Who, Where, How Matters: Evaluating Emotion Recognition under Context Shifts

Sayak Mukherjee
Delft University of Technology
Delft, Netherlands
sayak.mukherjee@tudelft.nl

Tom Viering
Delft University of Technology
Delft, Netherlands
t.j.viering@tudelft.nl

Bernd Dudzik
Delft University of Technology
Delft, Netherlands
b.j.w.dudzik@tudelft.nl

Abstract

Human emotion perception is inherently context-dependent, especially when observable behaviour is ambiguous. Vision-based emotion recognition systems learn to mimic emotion recognition from data. However, collecting a representative dataset is challenging due to the lack of consensus on what constitutes relevant contexts, as well as ethical and privacy constraints. This brings into question if such systems perform well across diverse real-world contexts. We consider three sources of context based on psychological theory: who expresses the emotion, where the emotion occurs, and how the data is collected. Our key contributions are: (a) Operationalising these contextual variations using EMOTIC, a widely used dataset; (b) Evaluating the supervised EMOTIC model when trained on one subgroup (e.g., women) and tested on another (e.g., men) within each contextual dimension (e.g. gender); and (c) Analysing the zero-shot performance of open-source (vision) language models under the same variations. Our results show that the EMOTIC model exhibits statistically significant performance gaps across contextual subgroups. We find large effect sizes for multiple metrics. These effects remain after controlling for dataset size, and some gaps even increase as the training sets grow. Zero-shot models underperform compared to EMOTIC, and also do not show consistent performance across contexts. These performance gaps are problematic for socially consequential settings, such as education, particularly regarding algorithmic discrimination and bias. Our work stresses the importance of context-aware data collection and evaluation for automated emotion recognition systems.

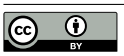
CCS Concepts

• **Human-centered computing** → **Human computer interaction (HCI)**; • **Computing methodologies** → **Cognitive science**.

Keywords

Context-aware Emotion Recognition, Context-sensitivity, Affective Computing, Generalisation

ACM Reference Format:
Sayak Mukherjee, Tom Viering, and Bernd Dudzik. 2026. Who, Where, How Matters: Evaluating Emotion Recognition under Context Shifts. In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI '26), October 05–09, 2026, Napoli, Italy*. ACM, New York, NY, USA, 25 pages. https://doi.org/10.1145/3776574.3831130



This work is licensed under a Creative Commons Attribution 4.0 International License.
ICMI '26, Napoli, Italy
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2318-6/26/10
https://doi.org/10.1145/3776574.3831130

1 Introduction

The ability of intelligent systems to infer users' emotional states is becoming increasingly important as they interact more closely with humans in complex social environments, such as in healthcare and education [39, 43]. Such capabilities further support modelling human experiences and values (or preferences), which are central to the development of *human-compatible* intelligent systems [48].

Traditional emotion recognition systems have been formulated as supervised pattern recognition tasks. Such systems learn to infer emotional states of users from their behavioural signals in isolation using third-party annotations as supervisory labels [55]. However, this formulation has significant limitations. First, behavioural signals alone may not be very expressive or have an emotionally ambiguous meaning [1, 35]. Second, behaviour is strongly modulated by the type of social interaction, cultural norms, and the personal circumstances of the individual expressing the emotion, among other factors [2, 47]. Lastly, the supervision labels reflect the outcome of a human perception process, which can vary depending on the annotator's own context and perspective as an observer [22]. These challenges have motivated the development of context-aware emotion recognition (CAER) systems that mirror the ability of humans to use contextual information when inferring others' emotional states [25, 28, 40].

Recently, deep-learning-based end-to-end CAER models have been proposed that learn to infer emotional states by considering the visual context of social interaction and the behaviours expressed within it [28, 36], supported by datasets designed to capture such information (e.g., EMOTIC [28] or HECO [60]). Such approaches implicitly assume that, given sufficiently large and diverse datasets, models can learn the variability required to generalise beyond the training distribution and operate reliably across real-world scenarios. However, in practice, collecting balanced and representative affective data remains challenging due to the lack of consensus on what constitutes relevant context [10], the subjective nature of human emotion perception [4, 8], and ethical and privacy constraints [38]. These challenges raise concerns about the performance of CAER models under contextual shifts, given their use in high-stakes settings such as education [44, 56] and healthcare [7, 16]. More broadly, emotion recognition systems have been shown to exhibit demographic biases, raising fairness concerns [24].

Based on the Modified Brunswikian Lens Model [50, 57], we consider three sources of contextual information that shape emotion perception [9]: who expresses the emotion (sender context), where the emotion occurs (situational context), and how the data is collected (dataset context). To operationalise variations within these dimensions, we use relevant subsets of the EMOTIC dataset

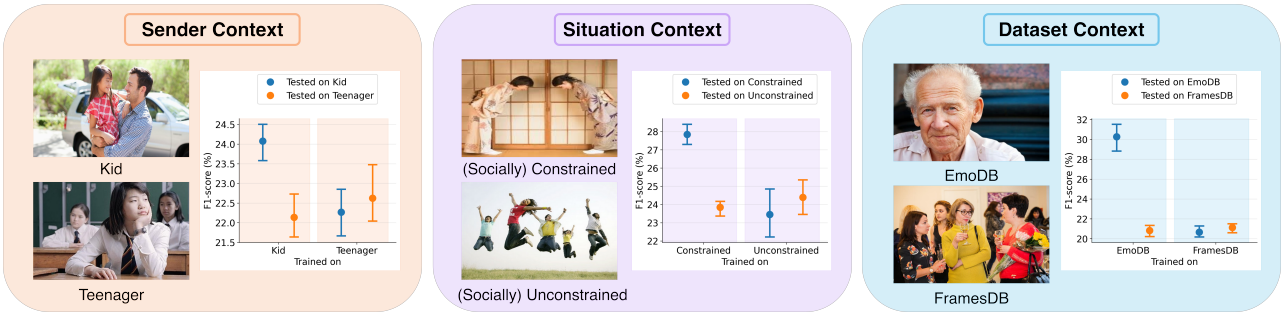


Figure 1: Variations across three context dimensions, namely who expresses emotion (sender), where emotion occurs (situation), and how data was collected (dataset context), illustrate the challenges of context-aware emotion recognition. The mean F1-scores with 95% confidence interval for the supervised EMOTIC model [28] show performance drops when trained on one subgroup (e.g., kid) and evaluated on another (e.g., teenager), revealing limited generalisation within contextual dimensions (e.g., sender).

[28], a widely used CAER benchmark [12, 13, 36, 59]. Then, we conduct a series of controlled experiments (Figure 1) to investigate:

> *How do variations in who expresses emotion (sender), where emotion occurs (situation), and how data was collected (dataset) influence the performance of CAER models?*

Furthermore, recent work has explored the use of large (vision) language models (LLMs/VLMs) for inferring emotions from behavioural signals while also considering contextual information. These studies mainly focused on the zero-shot setting, without task-specific training [3, 13, 53]. While promising, the analyses are rarely accompanied by an analysis of the model's performance across variations in contextual dimensions. To address this, we evaluate the performance of open-source zero-shot (vision) language models across the same contextual dimensions and compare them with supervised CAER models. By jointly evaluating both paradigms under identical conditions, our work provides a systematic analysis of how current emotion recognition systems perform under contextual shifts. Our contributions are summarised below:

- **Operationalising** context variations in EMOTIC informed by the Modified Brunswikian Lens Model [57], in Section 3.
- **Demonstrating** that for the supervised EMOTIC model, contextual shifts lead to performance gaps and asymmetric generalisations that cannot be fixed by adding more data. We report the effect size and statistical significance while carefully controlling for dataset size in Section 5.1.
- **Comparing** supervised EMOTIC model and zero-shot (vision–) language systems under identical contextual variations in Section 5.2, with results showing that even large-scale pre-training does not guarantee real-world robustness.

2 Background

Context in emotion recognition. Early computational models of vision-based emotion recognition were largely based on psychological models that assumed an unambiguous and universal relationship between emotional states and specific facial expressions [11, 15, 24]. However, findings from psychology increasingly challenge this view, highlighting that emotion perception emerges from a complex interplay between expressive behaviour and contextual information [2]. A growing body of work shows that contextual cues, such as visual background [57], vocal signals [45], and others' expressions [37], strongly influence how individuals perceive emotions. Moreover, the relative reliance on facial versus contextual information depends on the observer's perspective [22]. Empirical studies also demonstrate how contextual information shapes the perception of emotions conveyed by facial expressions [39]. These motivate a systematic examination of how such contextual variation impacts computational emotion recognition systems.

A principled framework for studying these influences is the Modified Brunswikian Lens Model (MBLM) of affect communication [50]. Building on this model, Wieser and Brosch [57] identifies three primary sources of contextual information: (i) *Sender context*. Characteristics of the individual expressing the emotion, such as goals, motivations, memories and psycho-physiological conditions. (ii) *Situation context*. Properties of the environment within which the sender is embedded. (iii) *Perceiver context*. Prior knowledge and biases of the observer, including cultural norms and expectations. We leverage this framework to study how contextual shifts affect CAER models. Specifically, we analyse variations in sender and situational context, and additionally consider the dataset context, which can be viewed as a special case of situation context. Dataset context captures differences in data-collection environments, participant populations, and annotation processes [8, 9].

Context-aware emotion recognition. Context-Aware Emotion Recognition (CAER) extends traditional facial expression analysis by explicitly modelling the joint influence of a sender's expression and the surrounding situational context. There are two main categories of CAER methods proposed in the literature.

The first group of methods conceptualises CAER as a two-stage reasoning process: (i) inferring emotion from observable expressions [11] and (ii) reasoning about emotions from the contextual information (e.g. event) in relation to the expressor's goals, beliefs, and values [46]. These stages are typically modelled separately and later combined through *emotional cue integration* [26, 40, 49]. Recent work has shown promising results using a two-stage model that leverages Large Language Models (LLMs) for contextual understanding [3, 53]. Nonetheless, how and when different information

sources should be integrated across individuals and social interaction contexts remains an open question [18, 21, 23, 51].

In contrast, the second group of CAER methods focuses on end-to-end deep learning models. Earlier supervised vision-based approaches employ dual-branch architectures: one branch extracts expressive cues from facial or bodily signals, while the other encodes contextual cues from the surrounding scene. A fusion module then integrates these representations to predict emotion labels. [17, 28–31, 36]. Although effective, such architectures require large, diverse, and well-annotated datasets to generalise across contexts. However, the lack of data from relevant target contexts due to privacy and logistical constraints raises concern about generalisability across populations or scenarios. Owing to the growth of Vision-Language Models (VLMs) and Multimodal Large Language Models (MLLMs), recent studies have shown significant improvements in CAER over supervised vision-based models [6, 13, 61]. While promising, these approaches also lack systematic evaluation under realistic contextual variations. We address these gaps by systematically analysing both supervised and zero-shot CAER models.

Datasets. CAER datasets are typically derived from two sources. The first consists of images from movies and television [29]. While these are often described as naturalistic images, media portrayals are deliberately crafted to elicit specific affective responses [19]. As a result, they may capture cinematic regularities rather than real-world contexts for emotion perception. The second source includes images of people *in the wild* collected from the internet, such as EMOTIC [28] and HECO [60], which reflect more closely real-world scenarios in which CAER systems are expected to operate. In this work, we focus on EMOTIC, as it is a widely used benchmark for contextual emotion recognition [12, 13, 36, 59].

3 Operationalising Context Variations

EMOTIC consists of a large collection of images that include both individuals and their surrounding visual context. Each image is labelled with a subset of 26 discrete emotion categories (i.e., multi-label). Specifically, it contains four sub-datasets: EmoDB, FramesDB, MSCOCO, and Ade20K. EmoDB and FramesDB comprise images manually collected from the web via Google search, using queries that combine activities, places, social environments, and emotional states [28]. In contrast, MSCOCO [33] and Ade20K [63] are derived from benchmark datasets developed originally for object recognition and scene understanding, respectively. The aggregation of data from multiple sources introduces substantial variability in scene composition, participating populations, and data-collection environments. This heterogeneity makes EMOTIC particularly well-suited to studying the effects of realistic contextual variation on CAER models. However, EMOTIC does not provide metadata on annotators, which limits our ability to analyse perceiver context.

In the following discussion, we analyse data imbalance (as shown in Figure 3) and examine how the categorical distributions of emotion labels vary within each contextual dimension. Owing to space constraints, the discussion in the main paper focuses on the most salient trends, while the corresponding detailed distribution plots are provided in Figures 6–11 in Appendix D. We quantify these distributional dissimilarities using Total Variation (TV) and Jensen-Shannon (JS) distances, which are bounded between 0 and 1.

Sender Context. For the sender context, we consider demographic attributes that are explicitly annotated in EMOTIC, namely age (adult, teenager, kid) and gender (male, female). These attributes are generally considered to account for variation in how emotions are expressed in individuals [5, 14].

Kids and teenagers, among the age subgroups, and females, among the gender subgroups, are substantially under-represented across all data splits (Figure 3). Such skewed representation may introduce bias in model learning, potentially limiting generalisation performance for under-represented age groups. We further analyse the distribution of the 26 categorical emotion labels across the subgroups defined by age and gender, as shown in Figure 6 and Figure 7 of the Appendix, respectively. The distributions reveal systematic differences in perceived emotions. Specifically, while kids are more frequently associated with emotions such as affection, happiness, and peace, teenagers are more often perceived as confident and excited. In contrast, for adults, labels such as disconnection, engagement, and esteem are more prevalent. Within gender subgroups, males are often labelled with emotions such as anticipation, confidence, and excitement, whereas females are perceived as portraying affection, happiness, pleasure, and sympathy.

Quantitatively, both JS and TV distances indicate that the largest divergence occurs between the Teenager and Kid (JS: 0.101, TV: 0.113) groups, whereas the Adult and Teenager groups exhibit the most similar label distributions (JS: 0.048, TV: 0.046). Furthermore, the JS and TV distances between the male and female subgroups are 0.092 and 0.113, respectively. Together, these findings suggest that age- and gender-related differences exist in annotations.

Situation Context. Since the situation context is not explicitly annotated in the EMOTIC dataset, we automatically derive it by first predicting scene categories using a ResNet-18 model pretrained on Places365, followed by grouping them into semantically meaningful, non-overlapping categories (e.g., social vs private) using GPT 5.2. Additional details regarding the annotation process can be found in Appendix A. For our analysis, we focus on variation at two levels of granularity. At a coarse level, we distinguish between socially constrained and unconstrained settings. While constrained settings mainly comprise indoor scenes (e.g., classrooms and restaurants), unconstrained settings primarily consist of outdoor scenes (e.g., farms and streets). Importantly, this distinction not only captures differences in lighting conditions but also broadly encompasses variations in social exposure, activity type, and environmental constraints that are known to influence emotion perception. At a finer level, we examine variation within each group. Within the socially constrained settings, we consider private, public, and commercial subcategories, reflecting differences in social norms and interaction dynamics. Furthermore, within the socially unconstrained settings, we consider rural, urban, and social environments that vary in crowd density and activity structure.

Figure 3 shows that the two coarse groups are nearly balanced, but notable imbalances exist within subgroups, with private and rural settings underrepresented. Label distributions (in Figure 8) indicate that unconstrained settings are dominated by high-arousal



Figure 2: Images from the subdatasets of EMOTIC.

emotions such as anticipation, confidence, and excitement, whereas constrained settings have a broader emotional spectrum, ranging from happiness and pleasure to confusion and disconnection. This difference is reflected quantitatively by JS distance of 0.118 and TV distance of 0.139. Further variation emerges within each setting. Within constrained environments, individuals are perceived as confident and engaged in public scenes, whereas in commercial scenes they are perceived as happy and pleased. In contrast, individuals are more frequently labelled as peaceful or fearful in private settings, as shown in Figure 9. The largest differences in label distributions occur between public scenes and commercial or private scenes (JS: 0.066; TV: 0.067), while commercial and private scenes exhibit the most similar distributions. Within unconstrained environments, individuals are often perceived as displaying positive, low-arousal emotions, such as affection and peace in rural scenes. Whereas those in social scenes are more frequently labelled with high-arousal emotions, including excitement and anticipation. Moreover, individuals in urban scenes are more often perceived as displaying negative or fatigued emotional states, such as confusion and sadness (Figure 10). The greatest distributional difference is observed between rural and social scenes (JS: 0.150; TV: 0.176), while social and urban scenes show the smallest difference.

While the situation labels considered in our analysis provide a tractable abstraction, they implicitly capture a rich and complex structure of real-world situations that shape how emotions are expressed and interpreted. Analysing performance across these hierarchical variations allows us to systematically evaluate how models respond to contextual shifts in realistic social settings.

Dataset Context. For dataset context, we treat each subset of the EMOTIC dataset, namely EmoDB, FramesDB and MSCOCO, as a distinct subgroup. These subsets differ substantially in how images are collected and composed, introducing variability in both expressive and contextual cues as shown in Figure 2. For example, EmoDB images often resemble curated or stock photographs, in which individuals are centrally positioned and exhibit clear expressions with minimal surrounding context. In contrast, MSCOCO and FramesDB images are more complex and unconstrained, frequently containing multiple people, occlusions, and richer scene context, where facial expressions may be less prominent. We hypothesise that such differences would influence how models rely on behavioural versus contextual cues for emotion inference. Specifically, models trained on EmoDB are likely to perform better on their corresponding test sets but struggle on other datasets due to limited contextual diversity. In contrast, models trained on MSCOCO and FramesDB may show lower performance on their respective test sets, given the complexity of these datasets, but adapt better to other datasets because they capture a wider range of contextual variation. Note that EMOTIC also consists of a subset of the Ade20K dataset. However, we exclude it from our analysis due to insufficient sample size.

Figure 3 shows a data imbalance, with MSCOCO dominating in size, while Figure 11 highlights differences in label distributions. Negative emotions such as pain, anger, and sadness are more prevalent in EmoDB. In contrast, MSCOCO and FramesDB are skewed toward excitement, confidence, anticipation, and engagement. Quantitatively, JS and TV distances confirm these trends. MSCOCO is most similar to FramesDB (JS: 0.141, TV: 0.162). Whereas EmoDB is distinct, with large JS and TV distances from MSCOCO (JS: 0.335, TV: 0.384) and FramesDB (JS: 0.288, TV: 0.346).

4 Experimental Setting

Models. We consider two classes of models: (i) supervised CAER models trained end-to-end on labelled data, and (ii) zero-shot large (vision) language models without task-specific fine-tuning.

We adopt the EMOTIC architecture [28] as our supervised baseline with the original hyperparameters provided by the authors. It consists of a two-branch convolutional neural network (CNN) designed to jointly model bodily expressions and contextual information. We use the original codebase[1] to train the model and to compute category-wise thresholds for assigning categorical labels to each image during test.

Among the zero-shot models, we use Llama3.1 8B [20] and LLaVA 7B [34], which are open-source large language and vision language models, respectively. For Llama3.1, we first generate narrative captions for each image using NarraCap [13]. Then the generated caption is provided to the model as part of the prompt. In contrast, the input to LLaVA is an image containing a red bounding box indicating the person whose emotion is to be inferred. Both models directly output the categorical labels. We follow the prompting strategy in [13] to allow for comparison with prior work. The experiments are run using Ollama [2] Python.

Metrics. For supervised CAER models, we evaluate predictive performance using mean average precision (mAP) for categorical emotion labels, a threshold-independent metric. Given that EMOTIC also provides continuous annotations, we further compute the mean absolute error for valence, arousal, and dominance (VAD) scores to assess regression performance. Following prior work [13], for both supervised and zero-shot models, we report the F1 score, the Hamming loss (i.e., fraction of incorrectly predicted labels per sample), and the subset accuracy (i.e., proportion of samples for which the predicted label set exactly matches the ground truth).

5 Results

In this section, we first examine how contextual shifts affect the supervised EMOTIC model. We then present results for zero-shot

[1]EMOTIC: https://github.com/Tandon-A/emotic
[2]Ollama: https://ollama.com/

Table 1: Overview of supervised EMOTIC model performance across the contextual subgroups.

(a) ID-OOD performance difference measured using mAP.

Context	Model-adj. Cohen's d	t(df)	p
Age: A, T, K	1.048	4.682 (41)	<.001***
Gender: M, F	1.065	2.996 (17)	<.01**
Situations: C, U	1.393	6.186 (13)	<.001***
C: Pu, Pr, Co	1.301	5.672 (37)	<.001***
U: Ru, Ur, So	0.533	1.811 (41)	0.077
Dataset: Ms, Emo, Fr	1.119	4.277 (41)	<.001***

(b) ID-OOD performance difference measured using F1-score.

Context	Model-adj. Cohen's d	t(df)	p
Age: A, T, K	0.458	1.534 (41)	0.133
Gender: M, F	0.546	1.238 (18)	0.231
Situations: C, U	1.188	4.450 (13)	<.001***
C: Pu, Pr, Co	1.022	3.843 (37)	<.001***
U: Ru, Ur, So	0.375	1.216 (41)	0.231
Dataset: Ms, Emo, Fr	1.270	6.535 (41)	<.001***

Note. Abbreviations - A: Adult, T: Teenager, K: Kid, M: Male, F: Female, C: Constrained, U: Unconstrained, Pu: Public, Pr: Private, Co: Commercial, Ru: Rural, Ur: Urban, So: Social, Ms: MSCOCO, Emo: EmoDB, Fr: FramesDB.

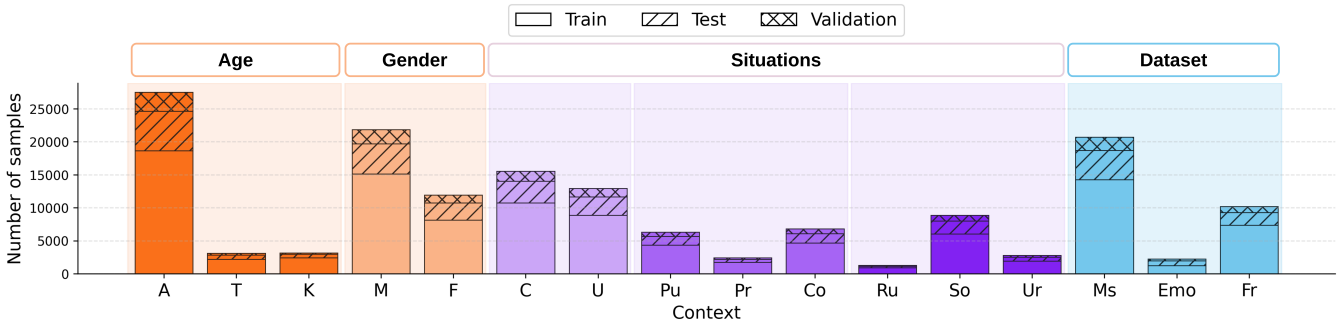


Figure 3: Data distribution across context factors.

(vision) language models under the same contextual variations and compare their behaviour to that of the supervised model.

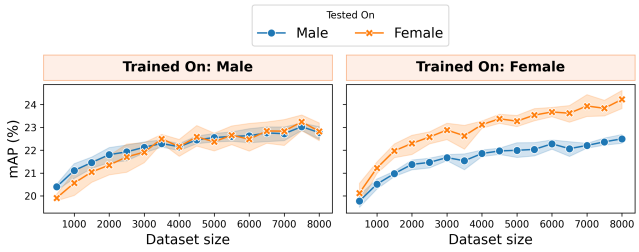
5.1 Analysis of Supervised Models

We evaluate the generalisation of the supervised EMOTIC dataset model by training it on a specific subgroup and testing it both on the same subgroup (in-distribution, i.e., ID) and on different subgroups (out-of-distribution, i.e., OOD) within each contextual dimension. To ensure a fair comparison despite dataset imbalances (Section 3), we control for training sample sizes by subsampling a fixed number of instances from each subgroup (e.g., male and female) of a contextual dimension (e.g., gender). Furthermore, the experiments are repeated across five seeds and multiple dataset sizes. This setup allows us to address two key questions for each contextual dimension: **Q1**. *How performance differs between ID and OOD at the largest training dataset size?* **Q2**. *How does the ID-OOD performance gap evolve with increasing training data?*

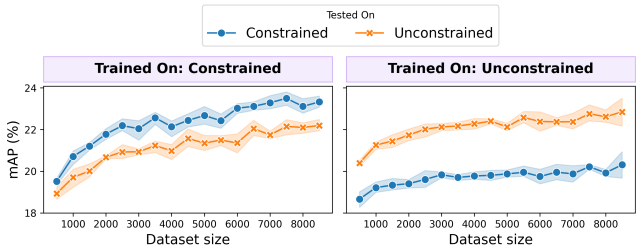
Q1: Generalisation gap. At a high level, the results show a consistent generalisation gap between ID and OOD settings, with substantially larger effects for mAP than F1-score. We report the t-statistics and significance of the ID–OOD performance gap at the largest dataset size in Tables 1a and 1b, estimated using a linear mixed model with training group and experimental seed as random blocking factors. We further quantify effect size using a model-adjusted Cohen's d (0.2 small, 0.5 medium, 0.8 large), which is obtained by standardising the outcome (mean $\mu = 0$, standard deviation $\sigma = 1$) and extracting the fixed-effect coefficient corresponding to the ID–OOD contrast. This reflects the shift in terms of standard deviations between conditions while accounting for variability due to training group and seed.

For mAP, all contexts exhibit large effect sizes, indicating very large and highly significant performance drops from ID to OOD. In contrast, for F1-scores, effect sizes are notably smaller. While situations and datasets remain significant with large effects, age and gender are not statistically significant. Furthermore, subgroups of the situation context reveal distinct patterns across both metrics. Within constrained settings, the generalisation gap is large and significant, indicating variability across private, public, and commercial contexts. In contrast, unconstrained settings show weaker, non-significant effects, suggesting greater transferability across rural, urban, and social environments.

We provide extended results in Appendix E, providing deeper insights. The results show asymmetric generalisation gaps across the subgroups of some context factors. For age (Table 4), models trained on kids or teenagers generalise better to adults than to each other, which aligns with the larger label distribution differences



(a) Across genders.



(b) Across situations.

Figure 4: Learning curves for EMOTIC model. The shaded region shows a 95% confidence interval.

between kids and teenagers. However, within the dataset context (Table 19), models trained on FramesDB perform well on MSCOCO, whereas the reverse transfer is weaker. This highlights that differences in label distribution alone do not fully explain generalisation. Furthermore, models trained on FramesDB and MSCOCO generalise well to EmoDB, while models trained on EmoDB transfer poorly. Also, within the unconstrained setting (Table 16), regardless of the subgroup the model is trained on, it performs better on the rural subgroup. Although we can hypothesise about the factors that might have helped or hindered the generalisability in the observations above, identifying them precisely remains an open problem.

Q2: Effect of training set size. Across all contexts, we observed a statistically significant ID-OOD performance gap at every dataset size. Moreover, increasing data either amplifies existing differences or leaves them largely unchanged, depending on the context. For the analysis, we perform a three-factor ANOVA with training group, testing group, and training size as fixed effects and seed as a random blocking factor, using mAP as the dependent variable. To directly analyse the generalisation gap across training set sizes, we further perform a planned contrast comparing ID and OOD performance, with significance assessed using a t-test. Detailed statistical results are provided in Appendix E.

For age, the performance gap initially increases with dataset sizes and then plateaus, as indicated by the increasing t-statistic in Table 6b. This indicates that additional data initially amplifies the separation between ID and OOD performance before stabilising, as also reflected in the learning curves (Figure 12). Furthermore, the ID performance consistently remains highest, and transfer from kids or teenagers to adults is stronger than transfer between kids and teenagers, suggesting stable generalisation patterns across dataset sizes. Across genders, the ID-OOD performance gap remains relatively stable regardless of dataset size (Table 9b). However, the learning curves (Figure 4a) reveal subgroup-specific asymmetry. The models trained on male data generalise well to females, whereas models trained on females exhibit a widening performance gap as more data are added, explaining the overall stable trend.

For the situational context, the ID-OOD performance gap shows a subtle increase with dataset size, as shown by the planned contrasts in Table 12b. However, the learning curve (Figure 4b) shows that this gap varies across subgroups. Models trained on the unconstrained subgroup exhibit a large but stable performance gap, whereas those trained on the constrained subgroup show a growing gap as the dataset size increases. Within subgroups of the constrained settings, the performance gap increases sharply with more data (Table 15b). Learning curves (Figure 15) further indicate that models trained on public and commercial subgroups can transfer between each other. However, when applied to the private subgroup, performance often degrades, worsening with larger training set sizes.

For the dataset context, the ID–OOD gap is consistently larger than in other context dimensions and increases with dataset size (Table 21b). While overall performance improves slightly, the relative patterns remain unchanged (Figure 16): (i) performance on EmoDB is always better than others, irrespective of the training dataset; (ii) models transfer well between MSCOCO and FramesDB; (iii) performance of the model trained on EmoDB drops significantly when applied to MSCOCO or FramesDB.

5.2 Analysis of Zero-shot Models

We evaluate the zero-shot performance of (vision) language models across subgroups within each contextual dimension by assessing them independently on each subgroup. Performance is estimated via bootstrap resampling, where test instances are sampled with replacement 100 times. Here, we focus only on the F1-score, as zero-shot models output categorical labels over which mAP cannot be computed. The supervised EMOTIC dataset model is trained on the full training set, with performance averaged over five random seeds. We then report the mean performance for all the models with 95% confidence interval in Figure 5. This setup enables us to address two key questions for each contextual dimension: **Q1**. *How zero-shot models compare to the supervised EMOTIC model?* **Q2**. *How performance varies across subgroups for both approaches.*

Q1: Zero-shot vs supervised. Across all contextual dimensions, the supervised EMOTIC model consistently outperforms the zero-shot vision–language models (Figure 5), suggesting that zero-shot models struggle to capture fine-grained affective cues without supervision. This indicates a clear advantage of task-specific training over large-scale pre-training. While LLaVA generally outperforms Llama3.1 among zero-shot approaches, both remain substantially below EMOTIC across all subgroups. This gap is particularly pronounced for kids and in socially constrained settings.

Q2: Performance across subgroups. Overall, both supervised and zero-shot models exhibit similar sensitivity to contextual variation, indicating limited robustness to real-world diversity. We also observe a largely consistent ranking across contextual subgroups,

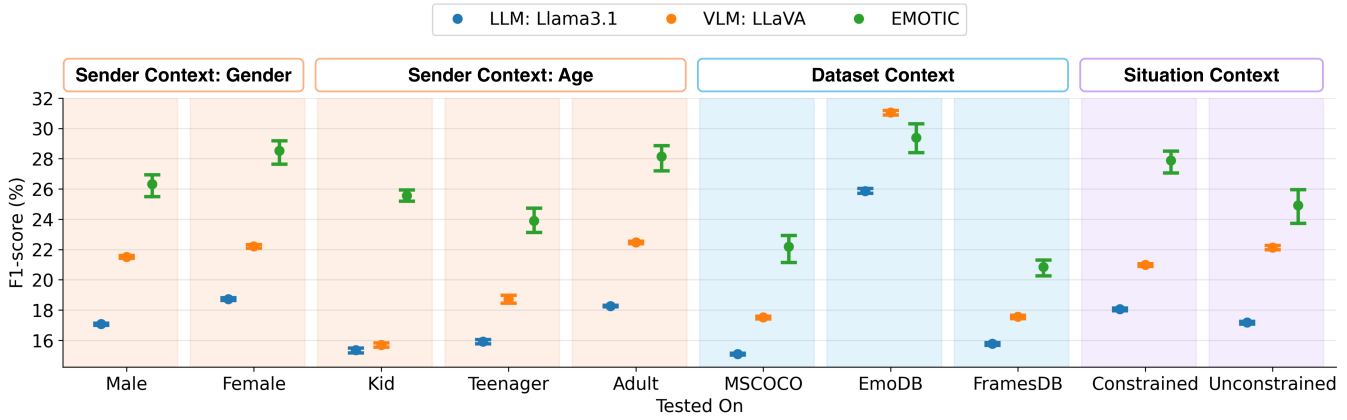


Figure 5: Comparison of supervised and zero-shot models. Error bars show 95% confidence intervals.

with EMOTIC performing best, followed by LLaVA, then Llama3.1. Across age groups, all models perform better for adults than for kids and teenagers. This may be attributed to the overrepresentation of adults in affective datasets. Moreover, while EMOTIC shows a larger decline for teenagers, zero-shot models struggle more with kids. Across genders, all models work better on female than male subgroups. For situational context, EMOTIC and Llama3.1 perform better in more constrained settings, whereas LLaVA shows an advantage in unconstrained environments. Lastly, across datasets, all models perform best on EmoDB, which aligns with its relatively lower complexity as discussed in Section 3. Notably, this is the only case where a zero-shot model, LLaVA, surpasses the supervised approach. This suggests that while zero-shot models can effectively leverage clear expressive cues, they struggle with more complex social scenarios that require integrating behavioural and contextual cues. Extended results are provided in Appendix E.

6 Discussion

We first summarise our key observations. Then we outline some limitations of our work and avenues for future research.

6.1 Key Insights

We jointly investigated the performance of two types of CAER models, namely the supervised EMOTIC model and zero-shot (vision) language models, while focusing on variations within three primary sources of contextual information: who expresses the emotion, where the emotion occurs, and how the data is collected

From the data analysis, we obtained three key insights. First, significant data imbalances occur across contextual dimensions, with under-representation of kids, teenagers, females, and private and rural settings. Such gaps likely stem from practical and ethical constraints in data collection. Nonetheless, such imbalances can bias model learning toward dominant subgroups. Second, label distributions vary substantially across subgroups, revealing that emotions are not uniformly expressed, perceived or captured across contexts. For example, within the sender context, kids are often associated with positive, low-arousal emotions (e.g., affection), teenagers with high-arousal states (e.g., excitement), and adults with more mixed or socially complex emotions (e.g., disconnection and engagement). Furthermore, the contextual structure is hierarchical, with differences emerging at both coarse- and fine-grained levels. For example, while constrained and unconstrained settings differ broadly, additional variation exists within each setting (e.g., public vs private, rural vs urban). Additionally, CAER datasets also introduce shifts in label distributions tied to how the data are collected and annotated. For instance, while EmoDB is skewed toward negative emotions, MSCOCO and FramesDB capture more positive emotional states. Third, the datasets differ not only in label distributions but also in image composition and complexity. While EmoDB is more controlled, featuring clear facial expressions, MSCOCO and FramesDB are more diverse and unconstrained.

From the model analysis, we observe persistent performance gaps across all contextual dimensions with both supervised and zero-shot models. The supervised CAER model consistently performs better in-distribution than out-of-distribution, with larger effects on mAP than on F1-scores. This divergence may be partly attributed to the threshold selection strategy and to the data and label distributions of the validation set on which the thresholds are computed. However, we do not explicitly account for these factors, as (i) ensuring balanced class representation in the validation set is non-trivial given multiple labels per instance [52], and (ii) formulating an effective threshold selection strategy requires dedicated analysis. Furthermore, we observe that generalisation is often asymmetric, indicating that transferability depends not only on label distributions but also on intrinsic data characteristics such as variability and scene composition. Additionally, increasing training data did not resolve the ID–OOD gap. Instead, it either amplified existing disparities (e.g. across genders) or left them largely unchanged (e.g. within the unconstrained settings), suggesting that scaling does not address the gap. Subgroup-level analysis further highlights context-specific transfer patterns across training set sizes, such as better generalisation to adults when trained on kids or teenagers and poor transfer of a model trained on EmoDB to other subdatasets, reinforcing the role of dataset complexity.

For zero-shot models, the results show a clear performance gap relative to supervised approaches, with LLaVA outperforming

Llama3.1, but both falling short of EMOTIC. Importantly, zero-shot models exhibit similar sensitivity to contextual variation, indicating that large-scale pre-training does not inherently offer robustness. Performance across subgroups follows consistent trends for both approaches (e.g., performance on EmoDB is consistently easier).

Overall, our findings underscore that context is a fundamental source of variability in emotion recognition, shaped by both demographic and environmental factors. They indicate that while CAER datasets can capture this variation to some extent, they also introduce shifts in intrinsic data properties and in label distributions tied to how the data is collected and annotated. We show that while the supervised model struggles to generalise across these variations, even with increasing dataset sizes, zero-shot models exhibit similar weaknesses. Overall, these findings indicate that robustness to contextual shifts remains an open challenge for CAER systems. Additionally, it suggests that scaling data collection efforts alone is insufficient to address it. Thus, these results motivate a shift toward context-aware data collection and model evaluation to develop reliable CAER systems for real-world applications.

6.2 Limitations and Future Work

Dataset and modelling paradigms. Although our analysis focuses on the EMOTIC dataset and a selected set of supervised and zero-shot models, it can be extended to other CAER datasets, such as HECO [60], to alternative supervised architectures such as EMOTICON [36] and CAER-Net [29] and to other zero-shot models like Qwen [58] and Gemma [54]. However, our goal was not to conduct an exhaustive survey of datasets and models. Instead, we aimed to provide empirical evidence that (i) even a widely used benchmark such as EMOTIC exhibits substantial imbalances across the contextual subgroups, and (ii) context-dependent performance gaps emerge across different modelling paradigms. Future work can build on this foundation to systematically examine data collection strategies and evaluation protocols in CAER.

Contextual factor. Our analysis considers two sender-contexts, namely age and gender, because these attributes are explicitly annotated in EMOTIC. Other relevant sender characteristics, such as ethnicity or cultural background, were beyond the scope of this study but constitute important directions for future research, particularly given their potential influence on emotion perception [27, 42]. With respect to situational context, our analysis relied on scene labels extracted using a pre-trained model. Manual inspection of the images suggests that this model can struggle when only limited background information is visible, and it occasionally produces inaccurate scene predictions. Such inconsistencies may introduce noise into the contextual grouping and, consequently, affect the analysis. Future work should therefore consider incorporating human-annotated situation labels to ensure more reliable contextual categorisation. Finally, due to the absence of annotator metadata in EMOTIC, we were unable to examine the role of the perceiver's context in emotion annotation. Future data collection efforts should record annotator characteristics to enable analysis of how perceiver-related factors affect the generalisability of CAER models trained on such datasets [9].

Multimodal CAER. Our study focuses exclusively on the visual modality. However, with the growing availability of multimodal emotion recognition datasets [32, 45], it is now feasible to examine how contextual factors operate across visual, textual, and audio modalities. Given that we observe systematic contextual variation and performance gaps even within a single modality, the robustness of multimodal CAER systems cannot be assumed and needs to be explicitly analysed. Thus, future research should extend this analysis to multimodal CAER settings to investigate which contextual factors most strongly affect generalisability and robustness across different modalities. Furthermore, it would be valuable to examine how cues can be reliably extracted, how they interact across vision, language, and audio, and whether consistent cross-modal patterns can be identified that are stable across subgroups within each contextual dimension (e.g. situation). Such insights are essential for developing reliable CAER applicable in real-world, multimodal environments where these systems are ultimately intended to support human decision-making.

7 Conclusion

In this work, we jointly investigated the performance of supervised and zero-shot Context-Aware Emotion Recognition (CAER) models across variations in sender (who), situational (where), and dataset (how) contexts. Our findings highlight that context is a fundamental and structured source of variability in emotion recognition. These contextual dimensions are not only unevenly represented but also differ systematically in how emotional meaning is expressed and annotated. As a result, models risk learning context-specific biases tied to dominant subgroups rather than transferable representations. We therefore advocate for context-aware data collection, transparent reporting of subgroup distributions, and systematic evaluation protocols that explicitly account for contextual variability. Future work should move beyond scale and focus on understanding the data properties that enable generalisation, alongside developing CAER models that effectively utilise contextual signals from multimodal data to achieve reliable performance in real-world settings.

8 Safe and Responsible Innovation Statement

Emotion recognition from multimodal signals is highly context-dependent. However, current CAER systems show inconsistent performance across different demographic groups, situations, and datasets. These limitations, often exacerbated by imbalanced datasets, may lead to biased or unreliable outcomes (i.e., algorithmic discrimination), particularly in sensitive domains such as healthcare.

Our work contributes to the development of reliable emotion-aware systems by systematically examining the generalisability of CAER models. It highlights the importance of context-aware data collection and careful consideration of how systems perform in diverse real-world settings.

To support ethical and safe deployment, we advocate for approaches that explicitly integrate contextual information and prioritise fairness and transparency throughout the system lifecycle. Aligning with regulatory frameworks such as the EU AI Act, this work reflects a broader commitment to designing AI technologies that are socially responsible and can meaningfully support human well-being in digital interactions.

Acknowledgments

This research was (partially) funded by the Hybrid Intelligence Center, a 10-year programme funded by the Dutch Ministry of Education, Culture and Science through the Netherlands Organisation for Scientific Research, Hybrid Intelligence Centre, grant number 024.004.022.

References

[1] Hillel Aviezer, Noga Ensenberg, and Ran R Hassin. 2017. The inherently contextualized nature of facial emotion perception. *Current opinion in psychology* 17 (2017), 47–54.

[2] Lisa Feldman Barrett. 2022. Context reconsidered: Complex signal ensembles, relational meaning, and population thinking in psychological science. *American Psychologist* 77, 8 (2022), 894.

[3] Joost Broekens, Bernhard Hilpert, Suzan Verberne, Kim Baraka, Patrick Gebhard, and Aske Plaat. 2023. Fine-grained affective processing capabilities emerging from large language models. In *2023 11th international conference on affective computing and intelligent interaction (ACII)*. IEEE, 1–8.

[4] Federico Cabitza, Andrea Campagner, and Martina Mattioli. 2022. The Unbearable (Technical) Unreliability of Automated Facial Emotion Recognition. *Big Data & Society* 9, 2 (July 2022), 205395172211295. doi:10.1177/20539517221129549

[5] Tara M Chaplin. 2015. Gender and emotion expression: A developmental contextual perspective. *Emotion Review* 7, 1 (2015), 14–21.

[6] Zebang Cheng, Zhi-Qi Cheng, Jun-Yan He, Kai Wang, Yuxiang Lin, Zheng Lian, Xiaojiang Peng, and Alexander Hauptmann. 2024. Emotion-llama: Multimodal emotion recognition and reasoning with instruction tuning. *Advances in Neural Information Processing Systems* 37 (2024), 110805–110853.

[7] Nicholas Cummins, Stefan Scherer, Jarek Krajewski, Sebastian Schnieder, Julien Epps, and Thomas F Quatieri. 2015. A review of depression and suicide risk assessment using speech analysis. *Speech communication* 71 (2015), 10–49.

[8] Bernd Dudzik, Tiffany Matej Hrkalovic, Chenxu Hao, Chirag Raman, and Masha Tsfasman. 2024. Indeterminacy in Affective Computing: Considering Meaning and Context in Data Collection Practices. In *2024 12th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*. 181–185. doi:10.1109/ACIIW63320.2024.00036

[9] Bernd Dudzik, Michel-Pierre Jansen, Franziska Burger, Frank Kaptein, Joost Broekens, Dirk KJ Heylen, Hayley Hung, Mark A Neerincx, and Khiet P Truong. 2019. Context in human emotion perception for automatic affect detection: A survey of audiovisual databases. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 206–212.

[10] Bernd Dudzik, Dennis Küster, David St-Onge, and Felix Putze. 2022. The 4th Workshop on Modeling Socio-Emotional and Cognitive Processes from Multimodal Data In-the-Wild (MSECP-Wild). In *Proceedings of the 2022 International Conference on Multimodal Interaction* (Bengaluru, India) *(ICMI '22)*. Association for Computing Machinery, New York, NY, USA, 803–804. doi:10.1145/3536221.3564029

[11] Paul Ekman. 1992. Are there basic emotions? (1992).

[12] Yasaman Etesam. 2025. Vision language models for environmental and emotional awareness. (2025).

[13] Yasaman Etesam, Özge Nilay Yalçın, Chuxuan Zhang, and Angelica Lim. 2024. Contextual emotion recognition using large vision language models. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 4769–4776.

[14] Yingruo Fan, Jacqueline CK Lam, and Victor OK Li. 2021. Demographic effects on facial emotion expression: an interdisciplinary investigation of the facial action units of happiness. *Scientific reports* 11, 1 (2021), 5214.

[15] Ali Pourramezan Fard, Mohammad Mehdi Hosseini, Timothy D Sweeny, and Mohammad H Mahoor. 2025. Affectnet+: A database for enhancing facial expression recognition with soft-labels. *IEEE Transactions on Affective Computing* (2025).

[16] Vanessa Farsadaki and Charla Griffy-Brown. 2025. AI affective computing and behavioral health. *Frontiers in Computer Science* 7 (2025), 1692728.

[17] Qinquan Gao, Hanxin Zeng, Gen Li, and Tong Tong. 2021. Graph reasoning-based emotion recognition network. *IEEE Access* 9 (2021), 6488–6497.

[18] Srishti Goel, Julian Jara-Ettinger, Desmond C Ong, and Maria Gendron. 2024. Face and context integration in emotion inference is limited and variable across categories and individuals. *Nature Communications* 15, 1 (2024), 2443.

[19] Clare Grall and Emily S Finn. 2022. Leveraging the power of media to drive cognition: A media-informed approach to naturalistic neuroscience. *Social Cognitive and Affective Neuroscience* 17, 6 (2022), 598–608.

[20] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783* (2024).

[21] Bin Han and Jonathan Gratch. 2025. Salience Adjustment for Context-Based Emotion Recognition. In *2025 IEEE 19th International Conference on Automatic Face and Gesture Recognition (FG)*. IEEE, 1–6.

[22] Bin Han, Jessie Hoegen, Su Lei, Gale Lucas, Danielle Shore, Brian Parkinson, and Jonathan Gratch. 2025. How Expression, Context and Perspective Determine Judgments of Emotion. *IEEE Transactions on Affective Computing* (2025), 1–12. doi:10.1109/TAFFC.2025.3618588

[23] Mirella Hladkỳ, Rúbia R Guerra, X Laura Cang, Karon E MacLean, Patrick Gebhard, and Tanja Schneeberger. 2024. Modeling the'Kiss my Ass'-Smile: Appearance and Functions of Smiles in Negative Social Situations. In *2024 12th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 116–124.

[24] Mohammad Mehdi Hosseini, Ali Pourramezan Fard, and Mohammad H Mahoor. 2025. Faces of fairness: Examining bias in facial expression recognition datasets and models. *arXiv preprint arXiv:2502.11049* (2025).

[25] Sean Dae Houlihan. 2022. *A computational framework for emotion understanding*. Ph. D. Dissertation. Massachusetts Institute of Technology.

[26] Sean Dae Houlihan, Max Kleiman-Weiner, Luke B Hewitt, Joshua B Tenenbaum, and Rebecca Saxe. 2023. Emotion prediction as computation over a generative theory of mind. *Philosophical Transactions of the Royal Society A* 381, 2251 (2023), 20220047.

[27] Birgit Koopmann-Holm and Jeanne L Tsai. 2014. Focusing on the negative: cultural differences in expressions of sympathy. *Journal of personality and social psychology* 107, 6 (2014), 1092.

[28] Ronak Kosti, Jose M Alvarez, Adria Recasens, and Agata Lapedriza. 2019. Context based emotion recognition using emotic dataset. *IEEE transactions on pattern analysis and machine intelligence* 42, 11 (2019), 2755–2766.

[29] Jiyoung Lee, Seungryong Kim, Sunok Kim, Jungin Park, and Kwanghoon Sohn. 2019. Context-aware emotion recognition networks. In *Proceedings of the IEEE/CVF international conference on computer vision*. 10143–10152.

[30] Weixin Li, Xuan Dong, and Yunhong Wang. 2021. Human emotion recognition with relational region-level analysis. *IEEE Transactions on Affective Computing* 14, 1 (2021), 650–663.

[31] Xinpeng Li, Xiaojiang Peng, and Changxing Ding. 2021. Sequential interactive biased network for context-aware emotion recognition. In *2021 IEEE International Joint Conference on Biometrics (IJCB)*. IEEE, 1–6.

[32] Zheng Lian, Haoyu Chen, Lan Chen, Haiyang Sun, Licai Sun, Yong Ren, Zebang Cheng, Bin Liu, Rui Liu, Xiaojiang Peng, Jiangyan Yi, and Jianhua Tao. 2025. AffectGPT: A New Dataset, Model, and Benchmark for Emotion Understanding with Multimodal Large Language Models. In *Forty-second International Conference on Machine Learning*. https://openreview.net/forum?id=xmbdACI0xu

[33] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*. Springer, 740–755.

[34] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems* 36 (2023), 34892–34916.

[35] Andreas Marpaung and Avelino Gonzalez. 2017. Can an affect-sensitive system afford to be context independent?. In *International and Interdisciplinary Conference on Modeling and Using Context*. Springer, 454–467.

[36] Trisha Mittal, Pooja Guhan, Uttaran Bhattacharya, Rohan Chandra, Aniket Bera, and Dinesh Manocha. 2020. Emoticon: Context-aware multimodal emotion recognition using frege's principle. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 14234–14243.

[37] Christian Mumenthaler, David Sander, and Antony SR Manstead. 2018. Emotion recognition in simulated social interactions. *IEEE Transactions on Affective computing* 11, 2 (2018), 308–312.

[38] Desmond C. Ong. 2021. An Ethical Framework for Guiding the Development of Affectively-Aware Artificial Intelligence. In *2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 1–8. doi:10.1109/acii52823.2021.9597441

[39] Desmond C Ong, Jamil Zaki, and Noah D Goodman. 2015. Affective cognition: Exploring lay theories of emotion. *Cognition* 143 (2015), 141–162.

[40] Desmond C Ong, Jamil Zaki, and Noah D Goodman. 2019. Computational models of emotion inference in theory of mind: A review and roadmap. *Topics in cognitive science* 11, 2 (2019), 338–357.

[41] Rameswar Panda, Jianming Zhang, Haoxiang Li, Joon-Young Lee, Xin Lu, and Amit K Roy-Chowdhury. 2018. Contemplating visual emotions: Understanding and overcoming dataset bias. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 579–595.

[42] BoKyung Park, Jeanne L Tsai, Louise Chim, Elizabeth Blevins, and Brian Knutson. 2016. Neural evidence for cultural differences in the valuation of positive facial expressions. *Social Cognitive and Affective Neuroscience* 11, 2 (2016), 243–252.

[43] Rosalind W Picard. 2000. *Affective computing*. MIT press.

[44] Kaśka Porayska-Pomsta, Manolis Mavrikis, Sidney D'Mello, Cristina Conati, and Ryan SJd Baker. 2013. Knowledge elicitation methods for affect modelling in education. *International Journal of Artificial Intelligence in Education* 22, 3 (2013), 107–140.

[45] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. MELD: A Multimodal Multi-Party Dataset

for Emotion Recognition in Conversations. doi:10.48550/arXiv.1810.02508 arXiv:1810.02508 [cs].
[46] Ira J Roseman and Craig A Smith. 2001. Appraisal theory. *Appraisal processes in emotion: Theory, methods, research* (2001), 3–19.
[47] James A Russell. 1997. into faces: Resurrecting a dimensional-contextual perspective. *The psychology of facial expression* (1997), 295.
[48] Stuart Russell. 2022. Human-Compatible Artificial Intelligence.
[49] Rebecca Saxe and Sean Dae Houlihan. 2017. Formalizing emotion concepts within a Bayesian model of theory of mind. *Current opinion in Psychology* 17 (2017), 15–21.
[50] Klaus R Scherer. 2003. Vocal communication of emotion: A review of research paradigms. *Speech communication* 40, 1-2 (2003), 227–256.
[51] Tanja Schneeberger, Mirella Hladkỳ, Ann-Kristin Thurner, Jana Volkert, Alexander Heimerl, Tobias Baur, Elisabeth André, and Patrick Gebhard. 2023. The deep method: Towards computational modeling of the social emotion shame driven by theory, introspection, and social signals. *IEEE Transactions on Affective Computing* 15, 2 (2023), 417–432.
[52] Konstantinos Sechidis, Grigorios Tsoumakas, and Ioannis Vlahavas. 2011. On the stratification of multi-label data. In *Proceedings of the 2011th European Conference on Machine Learning and Knowledge Discovery in Databases - Volume Part III* (Athens, Greece) *(ECMLPKDD'11)*. Springer-Verlag, Berlin, Heidelberg, 145–158. doi:10.1007/978-3-642-23808-6_10
[53] Ala N Tak and Jonathan Gratch. 2023. Is gpt a computational model of emotion?. In *2023 11th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 1–8.
[54] Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, et al. 2025. Kimi-vl technical report. *arXiv preprint arXiv:2504.07491* (2025).
[55] Yan Wang, Wei Song, Wei Tao, Antonio Liotta, Dawei Yang, Xinlei Li, Shuyong Gao, Yixuan Sun, Weifeng Ge, Wei Zhang, et al. 2022. A systematic review on affective computing: Emotion models, databases, and recent advances. *Information Fusion* 83 (2022), 19–52.
[56] Jacob Whitehill, Zewelanji Serpell, Yi-Ching Lin, Aysha Foster, and Javier R Movellan. 2014. The faces of engagement: Automatic recognition of student engagementfrom facial expressions. *IEEE Transactions on Affective Computing* 5, 1 (2014), 86–98.
[57] Matthias J Wieser and Tobias Brosch. 2012. Faces in context: A review and systematization of contextual influences on affective face processing. *Frontiers in psychology* 3 (2012), 471.
[58] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388* (2025).
[59] Dingkang Yang, Zhaoyu Chen, Yuzheng Wang, Shunli Wang, Mingcheng Li, Siao Liu, Xiao Zhao, Shuai Huang, Zhiyan Dong, Peng Zhai, et al. 2023. Context de-confounded emotion recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 19005–19015.
[60] Dingkang Yang, Shuai Huang, Shunli Wang, Yang Liu, Peng Zhai, Liuzhen Su, Mingcheng Li, and Lihua Zhang. 2022. Emotion Recognition for Multiple Context Awareness. In *Computer Vision – ECCV 2022*, Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (Eds.). Vol. 13697. Springer Nature Switzerland, Cham, 144–162. doi:10.1007/978-3-031-19836-6_9 Series Title: Lecture Notes in Computer Science.
[61] Jiaxing Zhao, Qize Yang, Yixing Peng, Detao Bai, Shimin Yao, Boyuan Sun, Xiang Chen, Shenghao Fu, Xihan Wei, Liefeng Bo, et al. 2025. Humanomni: A large vision-speech language model for human-centric video understanding. *arXiv preprint arXiv:2501.15111* (2025).
[62] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2018. Places: A 10 Million Image Database for Scene Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40, 6 (June 2018), 1452–1464. doi:10.1109/TPAMI.2017.2723009
[63] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. 2019. Semantic understanding of scenes through the ade20k dataset. *International journal of computer vision* 127, 3 (2019), 302–321.

A Annotating Situation Context

Table 2: Situation categories with context relevance.

Category	Description	Examples
Constrained–Private	Enclosed spaces associated with privacy and personal routines.	Bedroom, bathroom
Constrained–Public	Enclosed spaces designed for education, work, or shared social activity.	Classroom, conference room
Constrained–Commercial	Indoor environments centred on consumption, leisure, or performance.	Restaurant, coffee shop
Unconstrained-Urban	Outdoor environments dominated by built infrastructure and transport systems.	Street, parking lot
Unconstrained-Social	Outdoor urban spaces are designed for gathering, leisure, or cultural activity.	Plaza, amphitheatre
Unconstrained-Rural	Human-modified natural spaces tied to farming, livestock, and village life.	Farm, pasture

Following prior work [41, 59], we first utilised the ResNet-18 model pre-trained on the Places365 [62] dataset to predict the scene categories from images. Thereafter, we performed clustering on the labels using GPT-5.2 by giving it the following prompt:

> *Consider the places in the following list: <places365-labels>. Group them into non-overlapping categories, such as 'nature' and 'urban', that are relevant to emotion perception.*

The example list of categories ('nature', and 'urban') is based [62][Figure 1]. This resulted in 10 categorical labels as seen in Table 2. We further used GPT-5.2 to assign each of the 365 place labels to one of the categories using the following prompt:

> *Consider the following categories: <categories>. Assign the following labels to one of these categories. Make sure the lists are not overlapping.*

Then, having obtained categorical labels for all images in the dataset, we analyse below whether variations between the two largest categories, namely 'constrained' and 'unconstrained', impact the generalisability of the EMOTIC model. To better understand the effect, we further analyse the performance differences between the subgroups for both 'constrained' and 'unconstrained'.

B Additional information on zero-shot models

For Llama3.1 [20], we first generate narrative captions for each image using NarraCap [13]. Then the generated caption is provided to the language model with the following prompt:

> *<caption>. From <list-of-emotions>, pick the top labels that this person is feeling at the same time.*

For LLaVA [34], we provide the image along with a red bounding box indicating the target person, and prompt the model as follows:

> *Write a caption describing the person in the red bounding box. Then from <list-of-emotions>, pick the top labels that this person is feeling at the same time.*

C Additional information about the EMOTIC model

It consists of a two-branch convolutional neural network (CNN) designed to jointly model bodily expressions and contextual information. The first branch extracts body-related features from the cropped person image, while the second branch captures scene-level contextual cues from the full image. Both branches are initialised using ResNet-18 backbones pre-trained on ImageNet and Places365, respectively. After initialisation, these feature extractors are kept frozen. The extracted representations are concatenated and fed into a fusion network with two output heads: one predicting categorical emotion labels and another predicting continuous valence–arousal–dominance scores. In our experiments, we train only the fusion network using the original codebase[3].

[3]EMOTIC: https://github.com/Tandon-A/emotic

D Data and Label Distributions

Table 3: Data distribution across context factors.

Context	Subgroups	Number of data samples		
		train	test	validation
Age	Adult	18628	5999	2892
	Teenager	2216	645	247
	Kid	2422	559	176
Gender	Male	15137	4569	2133
	Female	8129	2634	1182
Situations	Constrained	10725	3284	1524
	Unconstrained	8879	2786	1282
Constrained	Public	4345	1345	603
	Private	1726	489	203
	Commercial	4654	1450	718
Unconstrained	Rural	931	230	112
	Social	6019	1963	877
	Urban	1929	593	293
Dataset	MSCOCO	14276	4397	2029
	EmoDB	1228	735	275
	FramesDB	7348	1911	937

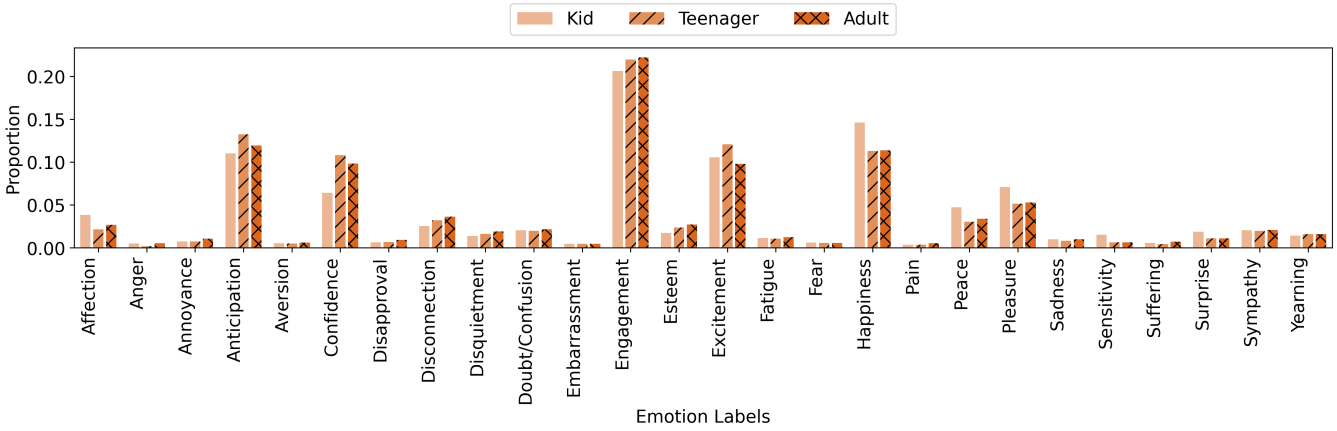


Figure 6: Label distribution across age groups.



Figure 7: Label distribution across genders.

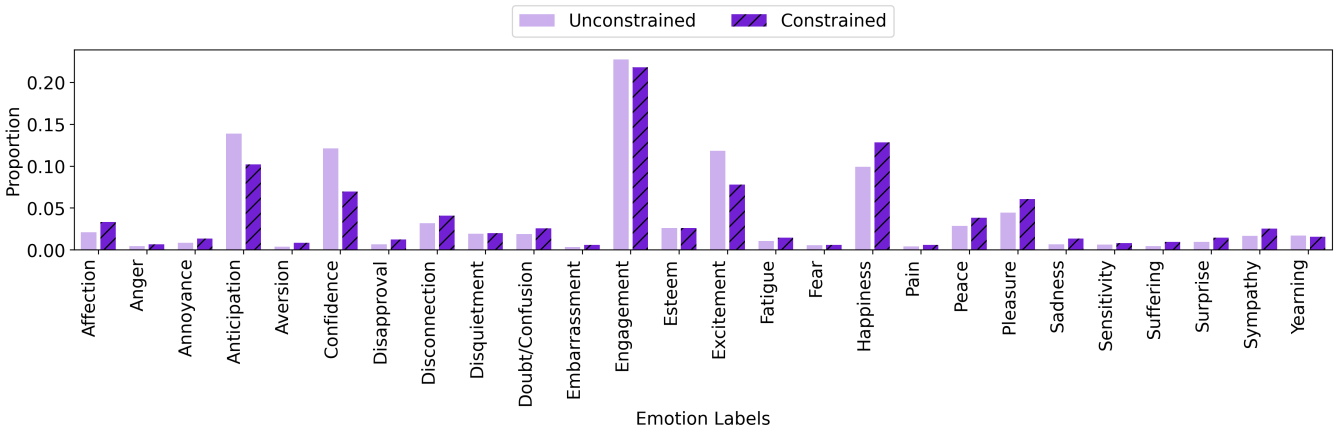


Figure 8: Label distribution across situations.

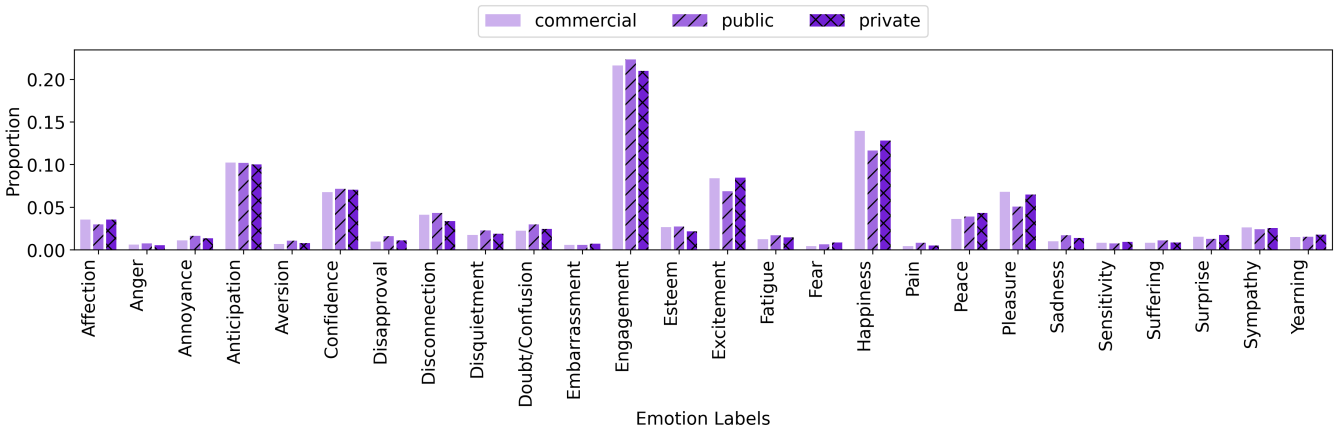


Figure 9: Label distribution within constrained settings.

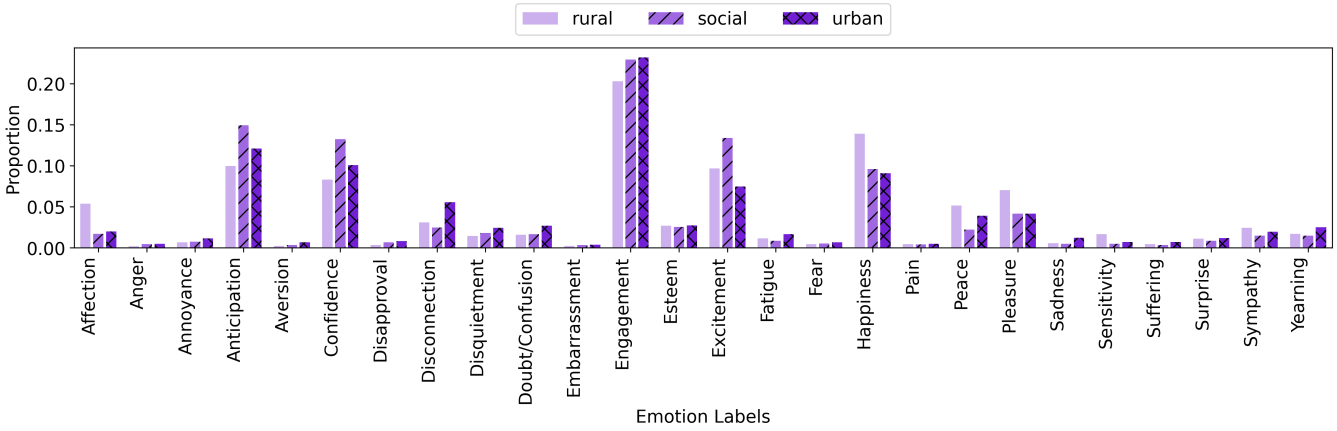


Figure 10: Label distribution within unconstrained settings.

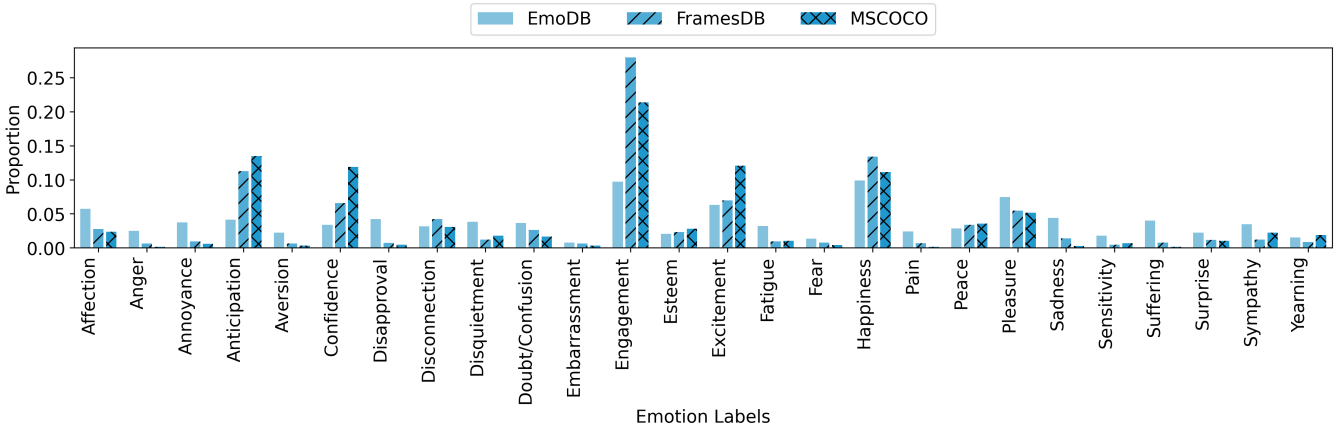


Figure 11: Label distribution across datasets.

E Extended Results

E.1 Results on Sender Context: Age

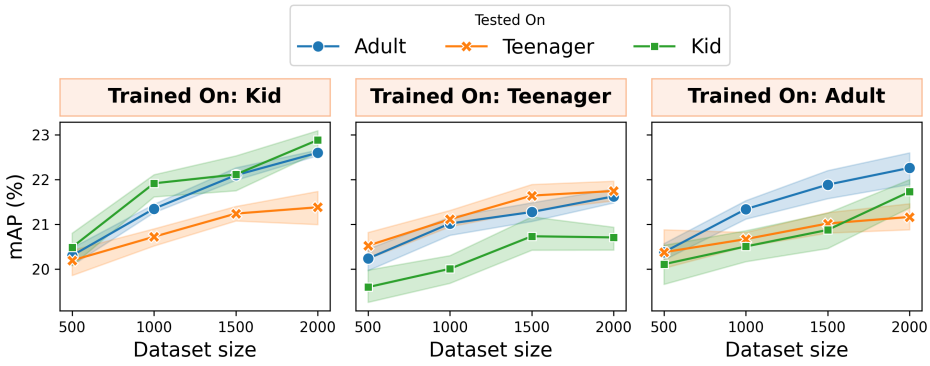


Figure 12: Performance of EMOTIC model across age groups and training set sizes.

Table 4: Performance of EMOTIC model across age groups. Style: Metric (± SE). ID and best performance.

Trained On	Tested On	Threshold-independent		Threshold-dependent		
		mAP (%) ↑	VAD ↓	F1-score (%) ↑	Hamming (%) ↓	S-acc (%) ↑
Kid	Kid	**22.881** (± 0.132)	**1.048** (± 0.005)	24.076 (± 0.259)	20.567 (± 0.388)	0.537 (± 0.16)
	Teenager	21.379 (± 0.21)	1.079 (± 0.004)	22.137 (± 0.304)	**19.252** (± 0.439)	**1.426** (± 0.283)
	Adult	22.597 (± 0.036)	1.067 (± 0.005)	**25.98** (± 0.15)	21.296 (± 0.507)	1.077 (± 0.215)
Teenager	Kid	20.705 (± 0.148)	1.046 (± 0.001)	22.267 (± 0.341)	23.389 (± 0.229)	0.716 (± 0.16)
	Teenager	**21.743** (± 0.127)	**1.014** (± 0.004)	22.624 (± 0.437)	**19.937** (± 0.357)	**1.612** (± 0.223)
	Adult	21.623 (± 0.092)	1.043 (± 0.002)	**23.625** (± 0.236)	21.384 (± 0.415)	1.164 (± 0.142)
Adult	Kid	21.726 (± 0.173)	1.064 (± 0.002)	23.179 (± 0.377)	21.804 (± 0.38)	1.073 (± 0.188)
	Teenager	21.162 (± 0.161)	1.036 (± 0.005)	22.468 (± 0.508)	**18.961** (± 0.174)	**1.674** (± 0.227)
	Adult	**22.258** (± 0.197)	**1.027** (± 0.001)	**25.125** (± 0.312)	20.37 (± 0.215)	1.477 (± 0.157)

Table 5: Performance of CAER models across age groups. Style: Metric (± SE). Best performance.

Framework	Model	Tested On	F1-score (%) ↑	Hamming (%) ↓	S-acc (%) ↑
Trained	EMOTIC	Kid	24.318 (± 0.321)	25.996 (± 1.705)	0.871 (± 0.193)
		Teenager	23.36 (± 0.326)	**22.873** (± 1.797)	**2.078** (± 0.447)
		Adult	**26.53** (± 0.442)	24.419 (± 1.733)	1.507 (± 0.335)
LLM (zero-shot)	NarraCap + Llama3.1	Kid	15.349 (± 0.084)	**23.331** (± 0.047)	0.345 (± 0.025)
		Teenager	15.917 (± 0.067)	23.816 (± 0.043)	**0.62** (± 0.035)
		Adult	**18.267** (± 0.027)	24.631 (± 0.011)	0.225 (± 0.007)
	NarraCap + Mistral	Kid	15.295 (± 0.064)	**24.81** (± 0.045)	**0.157** (± 0.017)
		Teenager	15.766 (± 0.077)	26.011 (± 0.045)	0.0 (± 0.0)
		Adult	**17.825** (± 0.03)	27.045 (± 0.013)	0.086 (± 0.004)
VLM (zero-shot)	LLaVA	Kid	15.696 (± 0.072)	20.11 (± 0.032)	0.705 (± 0.033)
		Teenager	18.729 (± 0.136)	**19.894** (± 0.037)	**0.781** (± 0.04)
		Adult	**22.468** (± 0.04)	21.723 (± 0.013)	0.332 (± 0.008)

Table 6: Statistical analysis across age groups

(a) ANOVA results.

Effect	df_n	df_d	F	p
trainedOn	2	140	43.284	<.001***
testOn	2	140	23.594	<.001***
datasetSize	3	140	157.925	<.001***
trainedOn ∗ testOn	4	140	40.922	<.001***
trainedOn ∗ datasetSize	6	140	3.324	<.01**
testOn ∗ datasetSize	6	140	3.764	<.01**
trainedOn ∗ testOn ∗ datasetSize	12	140	1.692	0.0745

(b) Planned contrast ID vs OOD.

datasetSize	t (df)	p
overall	11.138 (140)	<.001***
500	2.912 (140)	<.001***
1000	6.610262 (140)	<.001***
1500	5.984147 (140)	<.001***
2000	6.769701 (140)	<.001***

E.2 Results on Sender Context: Gender

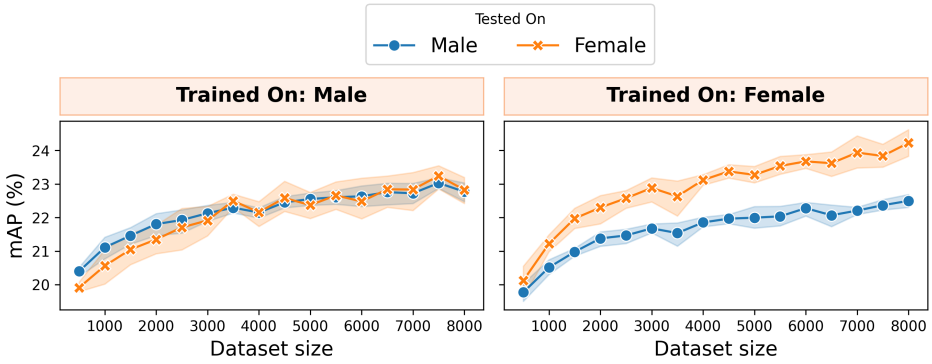


Figure 13: Performance of EMOTIC model across different genders and training set sizes.

Table 7: Performance of EMOTIC model across genders. Style: Metric (± SE). ID and best performance.

Trained On	Tested On	Threshold-independent		Threshold-dependent		
		mAP (%) ↑	VAD ↓	F1-score (%) ↑	Hamming (%) ↓	S-acc (%) ↑
Male	Male	22.771 (± 0.161)	**0.96** (± 0.004)	24.73 (± 0.32)	**19.037** (± 0.286)	**2.276** (± 0.182)
	Female	**22.827** (± 0.208)	1.027 (± 0.004)	**26.306** (± 0.44)	23.233 (± 0.314)	0.524 (± 0.056)
Female	Male	22.499 (± 0.111)	**1.008** (± 0.003)	24.371 (± 0.572)	**17.511** (± 0.291)	**2.127** (± 0.176)
	Female	**24.225** (± 0.223)	1.015 (± 0.003)	**27.859** (± 0.605)	21.22 (± 0.3)	0.721 (± 0.097)

Table 8: Performance of CAER models across different genders. Style: Metric (± SE). Best performance.

Framework	Model	Tested On	F1-score (%) ↑	Hamming (%) ↓	S-acc (%) ↑
Trained	EMOTIC	Male	24.93 (± 0.391)	**22.672** (± 1.811)	**2.084** (± 0.459)
		Female	**26.947** (± 0.432)	27.407 (± 1.613)	0.511 (± 0.115)
LLM (zero-shot)	NarraCap + Llama3.1	Male	17.073 (± 0.033)	**24.186** (± 0.014)	**0.431** (± 0.01)
		Female	**18.735** (± 0.05)	24.916 (± 0.02)	0.039 (± 0.004)
	NarraCap + Mistral	Male	17.009 (± 0.032)	26.91 (± 0.017)	**0.108** (± 0.005)
		Female	**18.114** (± 0.044)	**26.524** (± 0.021)	0.039 (± 0.004)
VLM (zero-shot)	LLaVA	Male	21.560 (± 0.046)	**21.392** (± 0.015)	0.271 (± 0.007)
		Female	**22.082** (± 0.063)	21.593 (± 0.021)	**0.63** (± 0.016)

Table 9: Statistical analysis across genders

(a) ANOVA results.

Effect	df_n	df_d	F	p
trainedOn	1	252	11.740	<.001***
testOn	1	252	143.873	<.001***
datasetSize	15	252	83.360	<.001***
trainedOn ∗ testOn	1	252	208.093	<.001***
trainedOn ∗ datasetSize	15	252	1.403	0.146
testOn ∗ datasetSize	15	252	2.528	<.01**
trainedOn ∗ testOn ∗ datasetSize	15	252	0.365	0.986

(b) Planned contrast ID vs OOD.

datasetSize	t (df)	p
overall	14.425 (252)	<.001***
500	2.251 (252)	<.001***
1000	3.365 (252)	<.001***
1500	3.775 (252)	<.001***
2000	3.691 (252)	<.001***
2500	3.565 (252)	<.001***
3000	3.809 (252)	<.001***
3500	2.385 (252)	<.001***
4000	3.379 (252)	<.001***
4500	3.419 (252)	<.001***
5000	3.915 (252)	<.001***
5500	3.877 (252)	<.001***
6000	4.148 (252)	<.001***
6500	3.936 (252)	<.001***
7000	4.334 (252)	<.001***
7500	3.368 (252)	<.001***
8000	4.481 (252)	<.001***

E.3 Results on Situation Context: Constrained vs Unconstrained

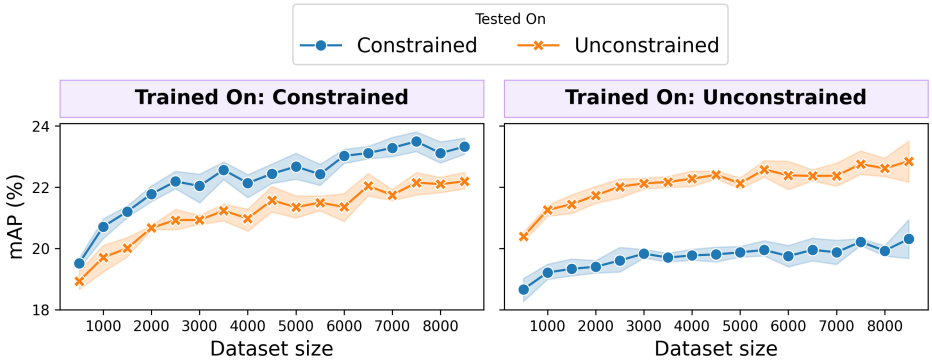


Figure 14: Performance of EMOTIC model across situations and training set sizes.

Table 10: Performance of EMOTIC model across different situations. Style: Metric (± SE). ID and best performance.

Trained On	Tested On	Threshold-independent		Threshold-dependent		
		mAP (%) ↑	VAD ↓	F1-score (%) ↑	Hamming (%) ↓	S-acc (%) ↑
Constrained	Constrained	**23.326** (± 0.158)	**1.004** (± 0.004)	**27.839** (± 0.325)	22.014 (± 0.407)	0.524 (± 0.075)
	Unconstrained	22.191 (± 0.16)	1.009 (± 0.012)	23.845 (± 0.238)	**18.29** (± 0.324)	**1.543** (± 0.166)
Unconstrained	Constrained	20.311 (± 0.358)	1.029 (± 0.003)	23.455 (± 0.769)	26.26 (± 0.975)	0.201 (± 0.037)
	Unconstrained	**22.849** (± 0.397)	**0.956** (± 0.007)	**24.393** (± 0.559)	**17.898** (± 0.671)	**2.915** (± 0.495)

Table 11: Performance of CAER models across different situations. Style: Metric (± SE). Best performance..

Framework	Model	Tested On	F1-score (%) ↑	Hamming (%) ↓	S-acc (%) ↑
Trained	EMOTIC	Constrained	**26.375** (± 0.427)	28.765 (± 1.61)	0.229 (± 0.051)
		Unconstrained	23.797 (± 0.375)	**20.497** (± 1.904)	**2.879** (± 0.643)
LLM (zero-shot)	NarraCap + Llama3.1	Constrained	**18.061** (± 0.044)	24.885 (± 0.019)	**0.34** (± 0.01)
		Unconstrained	17.176 (± 0.047)	**23.637** (± 0.019)	0.28 (± 0.009)
	NarraCap + Mistral	Constrained	**18.665** (± 0.048)	27.215 (± 0.019)	0.057 (± 0.004)
		Unconstrained	15.699 (± 0.041)	**25.8** (± 0.022)	**0.141** (± 0.008)
VLM (zero-shot)	LLaVA	Constrained	20.986 (± 0.048)	22.811 (± 0.024)	0.342 (± 0.01)
		Unconstrained	**22.133** (± 0.069)	**20.378** (± 0.02)	**0.419** (± 0.014)

Table 12: Statistical analysis across situations

(a) ANOVA results.

Effect	df_n	df_d	F	p
trainedOn	1	268	356.423	<.001***
testOn	1	268	210.661	<.001***
datasetSize	16	268	68.697	<.001***
trainedOn ∗ testOn	1	268	1732.480	<.001***
trainedOn ∗ datasetSize	16	268	8.216	<.001***
testOn ∗ datasetSize	16	268	0.568	0.906
trainedOn ∗ testOn ∗ datasetSize	16	268	1.429	0.102

(b) Planned contrast ID vs OOD.

datasetSize	t (df)	p
overall	41.623 (268)	<.001***
500	6.575 (268)	<.001***
1000	8.656 (268)	<.001***
1500	9.415 (268)	<.001***
2000	9.788 (268)	<.001***
2500	10.447 (268)	<.001***
3000	9.676 (268)	<.001***
3500	10.820 (268)	<.001***
4000	10.401 (268)	<.001***
4500	9.855 (268)	<.001***
5000	10.186 (268)	<.001***
5500	10.112 (268)	<.001***
6000	12.251 (268)	<.001***
6500	9.916 (268)	<.001***
7000	11.499 (268)	<.001***
7500	11.049 (268)	<.001***
8000	10.527 (268)	<.001***
8500	10.442 (268)	<.001***

E.4 Results on Situation Context: Public vs Private vs Commercial

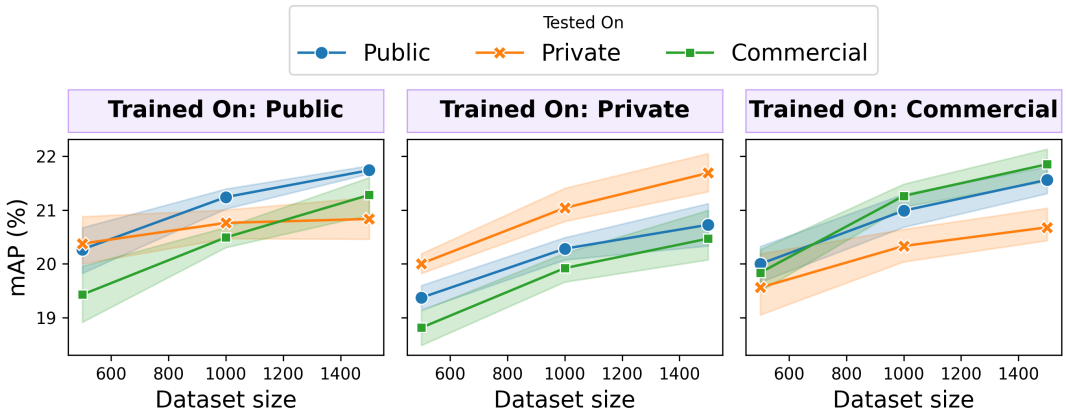


Figure 15: Performance of EMOTIC model within constrained settings and training set sizes.

Table 13: Performance of EMOTIC model within constrained settings. Style: Metric (± SE). ID and best performance.

Trained On	Tested On	Threshold-independent		Threshold-dependent		
		mAP (%) ↑	VAD ↓	F1-score (%) ↑	Hamming (%) ↓	S-acc (%) ↑
Public	Public	**21.74** (± 0.047)	1.083 (± 0.007)	**24.925** (± 0.228)	23.644 (± 0.217)	**0.431** (± 0.064)
	Private	20.835 (± 0.222)	1.157 (± 0.006)	23.19 (± 0.371)	23.802 (± 0.445)	0.409 (± 0.158)
	Commercial	21.28 (± 0.199)	**1.049** (± 0.003)	24.854 (± 0.164)	**22.546** (± 0.176)	0.359 (± 0.026)
Private	Public	20.727 (± 0.228)	1.108 (± 0.009)	23.573 (± 0.49)	24.218 (± 0.653)	**0.164** (± 0.064)
	Private	**21.689** (± 0.208)	1.146 (± 0.008)	**24.16** (± 0.293)	24.626 (± 0.528)	**0.164** (± 0.119)
	Commercial	20.469 (± 0.273)	**1.057** (± 0.01)	23.671 (± 0.477)	**23.31** (± 0.557)	0.138 (± 0.058)
Commercial	Public	21.561 (± 0.147)	1.089 (± 0.003)	24.387 (± 0.338)	23.06 (± 0.341)	**0.327** (± 0.08)
	Private	20.679 (± 0.185)	1.157 (± 0.006)	23.572 (± 0.501)	24.085 (± 0.31)	0.245 (± 0.119)
	Commercial	**21.85** (± 0.16)	**1.051** (± 0.005)	**25.935** (± 0.216)	**21.895** (± 0.155)	0.262 (± 0.04)

Table 14: Performance of CAER models within constrained settings. Style: Metric (± SE). Best performance.

Framework	Model	Tested On	F1-score (%) ↑	Hamming (%) ↓	S-acc (%) ↑
Trained	EMOTIC	Public	26.381 (± 0.491)	29.789 (± 1.59)	0.193 (± 0.048)
		Private	25.096 (± 0.41)	30.015 (± 1.519)	**0.423** (± 0.095)
		Commercial	**26.565** (± 0.398)	**27.394** (± 1.666)	0.198 (± 0.052)
LLM (zero-shot)	NarraCap + Llama3.1	Public	18.520 (± 0.065)	25.361 (± 0.031)	0.309 (± 0.017)
		Private	**18.595** (± 0.108)	**24.103** (± 0.039)	**0.618** (± 0.034)
		Commercial	16.741 (± 0.059)	24.669 (± 0.027)	0.271 (± 0.013)
	NarraCap + Mistral	Public	18.257 (± 0.068)	28.36 (± 0.034)	**0.084** (± 0.008)
		Private	**19.202** (± 0.122)	**26.083** (± 0.045)	0.0 (± 0.0)
		Commercial	18.237 (± 0.073)	26.517 (± 0.032)	0.071 (± 0.007)
VLM (zero-shot)	LLaVA	Public	20.094 (± 0.081)	23.946 (± 0.033)	0.315 (± 0.015)
		Private	**24.189** (± 0.131)	22.019 (± 0.048)	**0.472** (± 0.032)
		Commercial	20.205 (± 0.092)	**22.001** (± 0.026)	0.341 (± 0.015)

Table 15: Statistical analysis within constrained settings

(a) ANOVA results.

Effect	df_n	df_d	F	p
trainedOn	2	104	17.049	<.001***
testOn	2	104	6.837	<.01**
datasetSize	2	104	127.463	<.001***
trainedOn ∗ testOn	4	104	25.840	<.001***
trainedOn ∗ datasetSize	4	104	0.765	0.550
testOn ∗ datasetSize	4	104	3.213	<.05*
trainedOn ∗ testOn ∗ datasetSize	8	104	1.273	0.265

(b) Planned contrast ID vs OOD.

datasetSize	t (df)	p
overall	8.885 (104)	<.001***
500	3.430 (104)	<.001***
1000	5.531 (104)	<.001***
1500	6.427 (104)	<.001***

E.5 Results on Situation Context: Rural vs Urban vs Social

Table 16: Performance of EMOTIC model within unconstrained settings. Style: Metric (± SE). ID and best performance.

Trained On	Tested On	Threshold-independent		Threshold-dependent		
		mAP (%) ↑	VAD ↓	F1-score (%) ↑	Hamming (%) ↓	S-acc (%) ↑
Rural	Rural	**20.852** (± 0.563)	**1.225** (± 0.017)	**21.586** (± 0.369)	29.987 (± 1.102)	**0.261** (± 0.261)
	Urban	19.194 (± 0.078)	1.235 (± 0.011)	20.964 (± 0.269)	31.878 (± 1.126)	0.0 (± 0.0)
	Social	19.27 (± 0.09)	1.256 (± 0.021)	21.479 (± 0.385)	**28.612** (± 1.031)	0.031 (± 0.012)
Urban	Rural	**20.658** (± 0.164)	1.259 (± 0.022)	**21.425** (± 0.388)	24.973 (± 1.156)	0.261 (± 0.106)
	Urban	19.611 (± 0.133)	**1.256** (± 0.012)	21.333 (± 0.095)	23.273 (± 0.456)	0.27 (± 0.086)
	Social	18.901 (± 0.203)	1.319 (± 0.012)	21.016 (± 0.346)	**22.853** (± 0.923)	**0.316** (± 0.081)
Social	Rural	**21.082** (± 0.496)	1.201 (± 0.015)	22.043 (± 0.424)	23.458 (± 0.404)	0.348 (± 0.254)
	Urban	19.801 (± 0.351)	1.204 (± 0.014)	21.232 (± 0.574)	22.075 (± 0.382)	0.371 (± 0.083)
	Social	20.551 (± 0.076)	**1.2** (± 0.011)	**22.077** (± 0.069)	**17.943** (± 0.071)	**1.875** (± 0.208)

Table 17: Performance of CAER models within unconstrained settings. Style: Metric (± SE). Best performance.

Framework	Model	Tested On	F1-score (%) ↑	Hamming (%) ↓	S-acc (%) ↑
Trained	EMOTIC	Rural	**25.027** (± 0.547)	24.923 (± 1.815)	1.362 (± 0.306)
		Urban	22.255 (± 0.422)	23.114 (± 1.921)	0.675 (± 0.166)
		Social	23.538 (± 0.377)	**19.189** (± 1.913)	**3.722** (± 0.836)
LLM (zero-shot)	NarraCap + Llama3.1	Rural	16.475 (± 0.134)	24.07 (± 0.058)	0.0 (± 0.0)
		Urban	15.102 (± 0.091)	26.397 (± 0.039)	0.179 (± 0.016)
		Social	**17.32** (± 0.061)	**22.776** (± 0.025)	**0.345** (± 0.013)
	NarraCap + Mistral	Rural	**15.802** (± 0.126)	26.181 (± 0.062)	0.43 (± 0.043)
		Urban	15.272 (± 0.083)	27.938 (± 0.035)	**0.164** (± 0.017)
		Social	15.306 (± 0.044)	**25.142** (± 0.027)	0.097 (± 0.007)
VLM (zero-shot)	LLaVA	Rural	21.054 (± 0.202)	23.449 (± 0.059)	0.0 (± 0.0)
		Urban	17.902 (± 0.099)	23.11 (± 0.05)	0.295 (± 0.023)
		Social	**22.952** (± 0.087)	**19.238** (± 0.02)	**0.504** (± 0.016)

Table 18: Statistical analysis within unconstrained settings

(a) ANOVA results.

Effect	df_n	df_d	F	p
trainedOn	2	36	6.085	<.01**
testOn	2	36	19.485	<.001***
trainedOn * testOn	4	36	2.025	0.112

(b) Planned contrast ID vs OOD.

datasetSize	t (df)	p
overall	2.481 (32)	<.05*

E.6 Results on Dataset Context

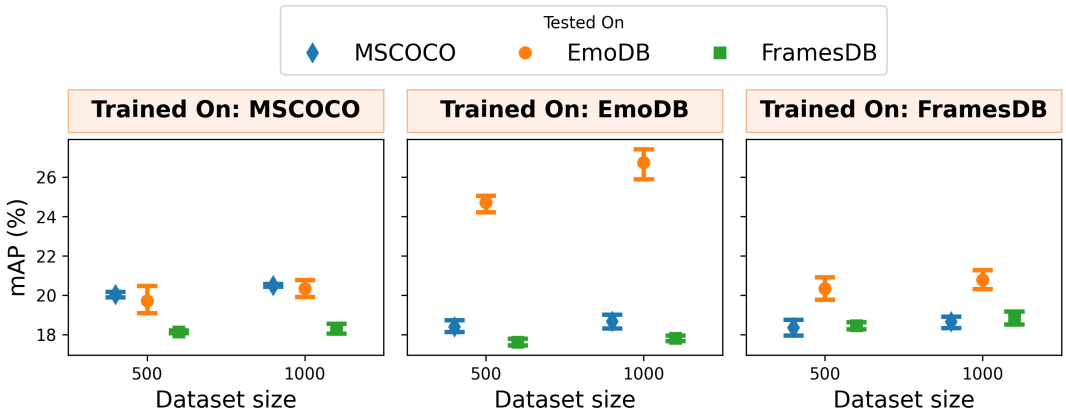


Figure 16: Performance of EMOTIC model across datasets, model types and training set sizes.

Table 19: Performance of EMOTIC model across datasets. Style: Metric (± SE). ID and best performance.

Trained On	Tested On	Threshold-independent		Threshold-dependent		
		mAP (%) ↑	VAD ↓	F1-score (%) ↑	Hamming (%) ↓	S-acc (%) ↑
MSCOCO	MSCOCO	**20.5** (± 0.039)	**1.043** (± 0.006)	**21.984** (± 0.127)	**19.259** (± 0.209)	**1.66** (± 0.168)
	EmoDB	20.341 (± 0.272)	1.509 (± 0.014)	19.049 (± 0.441)	28.87 (± 0.325)	0.0 (± 0.0)
	FramesDB	18.291 (± 0.143)	1.151 (± 0.027)	19.92 (± 0.273)	20.942 (± 0.298)	0.45 (± 0.054)
EmoDB	MSCOCO	18.695 (± 0.2)	1.331 (± 0.021)	21.225 (± 0.177)	**26.215** (± 0.86)	0.041 (± 0.013)
	EmoDB	**26.737** (± 0.429)	1.388 (± 0.013)	**30.273** (± 0.771)	29.262 (± 0.789)	**0.109** (± 0.079)
	FramesDB	17.833 (± 0.083)	**1.192** (± 0.019)	20.829 (± 0.32)	26.427 (± 0.912)	0.073 (± 0.035)
FramesDB	MSCOCO	18.675 (± 0.167)	1.301 (± 0.017)	20.281 (± 0.254)	20.886 (± 0.341)	0.35 (± 0.071)
	EmoDB	**20.784** (± 0.273)	1.515 (± 0.006)	20.662 (± 0.314)	29.372 (± 0.216)	0.0 (± 0.0)
	FramesDB	18.877 (± 0.197)	**1.072** (± 0.016)	**21.126** (± 0.272)	**20.365** (± 0.164)	**0.356** (± 0.061)

Table 20: Performance across different datasets. Style: Metric (± SE). Best performance.

Framework	Model	Tested On	F1-score (%) ↑	Hamming (%) ↓	S-acc (%) ↑
Trained	EMOTIC	MSCOCO	21.829 (± 0.303)	**21.53** (± 1.846)	**2.28** (± 0.504)
		EmoDB	**27.827** (± 0.478)	41.356 (± 1.084)	0.0 (± 0.0)
		FramesDB	20.471 (± 0.284)	24.424 (± 1.788)	0.405 (± 0.098)
LLM (zero-shot)	NarraCap + Llama3.1	MSCOCO	15.096 (± 0.03)	**23.989** (± 0.015)	0.298 (± 0.009)
		EmoDB	**25.868** (± 0.083)	24.485 (± 0.046)	**0.415** (± 0.027)
		FramesDB	15.773 (± 0.049)	25.435 (± 0.024)	0.207 (± 0.01)
	NarraCap + Mistral	MSCOCO	14.825 (± 0.034)	**26.637** (± 0.019)	0.085 (± 0.004)
		EmoDB	**24.458** (± 0.076)	27.41 (± 0.044)	**0.124** (± 0.012)
		FramesDB	16.478 (± 0.048)	26.877 (± 0.023)	0.05 (± 0.005)
VLM (zero-shot)	LLaVA	MSCOCO	17.516 (± 0.051)	**20.746** (± 0.016)	0.244 (± 0.008)
		EmoDB	**31.047** (± 0.073)	24.163 (± 0.055)	**0.639** (± 0.031)
		FramesDB	17.564 (± 0.056)	21.841 (± 0.022)	0.611 (± 0.02)

Table 21: Statistical analysis across datasets.

(a) ANOVA results.

Effect	df_n	df_d	F	p
trainedOn	2	68	72.859	<.001***
testOn	2	68	533.295	<.001***
datasetSize	1	68	27.918	<.001***
trainedOn ∗ testOn	4	68	200.676	<.001***
trainedOn ∗ datasetSize	2	68	2.094	0.131
testOn ∗ datasetSize	2	68	5.579	<.01**
trainedOn ∗ testOn ∗ datasetSize	4	68	3.079	<.05*

(b) Planned contrast ID vs OOD.

datasetSize	t (df)	p
overall	24.191 (68)	<.001***
500	15.053 (68)	<.001***
1000	19.158 (68)	<.001***