	53.08	66.97
29								26.68	24.22	31.84	66.80	67.04	46.86	65.02
30								23.14	25.14	34.19	57.40	63.43	50.00	65.71
31								30.09	29.20	24.20	60.40	62.83	50.88	63.50
32								29.49	26.40	34.08	59.80	65.17	55.62	66.29
33								27.27	32.23	35.51	57.90	64.74	52.34	65.29
34								23.34	22.11	28.37	66.30	66.09	54.55	66.58
35								25.85	28.74	34.12	56.30	65.46	47.83	65.94
36								18.87	19.18	37.58	51.90	64.15	50.94	63.52
FaH	**45.11**	**47.57**	**59.10**	**83.86[3]**	**83.91[2]**	**74.64**	**85.37[1]**	**25.16**	**24.67**	**30.82**	**57.04[3]**	**61.45[2]**	**48.52**	**62.37[1]**

baselines of models across the ***FEAR*** FI [9], ArtPhoto [11] and Emotion6 [10] respectively. Note that the SAE model [28] has two modes: SAE* projects features to the semantic space, while SAE projects from semantic to feature space.

As presented in Tables 4 and 5, our proposed method consistently outperforms LVLMs across both settings. This finding supports the hypothesis that incorporating affect-rich semantic attributes significantly improves the performance of the CLIP model [6]. Consequently, our

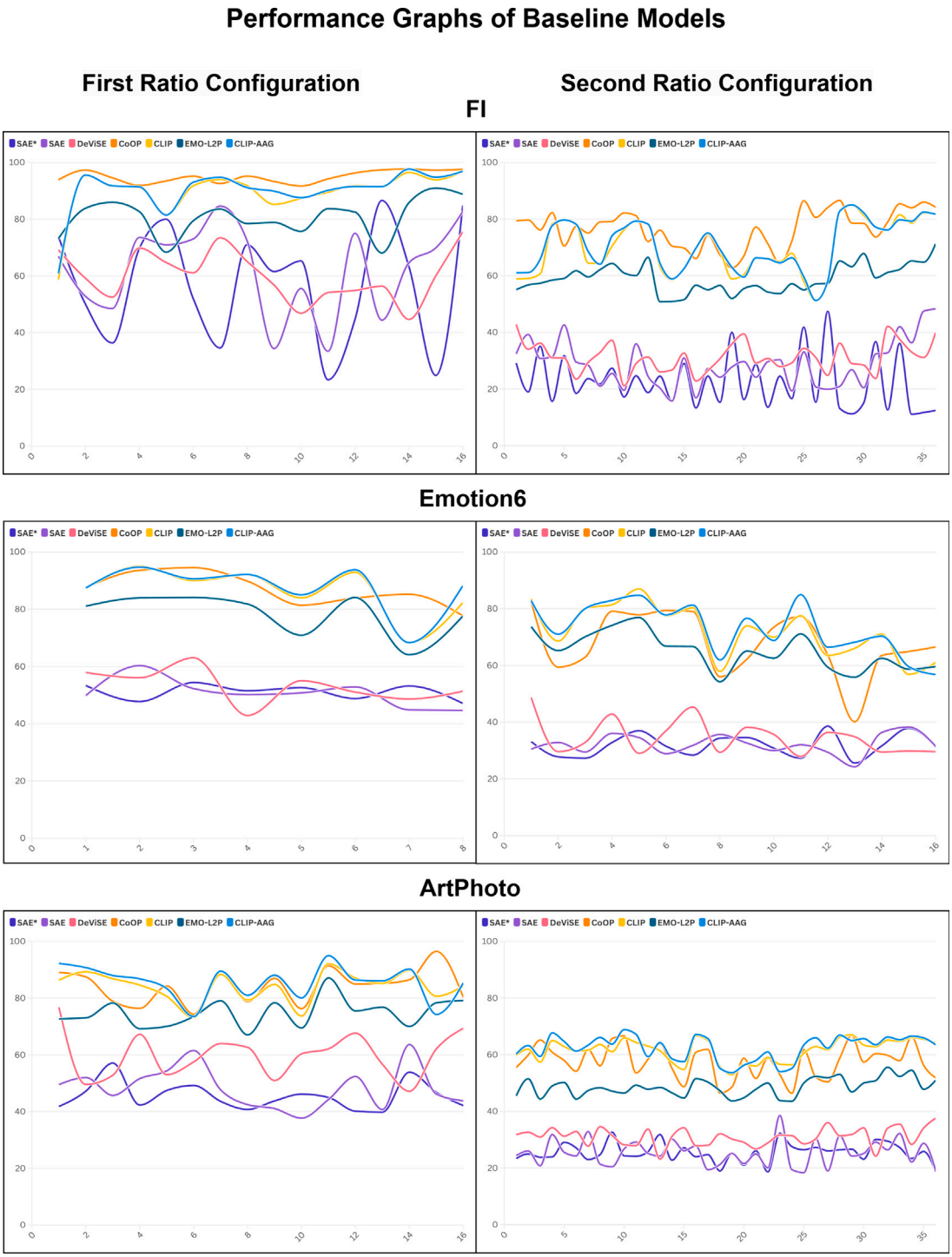


Fig. 5. Performance of various models in the conventional ZSL setting with the proposed FEAr split setting.

hypothesis is validated. That is, integrating emotional visual representations with semantic attributes strengthens the association between emotional semantics and their corresponding visual representations.

However, as shown in Table 3, CoOP [7] demonstrates the best performance, followed by our model when ***FEAr*** FI dataset is employed. The reason is that CoOP [7] employs a learnable vector in the text encoder, which effectively captures emotional features but requires large datasets to identify general emotional patterns. Our model, which relies on the mean of visual representations as a proxy for general emotional trends, performs well on smaller datasets but struggles with larger ones. Despite this, our model still surpasses CLIP [6], indicating that integrating visual representations with semantic attributes enhances performance. Nevertheless, this also reveals a limitation. Relying on the mean of visual representations may not adequately capture general emotional trends in larger datasets.

Our experiments validate that semantic attributes are crucial in zero-shot image emotion recognition. By adequately constructing affective semantic attributes that accurately represent emotions, the performance of zero-shot emotion recognition can be significantly enhanced. However, current methods primarily relying on generative approaches to create unseen emotion attributes may have limitations. In the long run, collaboration between psychologists and machine

Table 6
Performance of baselines for ***FEAr*** FI dataset in GZSL Settings. The notations [1], [2], and [3] represent the Top 1, 2, and 3 performances, respectively. Bottom row indicates FaH.

Baseline performance on *FEAr* FI dataset in GZSL												
Sets	Ratio configuration - I						Ratio configuration - II					
	SAE*	SAE	DeViSE	CLIP	EMO-L2P	CLIP-AAG	SAE*	SAE	DeViSE	CLIP	EMO-L2P	CLIP-AAG
1	0.00	56.97	0.00	15.76	39.03	12.24	0.00	27.33	5.02	15.84	34.10	12.28
2	0.00	44.35	7.20	16.75	34.42	11.75	0.00	36.63	2.72	15.01	37.05	10.92
3	0.00	38.31	0.93	10.58	37.99	9.48	0.00	27.49	0.00	10.65	37.69	10.65
4	0.00	52.94	0.00	12.60	40.93	12.04	0.00	29.76	10.94	10.00	34.26	9.17
5	0.00	29.88	0.00	22.87	34.30	21.10	0.00	29.28	10.73	14.39	35.08	13.92
6	0.00	17.09	5.71	22.79	31.60	19.74	0.00	24.58	2.09	12.09	38.53	8.37
7	0.00	68.04	4.63	17.38	34.90	22.52	0.00	27.16	8.76	29.74	41.15	24.12
8	0.00	35.75	0.00	21.84	36.47	22.49	0.00	21.84	7.63	27.18	44.87	27.67
9	0.00	27.14	0.00	37.12	49.09	19.17	0.00	24.06	0.00	28.50	45.65	23.78
10	0.00	36.61	6.04	32.05	41.58	32.55	0.00	20.91	8.81	22.94	40.96	22.19
11	0.00	38.00	1.77	30.30	45.87	28.52	0.00	35.61	7.27	33.13	41.22	32.60
12	0.00	46.48	0.00	37.10	49.28	36.09	0.00	25.46	2.31	25.55	44.54	24.97
13	0.00	31.78	0.00	64.87	44.00	59.94	0.00	19.05	7.93	42.18	40.39	43.70
14	0.00	39.46	7.59	42.99	38.93	22.00	0.00	17.45	10.36	38.69	42.47	31.98
15	0.00	66.57	3.86	24.47	41.96	5.90	0.00	25.86	0.00	42.29	44.27	43.67
16	0.00	49.73	0.00	51.34	46.30	50.02	0.00	19.92	9.93	30.65	38.82	31.21
17							0.00	26.75	9.66	39.82	40.35	37.59
18							0.00	24.17	3.28	34.22	42.40	32.13
19							0.00	23.92	8.30	33.96	38.17	40.16
20							0.00	28.95	6.05	36.27	43.18	34.29
21							0.00	20.05	0.00	41.29	44.25	40.07
22							0.00	35.73	7.52	30.05	38.69	25.25
23							0.00	28.96	8.59	37.39	39.47	35.85
24							0.00	18.11	1.42	36.70	43.97	35.18
25							0.00	26.18	9.37	21.29	36.96	18.87
26	Not Available						0.00	22.98	0.77	14.64	40.17	14.55
27							0.00	25.17	0.00	19.83	41.21	14.08
28							0.00	25.67	11.55	14.51	37.17	10.80
29							0.00	26.78	9.12	23.32	38.35	20.87
30							0.00	25.84	5.99	14.82	39.88	13.78
31							0.00	16.99	7.74	54.51	43.74	57.18
32							0.00	42.23	2.94	35.66	46.17	24.69
33							0.00	23.86	0.00	52.59	46.89	49.58
34							0.00	38.03	8.69	30.17	42.87	30.39
35							0.00	30.55	11.45	49.15	44.09	47.18
36							0.00	43.24	5.21	31.65	46.23	30.70
FaH	**0.00**	**37.88[2]**	**0.19**	**22.82[3]**	**39.84[1]**	**17.17**	**0.00**	**25.59[2]**	**0.54**	**23.83[3]**	**40.79[1]**	**21.12**

learning specialists will be essential to effectively designing and constructing emotion-related semantic attributes.

5.2.1. Baselines analysis

Fig. 5 presents line plots illustrating baseline performance on the emotion datasets under the ***FEAr*** framework across two ratio configurations. The left side of Fig. 5, the graphs for the ratio configuration - I across three different datasets are shown. The graphs indicate that traditional methods such as SAE [28] and DeVISE [29] generally exhibit lower performance than the LVLMs method. This demonstrates that LVLMs outperform traditional methods because they are trained end-to-end to align visual and semantic embeddings. Furthermore, the baseline model achieves a saturated level of performance in the ratio configuration - I, attaining 94.94% on FI [9], 85.27% on ArtPhoto [11], and 86.68% on Emotion6 [10]. This high performance suggests that further exploration in this configuration may not yield significant advancements.

In contrast, the right side of the Fig. 5 shows the performance of the baseline model on the various emotion datasets under the ***FEAr*** ratio configuration - II. The performance gap between traditional ZSL methods and LVLMs becomes more evident based on the three graphs for the ratio configuration - II. As the training dataset size decreases, traditional ZSL methods' performance degrades more significantly than LVLMs'. Additionally, compared to the ratio configuration - I, the overall performance of the models decreases, likely due to the introduction of a higher number of unseen classes. This suggests that there is substantial room for improvement and makes it a more compelling area for future research. Therefore, future studies should prioritize the ratio configuration - II, as it presents a more challenging and potentially rewarding problem.

6. Extended study

6.1. Generalized ZSL (GZSL)

Unlike conventional ZSL, GZSL simulates real-world scenarios by incorporating both seen and unseen attributes during testing, denoted as $A_i = A_i^s \cup A_i^u$ in our setup. This approach demonstrates the robustness of the baseline method under more realistic conditions.

Table 6, 7, 8 show the FaH for baseline models in GZSL settings across 3 emotion datasets with ***FEAr*** framework. In GZSL, EMO-L2P [5] outperforms the others. As in conventional ZSL, traditional ZSL models perform poorly compared to LVLMs.

In traditional ZSL, most models exhibit deficient performance, achieving accuracy below 30%. Surprisingly, in specific split settings, traditional ZSL models demonstrate 0% performance. Notably, SAE* achieves 0% across all split settings in both the first and ratio configuration - IIs when FI dataset [9] is being used. This highlights the limitations of traditional ZSL approaches, which rely on a few network layers to align visual and textual embeddings. These methods struggle to generalize unseen datasets and are heavily biased towards seen datasets.

Table 7
Performance of baselines for ***FEAr*** Emotion6 dataset in GZSL Settings. The notations [1], [2], and [3] represent the Top 1, 2, and 3 performances, respectively. Bottom row indicates FaH.

Baseline performance on *FEAr* EMOTION6 dataset in GZSL												
Sets	Ratio configuration - I						Ratio configuration - II					
	SAE*	SAE	DeViSE	CLIP	EMO-L2P	CLIP-AAG	SAE*	SAE	DeViSE	CLIP	EMO-L2P	CLIP-AAG
1	25.91	17.27	0.00	23.18	48.79	16.21	22.12	10.61	9.24	35.96	49.49	26.06
2	15.76	20.30	6.06	37.42	50.76	33.64	20.71	4.14	0.00	42.02	51.31	41.41
3	35.00	21.97	0.00	48.03	53.64	45.45	16.87	24.95	13.64	46.77	48.89	30.61
4	31.21	7.42	8.18	44.39	51.06	26.67	22.32	11.01	11.01	50.71	53.23	49.80
5	31.52	28.18	0.00	15.15	32.42	7.58	26.67	3.84	12.37	49.60	52.93	27.47
6	35.45	8.33	6.46	27.73	37.12	28.48	24.24	12.63	10.05	54.55	52.53	43.03
7	46.06	21.21	0.00	38.18	39.55	36.21	25.25	20.71	12.53	25.76	39.70	22.53
8	15.91	4.55	13.84	39.85	34.85	28.64	28.89	23.03	0.00	29.90	40.40	23.13
9							27.37	23.23	16.62	41.21	40.20	31.31
10							26.97	21.82	8.54	45.56	45.35	47.58
11							21.92	17.78	15.71	45.25	43.03	22.63
12	Not Available						22.63	22.93	13.99	47.68	44.34	47.47
13							17.88	13.64	0.00	21.52	37.47	23.33
14							31.72	27.78	13.48	29.29	43.13	31.11
15							21.62	22.63	0.00	38.28	40.10	40.61
16							11.62	12.32	15.51	33.54	41.11	38.38
FaH	**26.16[3]**	**11.41**	**0.20**	**30.13[2]**	**42.17[1]**	**21.18**	**21.91**	**12.27**	**0.39**	**37.38[2]**	**44.71[1]**	**31.70[3]**

Table 8
Performance of baselines for ***FEAr*** ArtPhoto dataset in GZSL Settings. The notations [1], [2], and [3] represent the Top 1, 2, and 3 performances, respectively.

Baseline performance on *FEAr* ArtPhoto dataset in GZSL												
Sets	Ratio configuration - I						Ratio configuration - II					
	SAE*	SAE	DeViSE	CLIP	EMO-L2P	CLIP-AAG	SAE*	SAE	DeViSE	CLIP	EMO-L2P	CLIP-AAG
1	38.64	30.91	0.12	45.45	33.18	12.27	22.75	10.45	12.05	39.75	28.69	31.35
2	25.09	11.07	11.52	36.90	22.88	36.90	18.11	9.44	5.89	39.80	27.81	32.40
3	41.71	27.43	3.04	19.43	16.00	13.71	14.29	7.27	0.00	36.84	31.58	30.58
4	7.14	23.08	0.00	23.08	24.73	21.98	19.86	18.28	13.14	33.18	24.83	22.57
5	26.73	9.68	0.00	55.30	41.01	52.53	19.33	12.00	11.30	32.89	26.44	30.44
6	8.96	13.06	11.67	41.04	27.24	40.67	23.45	18.08	10.18	24.29	22.88	25.42
7	14.53	11.05	4.82	30.81	24.42	18.60	12.72	11.18	12.27	41.89	33.11	40.57
8	18.44	8.38	0.00	32.96	30.17	44.13	15.56	6.94	10.36	35.56	34.17	42.50
9	28.65	3.24	0.00	55.14	44.86	59.46	26.16	10.08	0.00	38.42	37.60	35.15
10	26.27	1.27	11.52	40.25	37.71	3.81	12.17	11.68	12.50	30.66	25.55	17.03
11	20.71	12.86	3.93	25.71	33.57	12.14	13.64	16.03	12.42	34.21	28.71	36.36
12	21.77	6.12	0.00	33.33	30.61	23.13	21.12	10.56	4.46	25.16	25.47	28.88
13	7.41	9.72	0.00	54.63	44.44	51.39	26.90	19.30	11.75	41.07	33.47	40.86
14	13.86	5.99	11.45	44.94	32.96	45.32	12.02	7.67	10.71	37.85	33.25	30.43
15	21.05	2.34	2.50	29.82	26.32	24.56	21.61	3.02	0.00	39.20	36.43	38.44
16	6.74	6.74	0.00	35.39	42.13	36.52	15.84	7.24	12.79	32.58	26.92	24.89
17							20.94	3.12	12.12	36.75	29.62	31.18
18							20.11	15.30	6.61	24.93	23.23	25.50
19							17.00	6.40	10.69	46.36	37.31	47.02
20							17.09	5.60	9.11	45.66	38.94	45.94
21							20.60	18.68	0.00	46.98	36.26	45.60
22							7.11	11.03	12.63	39.95	30.15	38.48
23							31.08	9.64	12.27	41.45	36.63	35.18
24							21.00	16.61	6.61	33.54	27.27	34.17
25							24.59	14.26	8.81	46.49	33.68	49.59
26	Not Available						20.62	13.14	6.07	43.81	34.79	49.23
27							20.00	11.39	0.00	45.32	35.19	44.81
28							13.44	17.08	11.88	38.72	32.35	25.74
29							12.33	12.33	12.27	41.93	28.48	32.29
30							13.14	14.57	7.68	28.29	31.14	32.00
31							30.31	20.13	12.20	49.12	39.60	46.46
32							25.84	20.51	10.54	45.22	38.48	41.29
33							13.22	1.65	0.00	49.04	40.22	46.56
34							7.62	6.39	13.24	40.79	30.96	28.01
35							20.29	12.80	11.90	41.30	33.82	41.06
36							16.04	10.06	7.68	31.76	29.56	27.67
FaH	**15.06**	**5.98**	**0.21**	**34.64[1]**	**29.86[2]**	**18.78[3]**	**16.69**	**8.61**	**0.57**	**37.17[1]**	**31.17[3]**	**33.44[2]**

On the other hand, EMO-L2P [5] demonstrates superior performance in GZSL, attributed to the use of GPT-2 [24]-generated emotion-specific prompts that significantly enhance matching accuracy. Conversely, CLIP-AAG performs the worst among the LVLM models. This shortfall is primarily due to the AAG module being trained exclusively on seen emotion visual representations, leading to a bias in generating new semantic attributes. The issue is further exacerbated when different visual representations convey similar emotions, such as an image of a scene that can be categorized as both awe and contentment. This underscores the limitation of our model, where relying on the mean of visual representations as a general emotional trend is insufficient, primarily when we only trained with seen visual representation.



Fig. 6. Performance of various models in the generalized ZSL setting with the proposed FEAr split setting.

6.1.1. GZSL analysis

Fig. 6 illustrates the performance graphs for baseline models in the generalized ZSL setting. The graph reveals that nearly all models achieve performance levels below 60% in both ratio configurations, suggesting their results are close to random guessing. However, LVLMs still perform better than traditional ZSL models, underscoring the advantages of end-to-end training for aligning visual and textual embeddings.

The performance of ZSL models in the GZSL setting is significantly lower than in conventional ZSL scenarios. This apparent disparity highlights the limitations of current approaches in addressing the challenges posed by the GZSL framework, where models must balance performance on both seen and unseen classes. These findings suggest that existing methods fail to achieve reliable and generalized results in this domain.

Given these challenges, there is substantial potential for innovation and improvement in the GZSL paradigm. It is strongly recommended that future research efforts prioritize the exploration of novel methodologies, enhanced architectures, and more effective training strategies to overcome these limitations. Advancing the GZSL framework could pave the way for more robust and practical applications in real-world scenarios.

Table 9
Performance of baselines for ***FEAr*** in FI to ArtPhoto Cross-Validation setting. The notations [1], [2], and [3] represent the Top 1, 2, and 3 performances, respectively.

Baseline performance of FI → ArtPhoto														
Sets	Ratio configuration - I (6:2)							Ratio configuration - II (4:4)						
	SAE*	SAE	DeViSE	CoOP	CLIP	EMO-L2P	CLIP-AAG	SAE*	SAE	DeViSE	CoOP	CLIP	EMO-L2P	CLIP-AAG
1	47.73	71.36	60.62	83.60	75.91	74.09	84.55	21.52	33.61	25.00	57.80	51.64	63.73	62.50
2	38.75	42.44	50.00	89.70	84.13	63.84	90.04	17.86	39.80	25.72	59.40	53.06	51.28	61.48
3	40.00	60.00	61.67	86.30	90.86	90.86	87.43	25.56	32.83	26.88	52.90	50.63	51.88	59.15
4	57.69	60.44	55.84	86.30	87.91	87.91	85.71	15.80	39.50	25.00	59.40	61.63	48.08	66.37
5	62.21	70.05	60.69	82.00	76.04	54.84	78.80	23.11	33.33	25.00	47.80	58.00	62.22	62.44
6	38.06	62.31	50.00	71.60	72.01	76.49	72.39	19.77	29.38	25.00	53.10	57.63	40.68	60.45
7	40.70	77.91	52.45	87.80	86.05	72.67	89.53	23.03	33.55	25.00	50.90	49.56	55.92	63.60
8	54.19	54.75	59.14	81.60	76.54	84.36	80.45	24.44	38.06	26.19	61.70	53.33	64.44	63.61
9	76.76	72.97	67.58	90.80	84.86	84.32	87.57	28.61	38.69	26.70	51.20	50.95	44.14	61.31
10	29.66	47.46	50.00	71.20	77.12	81.36	71.61	20.68	36.98	29.92	66.40	64.72	46.47	67.64
11	50.00	62.86	50.00	95.70	91.43	97.14	93.57	26.56	33.73	25.00	64.10	63.88	39.23	64.83
12	68.03	53.74	59.55	85.00	86.39	81.63	86.39	25.78	31.06	25.32	58.10	55.59	60.87	61.80
13	66.67	63.43	57.27	89.80	85.65	93.98	85.65	21.56	33.47	25.00	63.70	56.06	46.00	62.83
14	37.83	70.79	50.00	89.50	87.64	56.93	88.39	18.41	35.04	25.00	48.30	53.20	40.66	57.80
15	40.94	60.23	55.94	88.30	88.89	86.55	85.96	24.37	39.20	26.27	54.00	53.77	52.01	51.51
16	57.30	57.30	59.25	86.50	84.27	90.45	85.96	15.84	36.88	30.35	66.50	64.03	59.28	66.97
17								24.72	34.74	25.00	64.10	65.92	64.37	63.47
18								19.83	39.09	25.00	48.40	55.24	51.27	52.41
19								17.00	33.33	25.00	34.90	44.81	34.66	52.54
20								21.29	45.38	25.98	53.20	47.90	63.59	56.86
21								27.20	43.68	33.07	45.10	49.73	37.36	56.87
22								19.85	40.44	25.84	62.50	55.88	49.02	57.60
23								26.75	39.76	25.00	53.50	52.53	40.72	59.28
24								23.51	36.36	24.90	53.30	44.51	50.16	57.05
25								21.69	33.47	25.00	55.40	57.64	74.38	60.74
26	Not Available							18.04	31.19	37.66	58.50	59.28	60.31	64.18
27								22.03	36.71	25.00	51.10	58.73	48.86	64.81
28								15.95	39.86	25.21	69.00	61.73	61.05	65.60
29								18.39	38.34	25.00	56.50	64.35	54.93	65.92
30								20.00	32.00	25.00	55.40	58.00	63.71	62.86
31								33.85	38.05	25.00	62.60	59.73	53.10	58.63
32								19.66	32.58	30.75	64.00	64.61	31.74	64.33
33								24.24	35.54	25.08	61.40	64.46	52.89	64.46
34								17.20	35.87	25.00	60.70	66.09	46.68	63.88
35								17.63	33.09	25.00	59.70	66.67	63.04	63.77
36								22.01	31.13	25.00	67.90	60.69	61.64	64.78
FaH	**47.27**	**60.38**	**55.87**	**84.96[1]**	**83.16[3]**	**77.78**	**84.29[2]**	**21.19**	**35.74**	**26.15**	**56.09[3]**	**56.54[2]**	**50.62**	**61.33[1]**

6.2. Cross-datasets validation

We conducted cross-dataset experiments to assess emotion recognition capabilities across datasets using shared classes from the FI and ArtPhoto datasets. We tested models trained on seen classes from one dataset on the unseen classes of the other, employing the ***FEAr*** split settings for both ratios in this evaluation.

Table 9 presents the performance of models trained on the FI dataset [9] and tested on the ArtPhoto dataset [11], denoted as **FI → ArtPhoto**. Similarly, Table 10 shows the performance of models trained on the ArtPhoto dataset [11] and tested on the FI dataset [9], represented as **ArtPhoto → FI**. It is important to note that the FI dataset [9] is significantly larger than the ArtPhoto dataset [11]. Following trends observed in conventional ZSL, LVLMs consistently outperform conventional ZSL models, with CLIP [6] emerging as the leading model, followed by CoOP [7] and CLIP-AAG. While CLIP [6] does not achieve the highest performance in the **FI → ArtPhoto** task, its results are closely aligned with those of CoOP, establishing CLIP as the overall top-performing model.

Interestingly, the CLIP-AAG model performs well in the **FI → ArtPhoto** setting, ranking second in the ratio configuration - I, very close to the top performer, and securing top performance in the ratio configuration - II. This is because FI, being a larger dataset, provides a mean of visual representation that captures a broader range of emotional trends, leading to better performance when tested on the smaller ArtPhoto dataset. Conversely, the limited visual data in ArtPhoto fails to capture the general emotional trends in the larger FI dataset, highlighting a limitation of our method. We must reiterate that our method has limitations. Despite utilizing a large set of seen images, our approach still struggles to generate comprehensive semantic attributes when faced with large unseen datasets.

6.2.1. Cross-datasets validation analysis

Fig. 7 presents the performance graphs for baseline models in the cross-dataset setting. The upper section of the figure illustrates the performance of models trained on the FI seen dataset [9] and tested on the unseen ArtPhoto dataset [11], while the lower section depicts the reverse scenario. The graphs exhibit a pattern similar to conventional ZSL results, with LVLMs outperforming traditional ZSL models, and a more pronounced performance gap in the ratio configuration - II.

Interestingly, models trained on the ArtPhoto dataset and tested on the FI dataset performed better than those trained on the FI dataset and tested on the ArtPhoto dataset. This could be attributed to ArtPhoto's smaller and more focused dataset characteristics, which may provide specific and coherent features that generalize effectively to the larger FI dataset. Conversely, the FI dataset, being significantly more extensive and diverse, might include noise or features less aligned with the narrower distribution of ArtPhoto, resulting in reduced performance in the reverse case.

In summary, the models' performance using the ratio configuration - I appears saturated. In contrast, the ratio configuration - II highlights opportunities for improvement, likely due to the inclusion of more unseen classes. Future research should prioritize exploring the ratio configuration - II, as it presents a more challenging and meaningful task for advancing the field.

Table 10
Performance of baselines for ***FEAr*** in ArtPhoto to FI Cross-Validation setting. The notations [1], [2], and [3] represent the Top 1, 2, and 3 performances, respectively.

Baseline performance of ArtPhoto → FI														
Sets	Ratio configuration - I (6:2)							Ratio configuration - II (4:4)						
	SAE*	SAE	DeViSE	CoOP	CLIP	EMO-L2P	CLIP-AAG	SAE*	SAE	DeViSE	CoOP	CLIP	EMO-L2P	CLIP-AAG
1	54.85	26.95	50.00	92.90	88.18	68.86	61.96	22.26	20.67	25.00	66.10	74.39	53.53	59.32
2	47.82	42.54	50.00	96.30	96.31	84.40	95.53	15.89	27.78	25.00	64.60	72.98	51.85	61.30
3	63.25	41.12	49.73	92.40	89.74	85.03	89.94	28.70	22.17	25.17	75.90	75.22	57.57	61.96
4	39.85	64.36	53.02	82.60	92.30	86.39	89.61	23.77	23.90	25.37	64.60	78.39	54.93	77.20
5	38.44	27.34	53.73	89.40	85.39	66.25	83.74	27.69	31.77	25.00	63.00	79.46	57.50	81.28
6	51.06	48.07	50.00	86.50	92.81	77.02	91.89	25.74	24.07	25.00	60.40	76.01	52.74	77.67
7	45.71	41.81	50.21	92.00	95.25	83.34	93.94	21.37	21.99	25.00	76.80	76.72	58.28	64.52
8	41.25	32.74	55.79	93.50	95.32	76.91	92.19	27.37	21.95	25.00	79.10	75.38	65.29	66.30
9	57.50	70.70	50.04	95.70	90.24	75.15	87.38	28.20	35.35	25.73	54.60	79.27	60.48	70.72
10	42.37	50.73	50.00	89.70	90.06	73.85	86.88	22.94	26.35	25.00	70.40	77.88	60.99	75.54
11	57.34	55.66	50.00	86.40	92.56	86.35	87.17	28.43	18.38	25.00	68.90	80.10	60.86	77.23
12	47.26	19.39	50.10	94.40	95.65	80.99	90.04	29.09	25.93	25.00	63.60	78.40	62.64	79.11
13	46.41	69.60	50.00	86.40	91.64	74.23	91.23	27.86	19.77	25.00	75.10	73.86	49.07	64.43
14	53.57	36.19	50.00	96.10	95.06	82.30	96.25	20.76	22.07	25.00	61.20	67.80	50.33	58.86
15	47.03	30.89	52.47	96.10	92.39	92.56	92.10	20.86	33.53	22.67	41.70	69.99	51.54	62.40
16	33.67	20.37	50.08	97.30	95.88	89.09	96.74	17.27	25.06	25.15	73.90	72.21	55.43	69.81
17								18.64	23.51	25.00	55.00	73.97	53.26	74.32
18								22.79	24.26	25.00	68.50	62.46	50.17	69.22
19								24.60	31.74	25.00	68.70	69.17	57.82	58.76
20								24.20	35.95	25.00	64.40	71.78	55.95	58.88
21								28.49	39.46	25.46	59.80	73.40	57.25	64.62
22								20.37	17.78	25.04	58.40	72.86	52.56	64.59
23								28.24	23.20	25.00	74.90	73.36	54.62	63.03
24								37.21	25.19	25.00	62.00	74.10	56.76	70.18
25								19.15	18.41	25.00	74.50	79.24	47.40	55.71
26	Not Available							28.01	21.02	25.58	59.20	77.45	59.50	56.18
27								19.36	15.24	25.00	71.00	81.39	56.33	60.10
28								25.83	27.53	24.97	62.80	82.60	61.82	83.02
29								39.35	29.74	25.00	80.30	84.88	58.52	82.91
30								24.60	33.08	25.00	74.50	83.07	65.53	83.01
31								22.60	19.59	25.00	74.60	79.34	55.38	73.26
32								25.55	18.44	25.00	85.70	77.21	60.88	74.10
33								34.03	26.98	25.00	78.90	83.33	61.80	80.34
34								16.37	12.29	25.07	77.30	79.88	61.39	78.70
35								37.90	25.30	25.00	79.60	82.43	65.51	81.68
36								24.05	24.64	25.00	86.60	81.45	68.20	80.27
FaH	**46.81**	**36.55**	**50.99**	**91.62[2]**	**92.42[1]**	**79.62**	**88.36[3]**	**24.21**	**23.50**	**25.10**	**67.46[3]**	**76.21[1]**	**57.00**	**69.02[2]**

Table 11
Ablation study of CLIP-AAG with Emotion6 dataset under the ***FEAr*** framework.

	Ratio configuration-I			Ratio configuration-II		
Method	Pretrained-CLIP	Finetuned-CLIP	CLIP-AAG	Pretrained-CLIP	Finetuned-CLIP	CLIP-AAG
Dataset 1	77.73	83.94	**85.00**	57.37	57.88	**61.92**
Dataset 2	91.21	92.88	**93.79**	73.84	73.94	**76.67**
Dataset 3	66.67	68.33	**68.33**	62.12	63.54	**66.46**
Dataset 4	87.73	82.27	**88.18**	66.16	66.06	**68.18**

6.3. Ablation study on CLIP-AAG components

CLIP-AAG comprises two key components: (i) task-specific finetuning of the CLIP model on seen emotion classes, and (ii) the proposed Affective Attribute Generator (AAG) that enriches semantic attributes with affective visual information. To quantify the individual contributions of these components, we conduct an ablation study by progressively enabling each module while keeping all other experimental settings unchanged.

Specifically, we evaluate three model variants: Pretrained-CLIP, which uses the original CLIP model without finetuning; Finetuned-CLIP, where CLIP is finetuned on the seen emotion classes using contrastive learning; and CLIP-AAG, which further integrates the AAG module to generate affectively enriched semantic attributes. All variants are evaluated under identical ***FEAr*** split settings using both ratio configurations.

Table 11 summarizes the ablation results on a subsampled set of ***FEAr*** for the Emotion6 dataset, presented for clarity of exposition. Finetuning CLIP consistently improves performance over the pretrained baseline across both ratio configurations, indicating the importance of adapting visual–textual representations to the emotion recognition task. More importantly, incorporating the AAG module yields additional and consistent gains over Finetuned-CLIP. The improvement is especially evident under Ratio Configuration-II, where a larger proportion of emotion classes are unseen. This observation suggests that affective attribute generation plays a critical role in enhancing generalization under more challenging zero-shot conditions by strengthening the alignment between semantic attributes and visual affective cues.

7. Conclusion

This paper addresses the challenges of inconsistent model performance stemming from unclear dataset split settings and the lack of robust semantic attributes for emotions in existing work. To overcome these issues, we present the ***Fair Evaluation for Affective Recognition*** (***FEAr***) framework and the FEAr-accuracy harmonization (FaH) evaluation metric, which provide standardized split settings for three widely used datasets and enable comprehensive performance evaluation. The

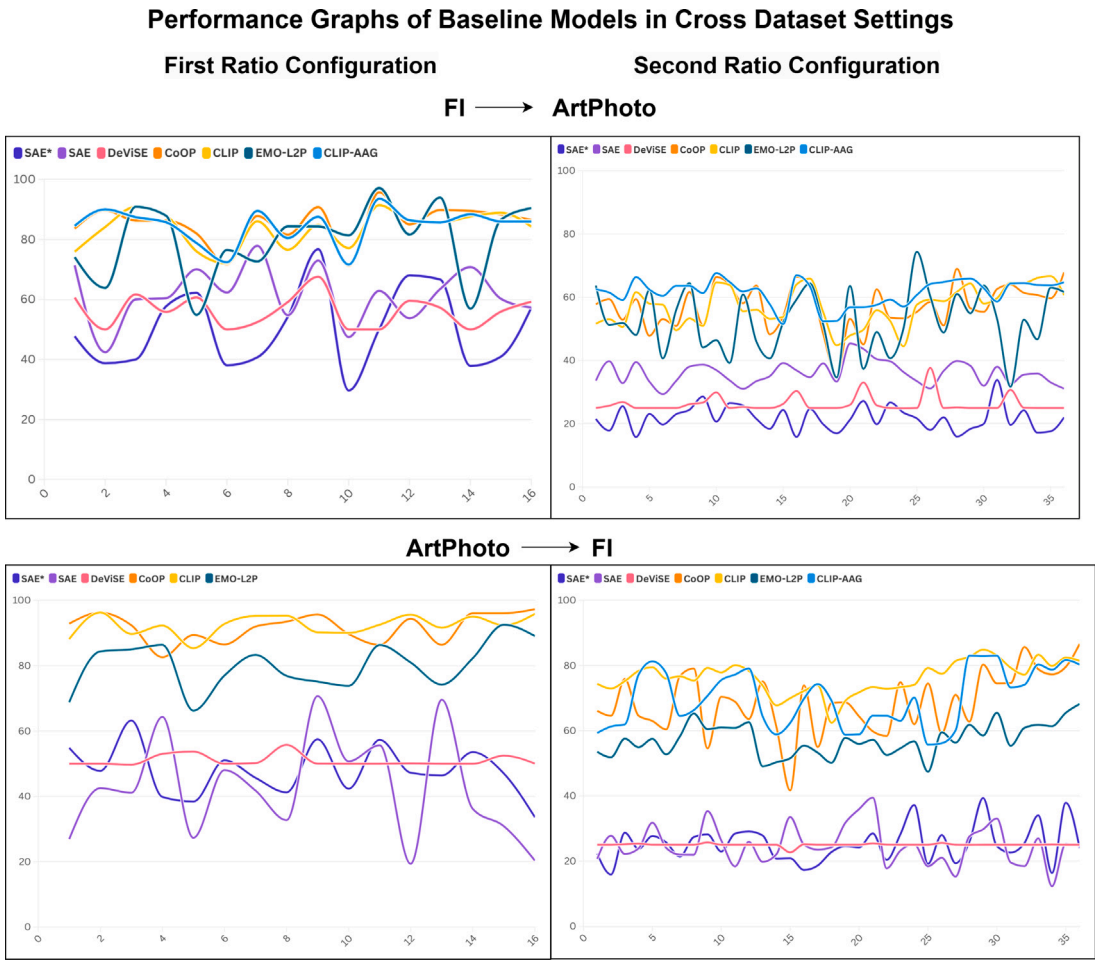


Fig. 7. Performance of various models in the cross-dataset setting with the conventional ZSL setting with proposed FEAr split setting.

FEAr framework enhances consistency by introducing multiple split configurations, while the FaH metric evaluates model performance through the harmonic mean of accuracy across these splits. Furthermore, the proposed CLIP-Affective Attribute Generator (CLIP-AAG) model generates semantic attributes enriched with emotional visual information, improving the alignment between semantic attributes and visual embeddings. Using FaH as a reference, we establish baseline models that provide a foundation for future research and offer a robust framework for evaluating model performance across diverse conditions. By addressing key limitations in the field, this work fosters greater reliability, transparency, and reproducibility in emotion recognition research, paving the way for meaningful advancements in this domain.

CRediT authorship contribution statement

Zhong Ken Hew: Writing – original draft, Visualization, Validation, Methodology, Investigation, Conceptualization. **Lai-Kuan Wong:** Writing – review & editing, Supervision, Methodology, Investigation, Conceptualization. **Chee Seng Chan:** Writing – review & editing, Supervision, Methodology, Investigation, Conceptualization.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work the author(s) used ChatGPT in order to enhance the quality of the writing. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the published article.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This research is supported by the Fundamental Research Grant Scheme (FRGS/1/2024/ICT02/UM/01/1), from the Ministry of Higher Education, Malaysia.

Data availability

Data will be made available on request.

References

[1] Wei Zhang, Xuanyu He, Weizhi Lu, Exploring discriminative representations for image emotion recognition with CNNs, IEEE Trans. Multimed. 22 (2020) 515–523.
[2] Dongyu She, Jufeng Yang, Ming-Ming Cheng, Yu-Kun Lai, Paul L. Rosin, Liang Wang, WSCNet: Weakly supervised coupled networks for visual sentiment classification and detection, IEEE Trans. Multimed. 22 (2020) 1358–1371.
[3] Stuti Jindal, Sanjay Singh, Image sentiment analysis using deep convolutional neural networks with domain specific fine tuning, in: 2015 International Conference on Information Processing, ICIP, 2015, pp. 447–451.
[4] Yongqin Xian, Christoph H. Lampert, Bernt Schiele, Zeynep Akata, Zero-shot learning—A comprehensive evaluation of the good, the bad and the ugly, IEEE Trans. Pattern Anal. Mach. Intell. 41 (9) (2019) 2251–2265.

[5] Hongxia Xie, Hua Chung, Hong-Han Shuai, Wen-Huang Cheng, Learning to prompt for vision-language emotion recognition, in: 2023 11th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos, ACIIW, 2023, pp. 1–4.
[6] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, Ilya Sutskever, Learning transferable visual models from natural language supervision, in: International Conference on Machine Learning, 2021.
[7] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, Ziwei Liu, Learning to prompt for vision-language models, Int. J. Comput. Vis. 130 (2021) 2337–2348.
[8] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, Jeffrey Dean, Distributed representations of words and phrases and their compositionality, NIPS '13, in: Proceedings of the 26th International Conference on Neural Information Processing Systems, vol. 2, Curran Associates Inc., Red Hook, NY, USA, 2013, pp. 3111–3119.
[9] Quanzeng You, Jiebo Luo, Hailin Jin, Jianchao Yang, Building a large scale dataset for image emotion recognition: The fine print and the benchmark, in: AAAI Conference on Artificial Intelligence, 2016.
[10] Kuan-Chuan Peng, Tsuhan Chen, Amir Sadovnik, Andrew C. Gallagher, A mixed bag of emotions: Model, predict, and transfer emotion distributions, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2015, pp. 860–868.
[11] Jana Machajdik, Allan Hanbury, Affective image classification using features inspired by psychology and art theory, in: Proceedings of the 18th ACM international conference on Multimedia, 2010.
[12] Wang Wei-ning, Yu Ying-lin, Zhang Jian-chao, Image Emotional Classification: Static Vs. Dynamic, vol. 7, 2004, pp. 6407 – 6411, vol. 7.
[13] Wei-Ning Wang, Yinglin Yu, Shengmin Jiang, Image retrieval by emotional semantics: A study of emotional space and feature extraction, in: 2006 IEEE International Conference on Systems, Man and Cybernetics, vol. 4, 2006, pp. 3534–3539.
[14] Jianbo Yuan, Sean Mcdonough, Quanzeng You, Jiebo Luo, Sentribute: image sentiment analysis from a mid-level perspective, in: Wisdom, 2013.
[15] Xiaohui Wang, Jia Jia, Jiaming Yin, Lianhong Cai, Interpretable aesthetic features for affective image classification, in: 2013 IEEE International Conference on Image Processing, 2013, pp. 3230–3234.
[16] Wenya Guo, Ying Zhang, Xiangrui Cai, L. Meng, Jufeng Yang, Xiaojie Yuan, LD-MAN: Layout-driven multimodal attention network for online news sentiment recognition, IEEE Trans. Multimed. 23 (2021) 1785–1798.
[17] Jing Zhang, Xinyu Liu, Zhe Wang, Hai Yang, Graph-based object semantic refinement for visual emotion recognition, IEEE Trans. Circuits Syst. Video Technol. 32 (2022) 3036–3049.
[18] Yurui Xie, Tiecheng Song, Jianying Yuan, Learning discriminative visual semantic embedding for zero-shot recognition, Signal Process., Image Commun. 115 (2023) 116955.
[19] Xiaoming Hu, Zilei Wang, Junjie Li, Learning complementary semantic information for zero-shot recognition, Signal Process., Image Commun. 115 (2023) 116965.
[20] Chi Zhan, Dongyu She, Sicheng Zhao, Ming-Ming Cheng, Jufeng Yang, Zero-shot emotion recognition via affective structural embedding, in: 2019 IEEE/CVF International Conference on Computer Vision, ICCV, 2019, pp. 1151–1160.
[21] Tao Chen, Damian Borth, Trevor Darrell, Shih-Fu Chang, DeepSentiBank: Visual sentiment concept classification with deep convolutional neural networks, 2014, arXiv abs/1410.8586.
[22] Yingrui Ye, Yuya Moroto, Keisuke Maeda, Takahiro Ogawa, Miki Haseyama, Affective embedding framework with semantic representations from tweets for zero-shot visual sentiment prediction, in: Proceedings of the 4th ACM International Conference on Multimedia in Asia, 2022.
[23] Bhuwan Dhingra, Zhong Zhou, Dylan J. Fitzpatrick, Michael Muehl, William W. Cohen, Tweet2Vec: Character-based distributed representations for social media, 2016, arXiv abs/1605.03481.
[24] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, Language models are unsupervised multitask learners, 2019.
[25] Joseph A. Mikels, Barbara L. Fredrickson, Gregory R. Larkin, Casey M. Lindberg, Sam J. Maglio, Patricia A. Reuter-Lorenz, Emotional category data on images from the international affective picture system, Behav. Res. Methods 37 (4) (2005) 626–630.
[26] Paul Ekman, An argument for basic emotions, Cogn. Emot. 6 (1992) 169–200.
[27] Margaret M. Bradley, Peter J. Lang, International affective picture system, in: Virgil Zeigler-Hill, Todd K. Shackelford (Eds.), Encyclopedia of Personality and Individual Differences, Springer International Publishing, Cham, 2017, pp. 1–4.
[28] Elyor Kodirov, Tao Xiang, Shaogang Gong, Semantic autoencoder for zero-shot learning, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2017, pp. 4447–4456.
[29] Andrea Frome, Gregory S. Corrado, Jonathon Shlens, Samy Bengio, Jeffrey Dean, Marc'Aurelio Ranzato, Tomas Mikolov, DeViSE: A deep visual-semantic embedding model, in: Neural Information Processing Systems, 2013.
[30] Yoav Goldberg, Omer Levy, Word2vec explained: deriving mikolov et al.'s negative-sampling word-embedding method, 2014, arXiv abs/1402.3722.
[31] Pavan Kumar Anasosalu Vasu, Hadi Pouransari, Fartash Faghri, Oncel Tuzel, CLIP with quality captions: A strong pretraining for vision tasks, 2024, arXiv abs/2405.08911.
[32] Martín Arjovsky, Soumith Chintala, Léon Bottou, Wasserstein generative adversarial networks, in: International Conference on Machine Learning, 2017.
[33] Omkar Gune, Biplab Banerjee, Subhasis Chaudhuri, Structure aligning discriminative latent embedding for zero-shot learning, in: British Machine Vision Conference, 2018.
[34] Olaf Ronneberger, Philipp Fischer, Thomas Brox, U-Net: Convolutional networks for biomedical image segmentation, in: Nassir Navab, Joachim Hornegger, William M. Wells, Alejandro F. Frangi (Eds.), Medical Image Computing and Computer-Assisted Intervention, MICCAI 2015, Springer International Publishing, Cham, 2015, pp. 234–241.