EMONET-FACE: An Expert-Annotated Benchmark for Synthetic Emotion Recognition

Christoph Schuhmann[1,2*] **Robert Kaczmarczyk**[1,3*] **Gollam Rabby**[2]

Felix Friedrich[4,5] **Maurice Kraus**[4] **Krishna Kalyan**[1] **Kourosh Nadi**[1]

Huu Nguyen[1,6] **Kristian Kersting**[4,5,7] **Sören Auer**[2,8]

[1]LAION e.V., [2]L3S Research Center and Leibniz University of Hannover,
[3]Technical University of Munich, [4]TU Darmstadt, [5]Hessian.AI, [6]Ontocord, [7]DFKI,
[8]TIB, Leibniz Information Centre for Science and Technology

Abstract

Effective human-AI interaction relies on AI's ability to accurately perceive and interpret human emotions. Current benchmarks for vision and vision-language models are severely limited, offering a narrow emotional spectrum that overlooks nuanced states (e.g., bitterness, intoxication) and fails to distinguish subtle differences between related feelings (e.g., shame vs. embarrassment). Existing datasets also often use uncontrolled imagery with occluded faces and lack demographic diversity, risking significant bias. To address these critical gaps, we introduce EMONET-FACE, a comprehensive benchmark suite, featuring: (1) A novel 40-category emotion taxonomy, meticulously derived from foundational research to capture finer details of human emotional experiences. (2) Three large-scale, AI-generated datasets (EMONET-FACE HQ, EMONET-FACE BINARY, and EMONET-FACE BIG) with explicit, full-face expressions and controlled demographic balance across ethnicity, age, and gender. (3) Rigorous, multi-expert annotations for training and evaluation. (4) We build EMPATHICINSIGHT-FACE, models that achieve human expert performance on our held-out benchmark and demonstrate strong real-to-sim generalization. The publicly released EMONET-FACE suite—taxonomy, datasets, and model—provides a robust foundation for developing and evaluating AI systems with a deeper understanding of human emotions.

1 Introduction

The advent of highly capable Large Language Models (LLMs) such as ChatGPT [34] has transformed human-computer interaction, a transformation further accelerated by advances in voice synthesis and multimodal reasoning [35]. Modern voice interfaces now enable emotionally expressive AI conversations [45, 25], leading to heightened expectations for AI systems with genuine emotional intelligence. As AI platforms like Replika [29], Character.ai [9], and Kajiwoto [22] are increasingly used for companionship and support [10], and as LLMs are consulted for mental health advice [48, 23, 16, 47], it becomes ever more critical for these systems to accurately interpret nonverbal emotional cues, especially facial expressions. Misreading users' emotions can erode trust and have negative psychological repercussions [8].

---

*Contributed equally and jointly supervised, correspondance to `contact@laion.ai`

Table 1: Comparison of facial emotion recognition datasets, including synthetic generation and expert annotation. Open license means CC-BY 4.0 or similar.

Dataset Name	Open License	Size	Emotions	Synthetic	Expert Labels	Continuous	Diversity
FERG-DB [2]	✗	55,767	7	✗	✗	✗	✗
FERG-3D-DB [1]	✗	39,574	7	✗	✗	✗	✗
SZU-EmoDage [19]	✓	840	8	✓	(✓)	✗	✗
EmotionNet [17]	✗	N/A	6	✗	✗	✗	N/A
EMOTIC [24]	✗	18,313	26	✗	✗	(✓)	(✓)
FindingEmo [51]	✓	25,869	8	✗	✗	✓	(✓)
AffectNet [33]	✗	450,000	8	✗	✗	✗	(✓)
EMONET-FACE BIG	✓	203,201	40	✓	✗	✗	✓
EMONET-FACE BINARY	✓	19,999	40	✓	✓	✗	✓
EMONET-FACE HQ	✓	2,500	40	✓	✓	✓	✓

(*ours* — EMONET-FACE BIG, EMONET-FACE BINARY, EMONET-FACE HQ)

At the same time, as AI systems become more embodied—through avatars and realistically synthesized faces—another crucial challenge emerges: enabling humans to correctly perceive and interpret the emotional states of these digital agents. For emotionally rich and affective interaction, the expressions generated by machine avatars must be legible and interpretable to human users, ensuring users understand not only the intentionality but also the emotional stance and nuance behind an AI's responses. This reciprocity is critical for building trust, rapport, and effective communication in human-AI partnerships. Truly emotionally aware AI thus requires a bidirectional capability: machines must recognize not only basic emotions [13] but also complex and subtle affective states [36, 12, 39], while synthetic emotions generated by AI avatars must be expressively rich and clearly understood by human counterparts. However, current benchmarks are limited in size, their spectrum of emotions, the diversity of groups included, and the quality of their annotations, constraining progress toward AI systems capable of authentic, context-sensitive empathy and expressiveness [11].

To address these gaps, we introduce EMONET-FACE, an expert-annotated data suite for fine-grained facial emotion recognition (FER) grounded in a novel, expansive 40-category emotion taxonomy. This taxonomy, developed by mining the "Handbook of Emotions"[27] and refined with psychological expertise, captures a broader and more detailed array of human emotional states than previous efforts. Using state-of-the-art text-to-image models [6, 32], we generated a series of structured, demographically balanced, high-quality synthetic datasets, covering a wide range of age, ethnicity, and gender. This includes a pretraining set of over 203,000 images (EMONET-FACE BIG), a fine-tuning dataset (EMONET-FACE BINARY) with almost 20k images and more than 62,000 human expert binary emotion annotations, and an evaluation benchmark (EMONET-FACE HQ) of 2,500 images meticulously rated by psychology experts using continuous scales across all 40 emotion categories. Alongside these datasets, we present EMPATHICINSIGHT-FACE, two models trained on our suite that achieve human-expert-level performance on the EMONET-FACE HQ benchmark and outperform concurrent models.

Our main contributions are as follows: **(i)** We introduce a comprehensive 40-category emotion taxonomy, refined through expert consultation, to capture fine-grained human emotional states; **(ii)** We construct diverse synthetic expert-annotated image datasets—spanning pretraining, fine-tuning, and benchmarking test sets—using state-of-the-art text-to-image models; **(iii)** We develop EMPATHICINSIGHT-FACE models that match human performance on fine-grained emotion recognition; **(iv)** We openly release our datasets and models to foster research on emotions in AI.[2]

2 Related Work

Understanding human emotion remains a complex task, grounded in diverse psychological theories and demanding robust data for computational modeling. Emotions are challenging as they are not universal—they are socially constructed, shaped by cultural norms and context (further elaborated in

---

[2]We openly release our datasets, models, training scripts, and open-source annotation tool at our HuggingFace and GitHub repositories. Generation prompts are included in the dataset metadata.

[3, 4] and App. A.1). This section outlines key theoretical frameworks, examines existing emotion taxonomies and datasets, and identifies the specific gaps that EMONET-FACE addresses.

Existing Emotion Taxonomies. Theories of emotion span a spectrum from universal, biologically rooted models to constructivist perspectives, with the latter emphasizing that emotions are shaped by context and culture rather than being innate, static programs [3, 4]. This conceptual diversity has direct consequences for how emotions are categorized and labeled in practice. Traditional taxonomies provide important foundations but often lack granularity. Ekman's model expanded to include emotions like amusement and shame [14], but remains limited to discrete, universal states. Parrott's hierarchical structure [38] and Plutchik's Wheel [40] offer more coverage but overlook distinctions critical for real-world applications. For example, Plutchik's dyadic model emphasizes emotional opposites and blends but may oversimplify ambiguous affective states. Other approaches include Panksepp's neurobiologically rooted systems (e.g., SEEKING, FEAR, LUST) [37], and Izard's emphasis on distinct expressive and neural signatures for emotions like joy, guilt, and interest [20, 21]. Recent efforts have enriched positive emotion taxonomies—e.g., Shiota et al.'s physiology-based framework [46] and Weidman and Tracy's focus on subjective feeling states [49]. Despite these contributions, existing emotion taxonomies often omit physical states and stances like fatigue, pain, numbness, or sourness, and do not clearly distinguish closely related states such as gratitude, affection, and lust. EMONET-FACE addresses these gaps with a broader, literature-informed taxonomy [27] expanding basic emotions with relevant physical states and stances to a 40-category taxonomy, tailored for high-resolution FER.

Emotion Recognition Datasets. Existing emotion recognition datasets vary significantly in terms of realism, diversity, and labeling accuracy. Early posed datasets (e.g., CK+ [28], JAFFE [30]) rely on actor-performed facial expressions, offering clear labels but limited real-world validity. In-the-wild datasets (e.g., AffectNet [33], EmotioNet [5], EMOTIC [24], FindingEmo [51]) provide more realistic expressions with contextual richness. However, they suffer from confounders like context-dependent perception effects (e.g., Kuleshov effect [7], background assimilation [50]), as well as occlusions and inconsistent image quality. In contrast, EMONET-FACE overcomes these limitations by offering high-resolution, synthetic facial images rendered under controlled conditions. To this end, our dataset offers deliberate and precise control over demographic diversity and background context for every emotion, in contrast to previous datasets, which might achieve some diversity incidentally by web scraping. Additionally, it is unique in offering expert annotations across 40 emotion categories. A detailed comparison with previous datasets is presented in Table 1.

3 Building EMONET-FACE: Fine-grained Expert Emotion Datasets

Having identified several critical gaps in existing emotion datasets—including limited taxonomic granularity, insufficient demographic and contextual diversity, image quality, and a lack of expert annotations—we turn next to how we constructed EMONET-FACE to address these challenges.

Introducing a Fine-grained Taxonomy. To enable more fine-grained affective understanding in AI, we developed a 40-category emotion taxonomy grounded in contemporary psychology and informed by the Theory of Constructed Emotion (TCE) [4]. Unlike prior work limited to basic emotions such as joy, anger, fear, sadness, and disgust, our taxonomy intentionally extends to a broad spectrum of fine-grained social, cognitive, and bodily states—including elation, contentment, gratitude, hope, pride, interest, awe, astonishment, relief, longing, teasing, embarrassment, shame, disappointment, contempt, sexual lust, doubt, confusion, envy, jealousy, bitterness, pain, fatigue, numbness, helplessness, and less common categories such as sourness and intoxication. Each category encompasses a cluster of semantically associated descriptive words, systematically extracted from the *Handbook of Emotions* [27] and refined via expert consultation (see App. Table 5 for full taxonomy).

To construct this taxonomy, we digitized the 946-page *Handbook of Emotions* via OCR, segmented the text into 500-word blocks, and used GPT-4 to extract candidate emotion nouns. After aggregation and deduplication, we identified 170 unique terms, which were then clustered through iterative rounds of independent listing, critical review, and expert-guided refinement with psychologists and researchers. Aligned with TCE, we do not claim biological universality; instead, our taxonomy is designed for context-aware, socially informed emotion interpretation in AI. Given the inherent ambiguity in facial emotion perception—e.g., high-arousal expression might plausibly reflect amusement, elation, or

(a) **Prompt:** An authentic, realistic closeup image of a Hispanic or Latino woman of 70 years who seems to experience triumph, superiority. Strong facial expression of triumph, superiority, [...]
(Midjourney v6)

(b) **Prompt:** A focused close-up photo of a Black or African American man of 30 years who seems to genuinely experience lighthearted fun, amusement, mirth, joviality, laughter, playfulness, silliness, [...]
(Flux Pro)
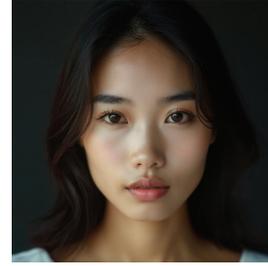
(c) **Prompt:** A closeup sharp, focused photo of a Southeast Asian woman of 40 years who seems to genuinely experience mild yearning, longing, pining, wistfulness, nostalgia, Craving, desire, Envy, [...]
(Flux Dev)

Figure 1: Samples from our EMONET-FACE datasets generated with different sota T2I models.

excitement—we designed the taxonomy for plausible multi-label annotations over rigid single-label assignments, supporting richer, context-aware emotion representations.

Image Generation and Selection Methodology. To construct the EMONET-FACE datasets, we generated synthetic facial images representing all 40 emotion categories using a controlled prompt-engineering pipeline. Each image was manually screened for quality, ensuring visual consistency and rejecting those with artifacts or ambiguous expressions. Critically, prompt templates were explicitly designed to ensure demographic diversity by balancing gender identity (45% man, 45% woman, 10% non-binary), capturing a wide age range (20–80 years, in 10-year increments), and representing 14 different ethnic backgrounds (e.g., "Middle Eastern," "South Asian"). Exemplary generated images are depicted in Figure 1. In total, we generated more than 200k images. While labor-intensive, this manual expert curation was essential to create a gold-standard benchmark, as automated filtering cannot reliably detect the subtle artifacts or ambiguous expressions that would compromise the integrity of a dataset for fine-grained emotion recognition.

Acquiring Expert Annotations. To ensure precise high-quality emotion labels in EMONET-FACE, we engaged psychology experts recruited via Upwork, selected for their verified academic degrees. Our team of annotators (13 in total), aged 18–44 and spanning diverse ethnic and linguistic backgrounds—e.g., Hispanic, European, South Asian, and Middle Eastern—with fluency in up to four languages, brought a diverse perspective. Annotation was conducted on our open-source platform (see App. C). Annotators were instructed to follow our detailed guidelines, rather than relying solely on personal intuition, following [42, 43], promoting coherence. Details on annotator demographics, recruitment, compensation, and adherence to ethical guidelines are provided in App. A.2 and checklist.

For EMONET-FACE HQ, we used a continuous emotion rating protocol: each expert independently assessed batches of 250 images drawn from the 2,500-image collection, scoring all 40 emotion categories for every image on a 0–7 intensity scale. This multi-rater approach, with four ratings per image, supported fine-grained, multi-label assignments for the ambiguous and overlapping expressions commonly found in facial emotion data. To construct EMONET-FACE BINARY, we implemented a rigorous multi-stage binary annotation protocol. In the affirmative sequence, images initially identified as positive for a target emotion by one annotator were reviewed by a second, and if confirmed again, by a third, resulting in emotion labels with triple positive consensus. Images with any annotator dissent were excluded to ensure high-confidence positives. Additionally, a contrastive batch was included, in which annotators reviewed images with the target emotion deliberately absent, ensuring the inclusion of high-quality true negatives (see Table 2). This combination provides robust verification of both the presence and absence of emotion. For EMONET-FACE BIG, we synthetically annotated over 200k images in a two-stage process with Gemini-2.5-Flash. Initially, the model identified and scored the five most salient emotional dimensions per image; subsequently, to boost detection of underrepresented emotions, we used a "hinting" strategy that prompted the model to arbitrarily consider specific emotions. This automated, iterative process continued until we achieved broad coverage across all 40 categories in our taxonomy.

Table 2: Annotation protocol and agreement rates for EMONET-FACE BINARY. "Affirmative" batches required triple positive agreement across annotators, while the "contrastive" batch ensured true negatives. Agreement indicates the proportion of images for which annotators consistently agreed on presence (affirmative) or absence (contrastive) of the target emotion, which is high overall.

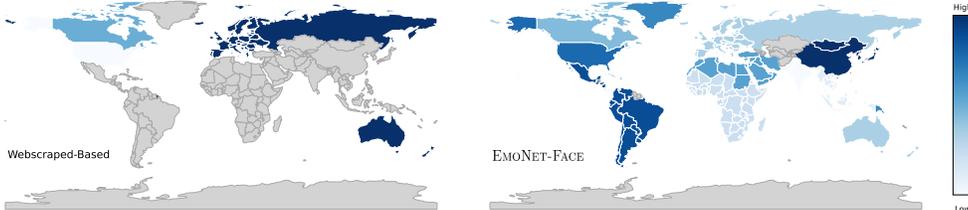| Batch	# Images	Agreement	# Annotators	Step
Batch 1 (affirmative)	19,999 (100.00%)	82.83%	3	Starting point
Batch 2 (affirmative)	14,670 (73.35%)	75.11%	4	Batch 1 positives
Batch 3 (affirmative)	11,018 (55.09%)	82.37%	4	Batch 2 positives
Batch 4 (contrastive)	19,999 (100.00%)	94.29%	4	Batch 1 contrastive

Figure 2: Approximate world map of demographic coverage and diversity in web-scraped datasets (left) compared to EMONET-FACE (right), which is much more diverse.

Inter-Annotator Agreement.   To quantify annotator consistency for EMONET-FACE, we evaluated agreement using both pairwise and group metrics for each emotion. For the continuous 0–7 scale in EMONET-FACE HQ, we used weighted kappa (κ_w, with quadratic weights) to capture how closely pairs of annotators rated images, considering only cases where both provided ratings. We also computed Krippendorff's α across all annotators for both EMONET-FACE HQ and EMONET-FACE BINARY. We also compute the conditional agreement for EMONET-FACE BINARY, following the sequential labeling, as described above.

The average pairwise κ_w between human annotators for EMONET-FACE HQ was 0.20 (see Figure 3, A, top part), indicating moderate agreement. Furthermore, for EMONET-FACE HQ, we depict Krippendorff's α per emotion in Figure 9. Across all 40 emotions, the mean α is 0.19 (95% CI [0.16, 0.22]), indicating moderate overall agreement. Per-emotion α varies widely: highest for Elation ($\alpha = 0.58$), Amusement (0.56) and Anger (0.46); lowest (even negative) for Interest (-0.08), Concentration (-0.02) and Contemplation (-0.02). This spread suggests that stimulus ambiguity, not annotator error, drives most disagreement.

For EMONET-FACE BINARY, we get overall $\alpha = 0.09$ (95% CI [0.09, 0.09]), as depicted in Figure 8, with the top five by binary α are Fatigue/Exhaustion ($\alpha = 0.49$), Pain (0.45), Malevolence (0.44), Teasing (0.43) and Intoxication (0.40). For conditional binary (presence/absence) agreement, consensus rates are reported in Table 2. High agreement was achieved across batches, with especially strong consensus in the contrastive (true-negative) setting, reflecting the comparative ease of judging when an emotion is absent, compared to present.

Taken together, these results demonstrate that while human agreement is strong for many emotions, some categories naturally elicit a wider range of interpretations, underscoring the nuanced nature of facial emotion expressions. Rather than indicating weak annotation quality, this pattern highlights the sensitivity of EMONET-FACE to the inherent complexity of affective perception. Our annotations reflect both the challenges and scientific opportunities of capturing emotional diversity.

Data Integrity, Safety, and Fairness.   All generated images were manually reviewed for obvious stereotypes, artifacts, or harmful content. This way, one inappropriate image was removed from EMONET-FACE BINARY. The EMONET-FACE dataset is released for research under ethical terms of use, supporting safe and fair study of affective AI. Moreover, we focused on a broad and diverse representation across identities. The development of fair and unbiased FER systems critically depends on diverse training and evaluation datasets. While some existing datasets (see Table 1) include varied identities, equitable coverage across groups and emotional categories remains limited to date. In
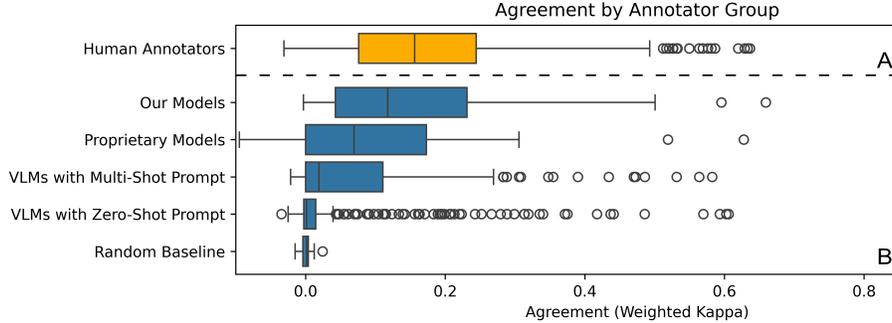
Figure 3: Weighted Kappa (κ_w) agreement scores by annotator group. **A** (top): Pairwise agreement between human annotators. **B** (below): Pairwise agreement between each human annotation and other sources, including 'Our Models' (EMPATHICINSIGHT-FACE), 'Proprietary Models' (HumeFace), 'VLMs (Multi-Shot and Zero-Shot Prompts)', and a 'Random Baseline'. Each box represents the interquartile range (IQR) of κ_w scores, with the median as center line.

creating our EMONET-FACE datasets, we explicitly addressed this by focusing on diversity across ethnicity, gender, and age groups. Figure 2 illustrates the approximate demographic diversity between average web-scraped datasets, predominantly White/Caucasian [31], and our datasets, which are much less Western-centric and more diversified across identities. For additional visualizations, refer to the supplementary files.

3.1 The EMONET-FACE Suite

After having established and evaluated its foundation, we release EMONET-FACE as three subsets:

(1) **EMONET-FACE HQ**: a *benchmark test dataset* with 2,500 images with multiple continuous expert annotators per image, resulting in 10,000 human expert annotations;

(2) **EMONET-FACE BINARY**: a *fine-tuning dataset* with 19,999 images with multiple binary expert annotators per image from a multi-stage process, resulting in 65,686 human expert annotations;

(3) **EMONET-FACE BIG**: a *pre-training dataset* with 203,201 images and generated labels.

4 Benchmarking Facial Emotion Recognition on EMONET-FACE

We now use EMONET-FACE to benchmark FER models, including our baselines, proprietary systems, and VLMs, against human expert annotations.

4.1 Experimental Setup

Developing Baseline Models: EMPATHICINSIGHT-FACE. In addition to establishing new FER datasets, we leverage them to train baseline models, which we term EMPATHICINSIGHT-FACE. Our architectural design is deliberately simple—a SIGLIP2 vision backbone followed by lightweight MLP regression heads—to demonstrate that strong performance can be achieved with a straightforward architecture when paired with high-quality data and robust embeddings. We chose continuous regression (predicting scores on a 0–7 scale) over classification for two key reasons: (1) *Capturing Intensity:* Regression naturally models the gradations of emotional expression, distinguishing mild amusement from intense elation, a nuance lost in discrete classification. (2) *Multi-Label Representation:* Faces often display blends of emotions. Our regression approach allows for predicting multiple emotions simultaneously (e.g., high sadness and moderate disappointment), providing a richer and more realistic affective profile. We pre-train on EMONET-FACE BIG and fine-tune on EMONET-FACE BINARY. Two model sizes are trained—EMPATHICINSIGHT-FACE LARGE (40 x 1.8M-parameter heads) and EMPATHICINSIGHT-FACE SMALL (40 x 151k-parameter heads). Full details in App. B.2.

State-of-the-Art Models. To benchmark SOTA FER, we evaluated several open- and closed-source VLMs—including Gemini (various versions), Claude 3.7 Sonnet, GPT-4o, Nova Pro, Pixtral Large,
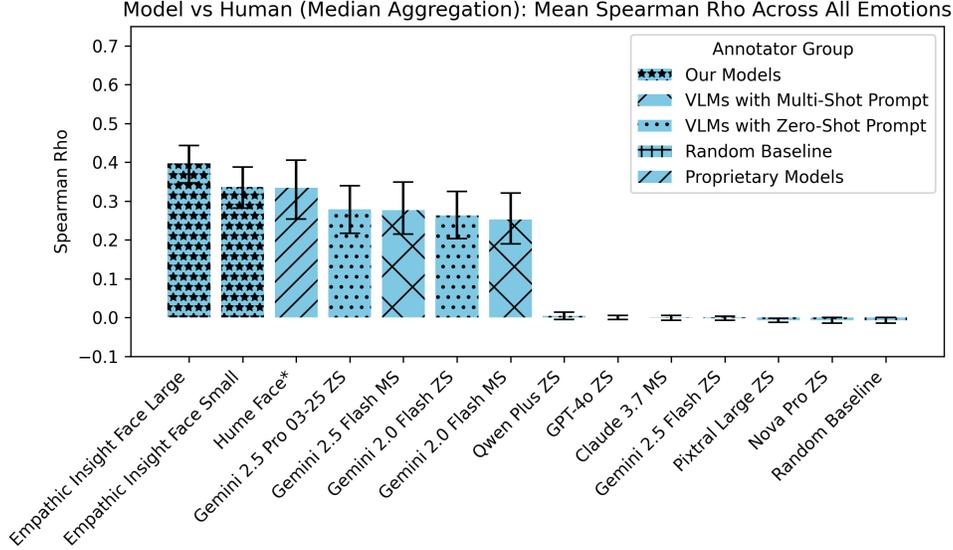
Figure 4: Mean Spearman's ρ correlation between various model annotators and human annotations. Human ratings were median-aggregated per emotion before correlation with model ratings. The bar heights represent the mean of these per-emotion Spearman's ρ values, calculated across all emotions for each model. Error bars indicate bootstrap 95% confidence intervals (N=1000 bootstraps) for these means. Model annotator groups, including our trained models (EMPATHICINSIGHT-FACE), VLMs with multi-shot or zero-shot prompting, proprietary models (HumeFace), and a random baseline, are distinguished by patterns as detailed in the legend

and Qwen Plus—on zero-shot and multi-shot prompt formats, and the proprietary HumeFace API. Model outputs were considered successful if they returned correctly formatted and parseable emotion labels; failures included refusals, empty responses, or unparseable outputs as reported in Table 3. The high failure rate of many SOTA VLMs is a significant finding of our benchmark, highlighting their brittleness for this task. For our performance analysis, any such failure was treated as a zero-agreement score, reflecting a complete inability to perform the task.

4.2 Do They See What We See?

We now examine how closely models—both specialized and general-purpose—recognize the same emotions as humans in synthetic faces, and how strongly their judgments align. To provide a clear and concise comparison across annotator groups, Figure 3 aggregates pairwise agreement scores (Weighted Kappa, κ_w) by model type and prompt format. This aggregation highlights several important trends: our models (EMPATHICINSIGHT-FACE) achieve consistently high agreement with human annotators; proprietary models (HumeFace) show only slight agreement; and, as expected, random baselines display negligible agreement. Most notably, VLM performance is highly variable and generally inconsistent—regardless of prompt format, VLMs on average do not offer a dependable improvement over random guessing. As illustrated in Table 3, this inconsistency is compounded by high rates of unparseable outputs or outright refusals, further hindering their reliability.

Strikingly, further statistical analyses (Pairwise Mann–Whitney U, bootstrapped 95% confidence intervals) show that EMPATHICINSIGHT-FACE models are statistically indistinguishable from human annotators ($\Delta = 0.019$, $p = 0.103$), while significantly ($p < 0.001$) outperforming proprietary models, both multi-shot and zero-shot VLMs, and random baselines. These findings underscore that with focused dataset construction and careful fine-tuning, state-of-the-art models can genuinely approach human-level reliability on synthetic FER tasks. By contrast, general-purpose VLMs continue to lag behind, particularly when faced with nuanced or ambiguous facial expressions, highlighting persistent limitations despite recent advances.

To provide a more granular perspective, Figure 4 disaggregates results by individual model and prompt setup, offering insight into the underlying variability. Here, Spearman's ρ measures rank

7

Table 3: Support for zero-shot and multi-shot emotion-analysis prompts (format only). *Gemini 2.5-pro-03-25 was discontinued after evaluating 1847 of 2500 images from EMONET-FACE HQ. ✓: successful parsing; ✗: refusal or unparseable output (e.g., mismatched labels)

Model name	Zero-shot prompt	Multi-shot prompt	Comment
Gemini 2.5 Pro 03-25	✗	✓*	Empty Response
Gemini 2.5 Flash	✓	✓	-
Gemini 2.0 Flash	✓	✓	-
Claude 3.7 Sonnet	✗	✓	Empty Response
GPT-4o	✓	✗	Rejection
Nova Pro	✓	✗	Non-parsable Output
Pixtral Large	✓	✗	Non-parsable Output
Qwen Plus	✓	✗	Non-parsable Output

correlation between model and human emotion ratings across all 40 categories. Our tailored models (EMPATHICINSIGHT-FACE) again match or exceed the performance of proprietary systems and VLMs, displaying consistency across runs. While a few VLMs (such as Gemini) achieve moderate agreement, many others fall well short, and there is no clear pattern favoring either zero-shot or multi-shot setups. This heterogeneity reinforces the "hit-or-miss" nature of current VLM annotators: depending on the model and prompt configuration, performance may range from reasonable to essentially random. As further illustrated in Table 3, prompt handling difficulties, model-internal safety constraints, and sensitivity to format further hinder VLM robustness—only two of nine VLMs consistently produce outputs under both prompt types. These results are also backed by direct comparison in App. Figure 6, which visually details the pairwise agreement landscape at the annotator level. This unpredictability highlights the need for careful selection and calibration when applying VLMs to FER tasks.

In summary, systematic evaluation of human, proprietary, and VLM-based annotators demonstrates that our new datasets and models reliably enable near-human performance, while exposing the current limitations of general-purpose AI models. EMONET-FACE thus establishes a rigorous new benchmark for progress in FER.

4.3 Generalization to Real-World Datasets

A key question for any benchmark based on synthetic data is how well models trained on it generalize to real-world, in-the-wild images. To assess this sim-to-real transfer, we evaluated our EMPATHICINSIGHT-FACE LARGE model on two established real-world datasets: the Facial Emotion Recognition Dataset (FERD; n=152) [41] and AffectNet (n=31,002) [33].

Before evaluation, a primary task is bridging the gap between our model's 40-category continuous output and the single discrete labels of the target datasets. To this end, we mapped our 40 labels to their 8-category taxonomies (details in App. D). After this mapping, we applied a softmax function across our model's 40 outputs to obtain a probability distribution. We then considered one fo the 8 basic emotion to be *present* if the probability of *any* of its constituent fine-grained emotions exceeded a threshold of $1.5 \times (1/40)$, which is 50% higher than a uniform random guess. This predicted label was then compared against the ground-truth label from the respective dataset to calculate accuracy.

As shown in Table 4, our model demonstrates strong generalization, achieving a mean accuracy of 78.29% on FERD and 75.72% on AffectNet. Performance is particularly high for categories like *Happy* and *Surprise*, where conceptual overlap in the category mapping is strong. Even with the inherent challenges of mapping between different emotion taxonomies, these results indicate that training on the controlled, diverse, and high-quality images of EMONET-FACE provides a robust foundation for real-world FER. This suggests that synthetic data can effectively bridge gaps in existing datasets, enabling the development of more capable and generalizable models.

5 Discussion

Challenges and Chances in FER. FER remains an intrinsically complex and nuanced endeavor, rooted in fundamental questions about the nature of emotion itself. Decades of affective science, including the Theory of Constructed Emotion (TCE), emphasize that emotions are not universally

Table 4: Performance of our EMPATHICINSIGHT-FACE LARGE model on real-world emotion datasets (FERD and AffectNet). We report accuracy (%) after mapping our 40-category output to their 8 basic emotion categories using our threshold-based prediction strategy.

	anger	contempt	disgust	fear	happy	neutral	sad	surprise	avg.
FERD	73.68	31.58	78.95	100.00	100.00	84.21	78.95	78.95	**78.29**
AffectNet	77.05	28.75	40.53	69.08	99.25	78.96	83.94	98.70	**75.72**

"read out" from facial muscle configurations, but rather interpreted by observers through a dynamic interplay of visual cues, individual experience, and situational context (see App. A.1). Our analysis of inter-annotator agreement (Section 3) underscores this reality: even among domain experts, there is marked variability in how expressions are labeled, particularly for subtle categories. This diversity should not be seen as noise, but as a core finding. It reflects the genuine psychological complexity of emotion perception and demonstrates that EMONET-FACE captures this ambiguity. A benchmark with perfect agreement would be unrealistic and likely less robust. Our findings motivate a paradigm shift away from seeking a single authoritative label, toward approaches that estimate distributions over plausible emotions, a challenge for which EMONET-FACE provides a foundational testbed.

This subjectivity and inter-annotator diversity, should not be seen solely as sources of noise or error; rather, they highlight the deeply constructive and situationally-contingent character of affective perception. Our findings motivate a paradigm shift away from only seeking a single authoritative label for facial images, toward approaches that estimate distributions over plausible emotion categories (see Figure 5) or defer to LLMs and VLMs capable of subsequent context-sensitive reasoning.

In response to these challenges, our study introduces EMONET-FACE —a comprehensive suite designed to systematically benchmark both human and model performance. By providing a large-scale, diverse, and fine-grained test bed, EMONET-FACE enables robust examination of inter-annotator variability, the limits of facial expression alone for conveying emotion, and the reliability of both domain-specific and general-purpose AI in this domain. We hope this resource lays the groundwork for future research that not only seeks technical progress, but also more faithfully engages with the rich psychological and contextual complexity of human affect.

Specialized Models Can Learn to See What Humans See. Despite the inherent challenges of FER, our EMPATHICINSIGHT-FACE models exhibit consistently high agreement with human annotators. This success is partly attributable to our use of SigLIP2, a powerful vision backbone that provides strong image embeddings highly suited for affective analysis. As shown in Table 7 and Figure 6, our EMPATHICINSIGHT-FACE LARGE model achieves human agreement scores significantly above state-of-the-art zero-shot VLMs, proprietary systems, and even several individual expert annotators. This level of agreement is significant, given our annotators' strong domain expertise, demonstrating that targeted model design and careful data curation can substantially close the gap between machine and human performance on this task. Yet it is important to recognize that such agreement does not necessarily equate to genuine emotional understanding. Current models, including our own, may accurately mimic or imitate human emotion recognition on a surface level by leveraging statistical regularities in the data, without actually grasping the underlying emotional states or meaning. True comprehension of emotion likely requires not only pattern matching in facial cues, but also deeper insight into the context, causes, and subjective experience behind each expression.

General-Purpose VLMs Hold Untapped Potential. Our broader benchmarking of general-purpose VLMs reveals substantial inconsistencies across models and prompt setups: while a few VLMs, such as Gemini, achieve moderate agreement for certain settings, their performance is generally unreliable and highly sensitive to prompt formatting (see Figures 3 and 4). Crucially, there is no current prompt or model configuration that consistently enables VLMs to excel at nuanced FER tasks. Furthermore, we observe a pronounced misalignment problem: different VLMs (or even the same VLM under different prompts) frequently produce divergent or incoherent predictions for the same input image. This lack of alignment and robustness presents a serious challenge for deploying these systems in interactive, real-world settings, as human users require dependable and interpretable model behavior for meaningful interaction [15, 18]. Robustness and alignment—consistency within a model and agreement across models—are therefore just as important as absolute performance when

building trustworthy and user-friendly affective AI. While the foundational capabilities of VLMs are rapidly improving and may eventually surpass specialized approaches like our EMPATHICINSIGHT-FACE, realizing this potential will require addressing both their alignment and robustness shortcomings through new techniques, such as improved prompting strategies, architectural innovations, or alignment-focused training protocols. Future work bridging these gaps—potentially by combining the flexibility of VLMs with task-specific supervision—could unlock even greater advances, and enable robust, reliable, and context-aware emotion understanding at scale.

Limitations. Despite these advances, several limitations should be considered. *Sim-to-Real Generalization:* While our new experiments (Section 4.3) show promising generalization to real-world images, the sim-to-real gap remains an important area for future research. Validating performance across a wider range of naturalistic datasets is crucial. *Static vs. Dynamic Cues:* Our work is restricted to static facial expressions. This omits temporal dynamics, such as the evolution of an expression or microexpressions, which are vital for nuanced emotion interpretation. EmoNet-Face provides a foundation, but future work should extend this to dynamic video-based benchmarks. *Cross-Cultural Scope:* Our 40-category taxonomy, while comprehensive, is grounded in Western-centric psychological literature. Its universality across different cultures is an open question. Although we controlled for diverse ethnicities in image generation, dedicated cross-cultural validation is a critical next step. *Context-Free Analysis:* By focusing on faces in isolation, we remove contextual confounders but also limit the scope of analysis. As highlighted by TCE, context is integral to emotion perception. Integrating situational and multimodal cues (e.g., speech, body language) is essential for the next generation of emotionally intelligent AI.

In summary, the EMONET-FACE suite establishes a rigorous and versatile new testbed for emotion recognition research. Our results demonstrate that focused, expert-informed modeling can yield near-human—and in some cases, super-human—consistency on challenging FER tasks. Addressing the outlined limitations, particularly by embracing richer contextual and multimodal information, will be key to the next generation of emotionally intelligent AI systems.

Ethical Considerations, Licensing, and Use. This work addresses concerns about uncalibrated AI by providing a benchmark for safer, more nuanced emotion recognition. All images in EMONET-FACE are synthetic, mitigating privacy risks associated with human data. We acknowledge the risk of misuse (e.g., for manipulation) and release our work to support transparent, safety-oriented research.

To ensure clarity for downstream users, we specify our licensing. The EMONET-FACE datasets and EMPATHICINSIGHT-FACE models are released under the CC BY 4.0 license. This is permissible because the source T2I models grant us the necessary rights. For images generated with a paid Midjourneysubscription, the terms of service grant full ownership of the assets to the creator. For images generated with FLUX.1[dev], the license explicitly states that the creators claim no ownership of the outputs and they may be used for any purpose, including commercial. However, this permissive license does not override legal and regulatory obligations. Downstream applications must comply with local laws, such as the EU AI Act, which prohibits emotion inference in sensitive contexts like workplaces and educational institutions (Article 5(1)(f)). EMONET-FACE is not intended for such uses; its application should be confined to legally permitted and ethically sound research contexts.

6 Conclusion

We introduced the EMONET-FACE suite—a comprehensive resource featuring a fine-grained 40-category taxonomy, richly annotated datasets, and specialized models that establish new state-of-the-art benchmarks for FER. The EMONET-FACE benchmark exposes significant limitations in current tools and models for emotion evaluation, underscoring both the complexity of the task and the need for future innovation. Leveraging this suite, we developed the EMPATHICINSIGHT-FACE models, which deliver leading performance on facial emotion recognition while also illuminating the substantial variability in human judgments and the inherent constraints of using facial expressions alone. We provide a thorough discussion of these challenges and outline promising directions for advancing human-like emotion understanding in AI—impacting not only computer science, but also contributing valuable insights to psychology and the social sciences.

7   Acknowledgements

References

[1] Deepali Aneja, Bhaswar Chaudhuri, Alex Colburn, Gary Faigin, Linda Shapiro, and Barbara Mones. Learning to generate 3d stylized character expressions from humans. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 160–169. IEEE, 2018.

[2] Deepali Aneja, Alex Colburn, Gary Faigin, Linda Shapiro, and Barbara Mones. Modeling stylized character expressions via deep learning. In *Asian Conference on Computer Vision*, pages 136–153. Springer, 2016.

[3] Lisa Feldman Barrett. Are emotions natural kinds? *Perspectives on Psychological Science*, 1(1):28–58, 2006.

[4] Lisa Feldman Barrett. *How Emotions Are Made: The Secret Life of the Brain*. Houghton Mifflin Harcourt, 2017.

[5] C. Fabian Benitez-Quiroz, Ramprakash Srinivasan, and Aleix M. Martinez. Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[6] Black Forest Labs. Flux.1-dev. `https://huggingface.co/black-forest-labs/FLUX.1-dev`, 2024. Accessed: 12-05-2025.

[7] Marta Calbi, Katrin Heimann, Daniel Barratt, Francesca Siri, Maria A. Umiltà, and Vittorio Gallese. How context influences our perception of emotional faces: A behavioral study on the kuleshov effect. *Frontiers in Psychology*, 10:611, 2019.

[8] C. Daryl Cameron and Michael Inzlicht. Empathy is hard work: People choose to avoid empathy because of its cognitive costs. *Journal of Experimental Psychology: General*, 2019.

[9] Inc. Character Technologies. Character.ai: Conversational ai agents. `https://character.ai`, 2023. Accessed: 2024.

[10] Rijul Chaturvedi, Sanjeev Verma, Ronnie Das, and Yogesh K. Dwivedi. Social companionship with artificial intelligence: Recent trends and future avenues. *Technological Forecasting and Social Change*, 2023.

[11] Alan S. Cowen and Dacher Keltner. Self-report captures 27 distinct categories of emotion bridged by continuous gradients. *Proceedings of the National Academy of Sciences*, 114(38):E7900–E7909, 2017.

[12] Richard J. Davidson. Affective neuroscience and psychophysiology: Toward a synthesis. *Psychophysiology*, 40(5):655–665, 2003.

[13] Paul Ekman. An argument for basic emotions. *Cognition and Emotion*, 1992.

[14] Paul Ekman. Basic emotions. In *Handbook of Cognition and Emotion*. American Psychologic Association, 1999.

[15] Severin Engelmann, Chiara Ullstein, Orestis Papakyriakopoulos, and Jens Grossklags. What people think ai should infer from faces. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1–14. ACM, 2022.

[16] Aaron Gilson, Chris Safranek, Thomas Huang, V Socrates, Joyce Chi, Richard Taylor, and James Yuan. How well does chatgpt perform on the united states medical licensing examination? the implications of large language models for medical education and knowledge assessment. *PLOS Digital Health*, 2023.

[17] Ralph Gross, Iain Matthews, Jeffrey Cohn, Takeo Kanade, and Simon Baker. Multi-pie. *Image and Vision Computing*, 28(5):807–813, 2010.

[18] Kimia Hamidieh, Haoran Zhang, Walter Gerych, Thomas Hartvigsen, and Marzyeh Ghassemi. Identifying implicit social biases in vision-language models. In *Proceedings of the 2024 AAAI/ACM Conference on AI, Ethics, and Society*, 2024.

[19] Shangfeng Han, Yanliang Guo, Xinyi Zhou, Junlong Huang, Linlin Shen, and Yuejia Luo. A chinese face dataset with dynamic expressions and diverse ages synthesized by deep learning. *Sci. Data*, 10(1):878, December 2023.

[20] Carroll E. Izard. *The Face of Emotion*. Appleton-Century-Crofts, 1971.

[21] Carroll E. Izard. *The Psychology of Emotions*. Springer, 1991.

[22] Kajiwoto. Kajiwoto: Create custom ai companions. `https://www.kajiwoto.com`, 2022. Accessed: 2025-05-13.

[23] Khondoker Tashya Kalam, Jannatul Mabia Rahman, Md Rabiul Islam, and Syed Masudur Rahman Dewan. Chatgpt and mental health: Friends or foes? *Health Science Reports*, 2024.

[24] Ronak Kosti, Jose M. Alvarez, Adria Recasens, and Agata Lapedriza. Emotic: Emotions in context dataset. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2017.

[25] Canopy Labs. Orpheus: Emotionally expressive speech ai. `https://canopylabs.ai/orpheus`, 2024. Accessed: 2024.

[26] Richard S. Lazarus. *Emotion and Adaptation*. Oxford University Press, 1991.

[27] M. Lewis, J. M. Haviland-Jones, and L. F. Barrett. *Handbook of Emotions*. Guilford Press, 4th edition, 2016.

[28] Patrick Lucey, Jeffrey F. Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 94–101, 2010.

[29] Inc. Luka. Replika: Ai companion. `https://replika.ai`, 2023. Accessed: 2024.

[30] Michael Lyons, Shigeru Akamatsu, Miyuki Kamachi, and Jiro Gyoba. Coding facial expressions with gabor wavelets. In *Proceedings of the Third IEEE International Conference on Automatic Face and Gesture Recognition*, 1998.

[31] Abhishek Mandal, Susan Leavy, and Suzanne Little. Dataset diversity: Measuring and mitigating geographical bias in image search and retrieval. In *Proceedings of the 1st International Workshop on Trustworthy AI for Multimedia Computing*, pages 19–25, 2021.

[32] Inc. Midjourney. Midjourney v6: Ai image generation. `https://midjourney.com`, 2024. Accessed: 2024.

[33] Ali Mollahosseini, Behzad Hasani, and Mohammad H. Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1):18–31, 2017.

[34] OpenAI. Chatgpt. `https://openai.com/chatgpt`, 2023. Accessed: 2024.

[35] OpenAI. Gpt-4o: Advanced multimodal ai. `https://openai.com/gpt-4o`, May 2024. Accessed: 2024.

[36] Andrew Ortony, Gerald L. Clore, and Allan Collins. *The cognitive structure of emotions*. Cambridge University Press, 1990.

[37] Jaak Panksepp. *Affective Neuroscience: The Foundations of Human and Animal Emotions*. Oxford University Press, 1998.

[38] W. Gerrod Parrott. *Emotions in Social Psychology*. Psychology Press, 2001.

[39] Rosalind W. Picard. *Affective Computing*. MIT Press, 1997.

[40] Robert Plutchik. *A General Psychoevolutionary Theory of Emotion*. Academic Press, 1980.

[41] Kucev Roman. Facial emotion recognition dataset, August 2023.

[42] Paul Röttger, Bertie Vidgen, Dirk Hovy, and Janet Pierrehumbert. Two contrasting data annotation paradigms for subjective nlp tasks. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2022.

[43] Paul Röttger, Giuseppe Attanasio, Felix Friedrich, Janis Goldzycher, Alicia Parrish, Rishabh Bhardwaj, Chiara Di Bonaventura, Roman Eng, Gaia El Khoury Geagea, Sujata Goswami, Jieun Han, Dirk Hovy, Seogyeong Jeong, Paloma Jeretič, Flor Miriam Plaza del Arco, Donya Rooein, Patrick Schramowski, Anastassia Shaitarova, Xudong Shen, Richard Willats, Andrea Zugarini, and Bertie Vidgen. Msts: A multimodal safety test suite for vision-language models, 2025.

[44] Stanley Schachter and Jerome E. Singer. Cognitive, social, and physiological determinants of emotional state. *Psychological Review*, 69(5):379–399, 1962.

[45] Sesame. Sesame: Conversational speech model. `https://sesame.com`, 2024. Accessed: 2024.

[46] Michelle N. Shiota, Samantha L. Neufeld, Alexander F. Danvers, Elizabeth A. Osborne, Oliver Sng, and Cindy I. Yee. Positive emotion differentiation: A functional approach. *Social and Personality Psychology Compass*, 2017.

[47] Michel Valstar, Björn Schuller, Nicholas Cummins, and Maja Pantic. Ai for mental health: A systematic review of affect recognition techniques. *IEEE Transactions on Affective Computing*, 14(1):3–20, 2023.

[48] Yaohui Wei, Lei Guo, Cheng Lian, and Jue Chen. Chatgpt: Opportunities, risks and priorities for psychiatry. *Asian Journal of Psychiatry*, 2023.

[49] Aaron C. Weidman and Jessica L. Tracy. Picking up good vibrations: Discrete positive emotions and their experiential content. *Emotion*, 2020.

[50] Ying Wu, Wen Chen, Hong Li, and Xiaolan Fu. The influence of background on facial emotion perception: A psychophysical study. *i-Perception*, 14(2), 2023.

[51] Jialin Yang, Ru Li, Shuai Xiao, Wenming Li, Wei-Long Zheng, and Bao-Liang Lu. Findingemo: An image dataset for emotion recognition in the wild. *IEEE Transactions on Multimedia*, 23:2033–2045, 2021.

A Appendices

A.1 Theories of Emotion: From Universal Programs to Constructed Experiences.

Emotion theories span a spectrum from biologically grounded to constructivist accounts. Basic Emotion Theory (BET) posits a universal set of evolved emotions—anger, disgust, fear, happiness, sadness, and surprise—each with distinct physiological and facial expressions [13], though it has been critiqued for oversimplifying emotion and overlooking cultural variability [3]. Cognitive appraisal theories emphasize interpretation, with the Schachter-Singer model combining physiological arousal with contextual labeling [44], and Lazarus's framework highlighting evaluations of relevance and coping potential as key drivers of emotional response [26]. The Theory of Constructed Emotion (TCE) offers a constructivist view, arguing that emotions are not innate but are constructed through predictive brain mechanisms using culturally learned concepts [4]. TCE holds that emotional experiences emerge from interpreting interoceptive signals (valence and arousal) via culturally shaped categories, rejecting the idea of fixed "emotion circuits" and emphasizing variability across instances. This predictive, concept-driven view has major implications for AI, calling for models that interpret affect in context-sensitive, culturally informed ways rather than relying on static facial-emotion mappings. This is further visualized in Figure 5.

A.2 Detailed Expert Annotator Recruitment and Process

To ensure the highest quality and precision of emotion labels in EMONET-FACE, we relied on expert annotation by psychology professionals, recruited through Upwork and selected for their verified academic degrees or substantial experience in the field. Our broad team of annotators, aged 18–44 and spanning diverse ethnic and linguistic backgrounds—e.g., Hispanic, European, South Asian, and Middle Eastern—with fluency in up to four languages, brought an inclusive perspective to the annotation process. All annotators (13 in total) participated voluntarily after giving informed consent and were fairly compensated at rates well above local standards, in keeping with the NeurIPS Code of Ethics. Annotators were instructed to follow our detailed, standardized guidelines, rather than relying solely on personal intuition, following [42, 43], promoting coherence and reproducibility. Open communication allowed for regular clarification of any ambiguities during annotation.

A.3 EMONET-FACE Data Preprocessing

The preprocessing pipeline included: (1) Parsing Ratings—extracting structured emotion scores; (2) Emotion Extraction—identifying all unique emotion labels to define the target space; (3) Long Format Transformation—reshaping the dataset into (image, annotator, emotion) triplets, with missing ratings filled as 0 to reflect the annotation platform's assumption that unrated emotions indicate absence; and (4) Annotator Typing—labeling annotators as Human or VLM based on a binary flag, enabling separate analyses. All ratings were standardized on a 0–7 scale.

A.4 Statistical Analysis and Visualization

The calculated agreement metrics were further analyzed and visualized:

- **Summary Statistics:** Mean, median, standard deviation, and count were calculated for Weighted Kappa and Spearman's Rho, grouped by the type of annotator pair (Human-Human, Human-Model, Model-Model). Similar summaries were computed for Krippendorff's Alpha, grouped by annotator group (Human, Model, All).
- **Visualizations:**
  - Box plots were generated to show the distribution of pairwise Weighted Kappa and Spearman's Rho across the different pair types.
  - A heatmap visualized the average Weighted Kappa for every annotator pair across all emotions. Annotators were ordered with human annotators listed first (alphabetically), followed by model annotators (alphabetically), for clarity.
  - Bar plots displayed Krippendorff's Alpha for the top N emotions (ranked by human annotator agreement) for the Human group, including reference lines indicating common interpretation thresholds (e.g., 0.80 for excellent, 0.67 for acceptable agreement).
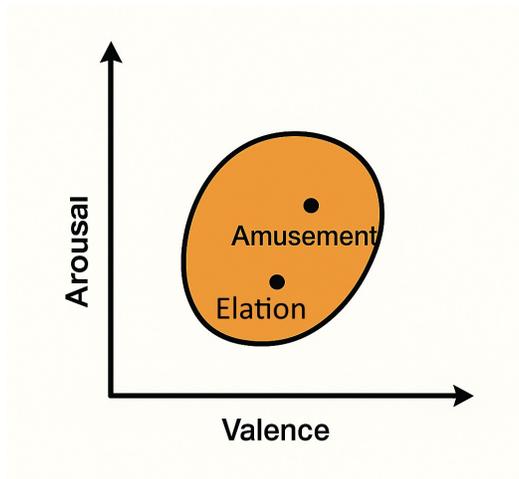
14

Figure 5: Rather than mapping each face to a single label, we estimate a distribution over plausible emotion categories.

- – Bar plots showed the average pairwise Weighted Kappa per emotion, faceted by pair type.
  - – Statistical comparisons were performed using the Mann-Whitney U test to evaluate significant differences in Weighted Kappa distributions between the different pair types (e.g., Human-Human vs. Human-Model). The results, including the difference in means, bootstrapped 95% confidence intervals for the difference, and p-values, were annotated onto the Kappa distribution box plot using the 'statannotations' library.
- • **Output Files:** Detailed results (pairwise scores, Alpha scores, summaries) were saved to CSV files in the 'output/tables' directory, and all generated plots were saved as PNG files in the 'output/figures' directory.

This comprehensive analysis allows for a detailed understanding of the reliability of the emotion ratings and the degree of alignment between human perception and VLM assessments within the EMO+ dataset.

The analysis was performed using Python (v3.10.12) with the libraries pandas (v2.2.2), numpy (v1.25.2), scikit-learn (v1.6.1), scipy, krippendorff (v0.8.1), seaborn (v0.13.2), matplotlib (v3.7.2), and statannotations (v0.7.2).

Table 5: The EmoNet-Face 40-Category Emotion Taxonomy. Each category represents a cluster of semantically related emotion words derived from the Handbook of Emotions and refined through expert consultation.

Category Name	Associated Descriptive Words
1. Amusement	'lighthearted fun', 'amusement', 'mirth', 'joviality', 'laughter', 'playfulness', 'silliness', 'jesting'
2. Elation	'happiness', 'excitement', 'joy', 'exhilaration', 'delight', 'jubilation', 'bliss', 'Cheerfulness'
3. Pleasure/Ecstasy	'ecstasy', 'pleasure', 'bliss', 'rapture', 'Beatitude'
4. Contentment	'contentment', 'relaxation', 'peacefulness', 'calmness', 'satisfaction', 'Ease', 'Serenity', 'fulfillment', 'gladness', 'lightness', 'serenity', 'tranquility'
5. Thankfulness/Gratitude	'thankfulness', 'gratitude', 'appreciation', 'gratefulness'
6. Affection	'sympathy', 'compassion', 'warmth', 'trust', 'caring', 'Clemency', 'forgiveness', 'Devotion', 'Tenderness', 'Reverence'

Continued on next page

Table 5 – continued from previous page

Category Name	Associated Descriptive Words
7. Infatuation	'infatuation', 'having a crush', 'romantic desire', 'fondness', 'butterflies in the stomach', 'adoration'
8. Hope/Optimism	'hope', 'enthusiasm', 'optimism', 'Anticipation', 'Courage', 'Encouragement', 'Zeal', 'fervor', 'inspiration', 'Determination'
9. Triumph	'triumph', 'superiority'
10. Pride	'pride', 'dignity', 'self-confidently', 'honor', 'self-consciousness'
11. Interest	'interest', 'fascination', 'curiosity', 'intrigue'
12. Awe	'awe', 'awestruck', 'wonder'
13. Astonishment/Surprise	'astonishment', 'surprise', 'amazement', 'shock', 'startlement'
14. Concentration	'concentration', 'deep focus', 'engrossment', 'absorption', 'attention'
15. Contemplation	'contemplation', 'thoughtfulness', 'pondering', 'reflection', 'meditation', 'Brooding', 'Pensiveness'
16. Relief	'relief', 'respite', 'alleviation', 'solace', 'comfort', 'liberation'
17. Longing	'yearning', 'longing', 'pining', 'wistfulness', 'nostalgia', 'Craving', 'desire', 'Envy', 'homesickness', 'saudade'
18. Teasing	'teasing', 'bantering', 'mocking playfully', 'ribbing', 'provoking lightly'
19. Impatience and Irritability	'impatience', 'irritability', 'irritation', 'restlessness', 'short-temperedness', 'exasperation'
20. Sexual Lust	'sexual lust', 'carnal desire', 'lust', 'feeling horny', 'feeling turned on'
21. Doubt	'doubt', 'distrust', 'suspicion', 'skepticism', 'uncertainty', 'Pessimism'
22. Fear	'fear', 'terror', 'dread', 'apprehension', 'alarm', 'horror', 'panic', 'nervousness'
23. Distress	'worry', 'anxiety', 'unease', 'anguish', 'trepidation', 'Concern', 'Upset', 'pessimism', 'foreboding'
24. Confusion	'confusion', 'bewilderment', 'flabbergasted', 'disorientation', 'Perplexity'
25. Embarrassment	'embarrassment', 'shyness', 'mortification', 'discomfiture', 'awkwardness', 'Self-Consciousness'
26. Shame	'shame', 'guilt', 'remorse', 'humiliation', 'contrition'
27. Disappointment	'disappointment', 'regret', 'dismay', 'letdown', 'chagrin'
28. Sadness	'sadness', 'sorrow', 'grief', 'melancholy', 'Dejection', 'Despair', 'Self-Pity', 'Sullenness', 'heartache', 'mournfulness', 'misery'
29. Bitterness	'resentment', 'acrimony', 'bitterness', 'cynicism', 'rancor'
30. Contempt	'contempt', 'disapproval', 'scorn', 'disdain', 'loathing', 'Detestation'
31. Disgust	'disgust', 'revulsion', 'repulsion', 'abhorrence', 'loathing'
32. Anger	'anger', 'rage', 'fury', 'hate', 'irascibility', 'enragement', 'Vexation', 'Wrath', 'Peevishness', 'Annoyance'
33. Malevolence/Malice	'spite', 'sadism', 'malevolence', 'malice', 'desire to harm', 'schadenfreude'
34. Sourness	'sourness', 'tartness', 'acidity', 'acerbity', 'sharpness'
35. Pain	'physical pain', 'suffering', 'torment', 'ache', 'agony'
36. Helplessness	'helplessness', 'powerlessness', 'desperation', 'submission'
37. Fatigue/Exhaustion	'fatigue', 'exhaustion', 'weariness', 'lethargy', 'burnout', 'Weariness'
38. Emotional Numbness	'numbness', 'detachment', 'insensitivity', 'emotional blunting', 'apathy', 'existential void', 'boredom', 'stoicism', 'indifference'
39. Intoxication/Altered States	'being drunk', 'stupor', 'intoxication', 'disorientation', 'altered perception'
40. Jealousy & Envy	'jealousy', 'envy', 'covetousness'

A.5 Detailed Image Generation Models and Reproducibility

For transparency and reproducibility in the generation of the EMONET-FACE datasets, all prompt texts and specific model identifiers used for image creation will be released alongside the dataset.
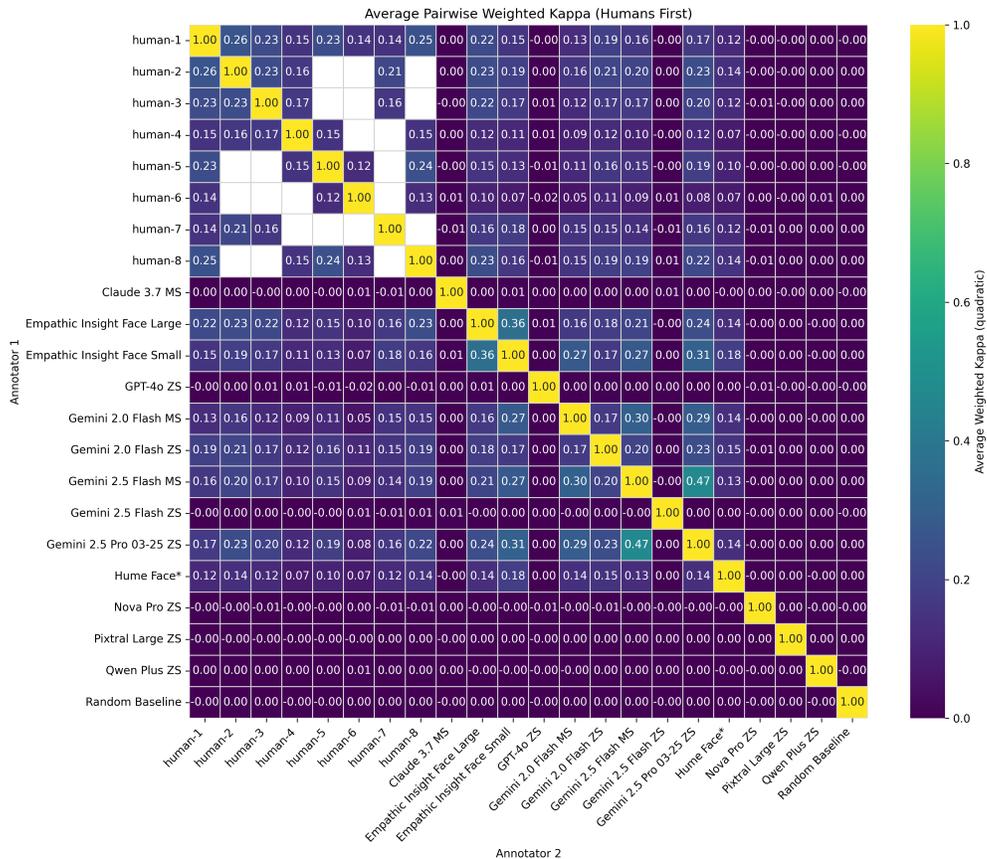
Figure 6: Heatmap of Average Pairwise Weighted Kappa (κ_w, quadratic weights) Across All Emotions. Annotators are grouped with humans listed first (alphabetically), followed by models (alphabetically). Values closer to 1 indicate higher agreement. The distinct block of warmer colors for Human-Human pairs highlights their higher relative agreement compared to Human-Model and Model-Model pairs.

Table 6: Internal Agreement for Repeated Annotations

Annotator	Images Annotated	Internal Agreement
Annotator 1	2000	86.10%
Annotator 2	2000	96.20%
Annotator 3	2000	84.77%

The EMONET-FACE HQ dataset (2,500 images) was generated using a combination of MidJourney v6 [32] and Flux-Dev and -Pro [6]. The larger EMONET-FACE BINARY (19,999 images) and EMONET-FACE BIG (over 203,000 images) datasets were produced exclusively with Flux-Dev.

We deliberately selected multiple state-of-the-art (SOTA) text-to-image (T2I) models to minimize potential confounders and ensure high image quality across the diverse range of emotions and demographic attributes. During this selection process, we intentionally excluded models such as Juggernaut-XL or Kolors from our final pipeline due to their propensity for frequent visual artifacts, including exaggerated wrinkles and unnatural facial features, which were not easily mitigated through prompt engineering. The prompts for all models specified sharp, close-up portraits with neutral backgrounds to maximize emotion clarity and consistently maintain demographic diversity across all generated datasets.

Category	General Emotion	Example Emotions
Positive High-Energy Emotions	Amusement	lighthearted fun, amusement, mirth, joviality, laughter, playfulness, silliness, jesting, delight
	Elation	happiness, excitement, joy, exhilaration, laughter, jubilation, cheerfulness
	Pleasure/Ecstasy	ecstasy, pleasure, bliss, beatitude
	Hope/Enthusiasm/Optimism	hope, enthusiasm, optimism, anticipation, courage, encouragement, zeal, fervor, inspiration, determination
	Triumph	triumph, superiority, hubris
	Interest	interest, fascination, curiosity, need for cognition
	Awe	awe, awestruck, wonder
	Astonishment/Surprise	astonishment, surprise, amazement, shock, startlement
	Teasing	teasing, bantering, mocking playfully, ribbing, provoking lightly
Positive Low-Energy Emotions	Contentment	contentment, relaxation, peacefulness, calmness, satisfaction, ease, serenity, bliss, fulfillment, gladness, lightness, serenity, tranquility
	Thankfulness/Gratitude	thankfulness, gratitude, appreciation, gratefulness
	Affection	sympathy, compassion, warmth, trust, caring, forgiveness, devotion, tenderness, reverence
	Relief	relief, respite, alleviation, solace, comfort, liberation
	Contemplation	contemplation, thoughtfulness, pondering, reflection, meditation, brooding, pensiveness
	Pride	pride, dignity, self-confidently, honor, self-consciousness
Negative High-Energy Emotions	Fear	fear, terror, dread, apprehension, alarm, horror, panic, nervousness
	Anger	anger, rage, fury, hate, irascibility, enragement, wrath, annoyance
	Malevolence/Malice	spite, sadism, malevolence, malice, malicious envy, revenge, desire to harm, schadenfreude
	Disgust	disgust, revulsion, repulsion, abhorrence, loathing
	Impatience and Irritability	impatience, irritability, irritation, restlessness, short-temperedness, exasperation
	Distress	worry, anxiety, unease, anguish, trepidation, concern, upset, pessimism, foreboding
Negative Low-Energy Emotions	Sadness	sadness, sorrow, grief, melancholy, dejection, despair, self-pity, sullenness, heartache, mournfulness, misery
	Bitterness	resentment, acrimony, bitterness, cynicism, rancor, malicious envy
	Contempt	contempt, disapproval, scorn, disdain, loathing, detestation
	Disappointment	disappointment, regret, dismay, letdown, chagrin
	Shame	shame, guilt, remorse, humiliation, contrition
	Emotional Numbness	numbness, detachment, insensitivity, emotional blunting, apathy, existential void, boredom, stoicism, indifference
	Doubt	doubt, distrust, suspicion, skepticism, uncertainty, pessimism
	Jealousy & Envy	jealousy, envy, covetousness
	Embarrassment	embarrassment, shyness, mortification, discomfiture, awkwardness, self-consciousness
	Helplessness	helplessness, hopelessness, powerlessness, desperation, submission
Cognitive States and Processes	Concentration	concentration, deep focus, engrossment, absorption, attention
	Confusion	confusion, bewilderment, disorientation, perplexity
Physical and Exhaustive States	Pain	physical pain, suffering, torment, ache, agony
	Fatigue/Exhaustion	fatigue, exhaustion, weariness, lethargy, burnout, Weariness
	Intoxication/Altered States of Consciousness	being drunk, stupor, intoxication, disorientation, altered perception
	Sourness	sourness, tartness, acidity, acerbity, sharpness
Longing & Lust	Longing	yearning, longing, pining, wistfulness, nostalgia, craving, desire, benign envy, homesickness
	Sexual Lust	sexual lust, carnal desire, lust, feeling horny, feeling turned on
	Infatuation	infatuation, having a crush, romantic desire, fondness, butterflies in the stomach, adoration

Figure 7: Emotion classification taxonomy showing hierarchical relationships between primary categories, general emotions, and specific descriptive terms. The complete classification (all levels) was used for `EmoNetHQ` annotations, while `EmoNetBinary` annotations utilized only the descriptive terms (bottom level).

B Dataset Construction Details

B.1 Construction of EMONET-FACE BIG

The EMONET-FACE BIG dataset, designed primarily for large-scale model pre-training, was constructed through a multi-stage process. This involved an initial phase of broad-spectrum image generation and annotation, followed by targeted generation and a refined annotation strategy to ensure comprehensive coverage across our 40-emotion taxonomy.

Initial Data Generation and Annotation The foundation of EMONET-FACE BIG was formed by incorporating all images from the EMONET-FACE BINARY dataset, augmented with an additional 20,000 images generated using the Flux text-to-image model. This combined set of images was then annotated using the Gemini 2.5 Flash model. The annotation prompt directed Gemini 2.5 Flash to identify the five most salient emotional dimensions present in each image and to assign an intensity score ranging from 0 to 7 for each of these selected dimensions. While this initial phase yielded a substantial volume of annotations, a qualitative review revealed that certain emotion categories within our taxonomy—such as infatuation, embarrassment, and sexual lust—received disproportionately few non-zero annotations from Gemini 2.5 Flash.

Table 7: Agreement summary (average weighted Kappa κ_w with quadratic weights) for all annotators: mean, standard deviation, median, and top emotion (with highest agreement) per annotator. For humans, agreement is with all other humans; for models, agreement is with all humans. Table is sorted in descending order of the mean.

Annotator	mean	std	median	top-1
human-2	0.2185	0.1541	0.2010	Elation (0.6330)
human-1	0.2012	0.1266	0.2139	Elation (0.5259)
human-3	0.1972	0.1502	0.1832	Elation (0.5766)
human-8	0.1908	0.1349	0.1562	Elation (0.6365)
human-5	0.1857	0.1277	0.1646	Elation (0.5866)
Empathic Insight Face Large	0.1795	0.1217	0.2008	Astonishment/Surprise (0.4981)
human-7	0.1720	0.1452	0.1546	Anger (0.5689)
Gemini 2.5 Pro 03-25 ZS	0.1707	0.1651	0.1192	Elation (0.6025)
Gemini 2.0 Flash ZS	0.1624	0.1562	0.1621	Elation (0.6059)
human-4	0.1540	0.1448	0.1210	Amusement (0.6290)
Gemini 2.5 Flash MS	0.1502	0.1599	0.0855	Elation (0.5824)
Empathic Insight Face Small	0.1443	0.1715	0.0806	Elation (0.6592)
human-6	0.1309	0.1057	0.1064	Elation (0.4198)
Gemini 2.0 Flash MS	0.1195	0.1522	0.0512	Elation (0.5639)
Hume Face*	0.1087	0.1429	0.0693	Elation (0.6276)
Qwen Plus ZS	0.0024	0.0047	0.0002	Contentment (0.0167)
Gemini 2.5 Flash ZS	0.0007	0.0085	0.0002	Amusement (0.0219)
Random Baseline	0.0003	0.0075	0.0015	Elation (0.0243)
Claude 3.7 MS	-0.0002	0.0101	-0.0000	Pride (0.0198)
Pixtral Large ZS	-0.0004	0.0055	-0.0008	Amusement (0.0145)
GPT-4o ZS	-0.0022	0.0078	-0.0007	Anger (0.0108)
Nova Pro ZS	-0.0030	0.0100	-0.0029	Distress (0.0140)

Targeted Generation and Annotation with Hinting Strategy To address the underrepresentation of these specific emotion categories and to enrich the dataset with more diverse examples for these less frequently annotated states, we implemented a revised two-step procedure:

1. **Targeted Image Generation:** We utilized the Flux-Dev model to generate new images. The prompts for Flux-Dev were specifically designed to elicit facial expressions corresponding to the underrepresented emotions. For instance, prompts targeted the generation of images depicting "embarrassment" or "infatuation."

2. **Annotation with a Hinting Strategy:** These newly generated, targeted images were subsequently annotated using Gemini Flash 2.0. The selection of Gemini Flash 2.0 for this phase was partly influenced by more favorable API rate limits available at the time of data collection, which was crucial for the iterative process. The core annotation instruction remained to score the top five most evident emotional dimensions on the 0-7 scale. However, a key modification was introduced: a "hint" was provided to the model. For example, if an image was generated with "infatuation" as the target emotion in the prompt, the annotation prompt for Gemini Flash 2.0 would include a suggestion that the image *might* contain "infatuation." This was intended to encourage the model to consider this specific dimension more readily, without mandating its selection or biasing the intensity score.

This hinting strategy proved effective. Qualitative inspection of the results indicated that for approximately 50% of the images where a hint was provided for a specific target emotion, Gemini Flash 2.0 assigned a non-zero annotation to that particular dimension. The quality of these hinted annotations was deemed reasonable and sufficient for the purposes of large-scale pre-training.

Iterative Refinement and Final Dataset Composition We iteratively applied this targeted image generation (with Flux-Dev) and hinted annotation (with Gemini Flash 2.0) process. The primary objective of this iterative refinement was to ensure that each of the 40 emotional dimensions in our taxonomy received approximately 5,000 samples with a non-zero annotation score attributed by Gemini Flash 2.0 (under the described hinting strategy). This iterative process continued until the desired representation was achieved for all emotion categories.

The final EMONET-FACE BIG dataset comprises a total of 203,201 images, each with synthetic emotion annotations derived through this combined approach of broad-spectrum initial generation/annotation and subsequent targeted generation/hinted-annotation.

B.2 Training of Baseline Models

To establish baseline performance for emotion dimension prediction, we developed and trained several Multi-Layer Perceptron (MLP) models. The process began with feature extraction: we computed SIGLIP2-400M image embeddings (1152 dimensions) for all images within the EMoNet-Face Big, and EmoNetFace Binary datasets. These embeddings served as the input features for our models.

Our initial approach involved training 40 distinct MLP models, one for each of the 40 emotion dimensions. Each MLP was designed to take an image's SIGLIP2-400M embedding as input and predict the rating for its specific emotion dimension as a continuous score (ranging from 0 to 7). To calibrate these continuous predictions, we addressed potential biases in the models' interpretation of neutral expressions. For each of the 40 MLP models, we performed inference on 1700 images generated by FluxDev. These images depicted faces with neutral emotional expressions, created using prompts randomly varied across the complete template distribution for different ages, ethnic backgrounds, and genders. We then calculated the average rating assigned by each MLP model to these neutral faces. This average neutral rating, which varied per model, was subsequently subtracted from all predictions made by that specific MLP to correct for its baseline offset on neutral expressions.

We first developed a "small model" configuration. For these small models, training was conducted exclusively on the EMoNet-Face Big dataset. To address class imbalance inherent in the 0-7 rating scale for each emotion dimension, we employed a specific sampling strategy, referred to as the "stumbling strategy," to construct balanced training datasets. For a given emotion dimension, samples were first grouped into eight buckets corresponding to ratings 0 through 7. We then calculated the average number of images across these eight buckets. The training set for that dimension was constructed by sampling up to 25 percent of this average count from each bucket. This approach was adopted because the bucket for rating '0' (emotion not present) typically contained a significantly larger number of samples (e.g., 180,000-190,000) compared to other rating buckets. Consequently, the average number of samples per bucket might be around 20,000, leading to approximately 5,000 samples from the '0' bucket and proportionally fewer (e.g., around 1,000) from each of the other buckets (ratings 1-7). A validation set was created by reserving 5 percent of the samples from each bucket using the same principle. The small MLP architecture consisted of an input layer (1152 features), a hidden layer with 128 neurons (ReLU, Dropout 0.1), a second hidden layer with 32 neurons (ReLU, Dropout 0.1), and an output layer with 1 neuron. This architecture has 151,745 parameters. The checkpoint performing best on the validation set was saved.

Following experiments with the small model, a grid search over different model sizes indicated that a larger MLP architecture could yield modest improvements in validation accuracy. This led us to transition our experiments to a "big model." The big MLP architecture is defined as follows: an input layer (1152 features); a first hidden layer with 1024 neurons (ReLU, Dropout 0.2); a second hidden layer with 512 neurons (ReLU, Dropout 0.2); a third hidden layer with 256 neurons (ReLU, Dropout 0.2); and an output layer with 1 neuron. This larger architecture comprises 1,837,057 parameters.

Initially, this big model was trained on EMoNet-Face Big using the same strategy of balancing datasets for each emotion dimension by sampling from each 0-7 rating bucket up to a limit (25 percent of the average samples per bucket) to reduce the dominance of the overrepresented zero-rating for continuous score prediction as the small model. However, after these initial explorations, we shifted the dataset composition and prediction task. For each emotion dimension, we transformed the problem into a binary classification task. Images with a rating of 0 were assigned to a '0' bucket (emotion not present). All images with ratings from 1 to 7 were merged into a single '1' bucket (emotion present, irrespective of intensity). The big model was then trained to predict either 0 or 1 via regression. The EMoNet-Face Big dataset was prepared for this binary task by binarizing its ratings as described. Due to the '0' bucket being substantially larger than the '1' bucket even after binarization, we balanced the EMoNet-Face Big training data by downsampling the '0' bucket to match the number of samples in the '1' bucket. The human-annotated EmoNetFace Binary dataset was also prepared by binarizing its ratings in the same manner.

We explored different training regimes for the big binary model. This included training solely on the (binarized) human-annotated EmoNetFace Binary dataset and training solely on the (binarized and balanced) synthetic EMoNet-Face Big dataset. Through experiments on a validation set (sampled as 5 percent from each bucket of the respective training data), we determined that the optimal performance was achieved by first pre-training the model on the large, synthetically annotated EMoNet-Face Big dataset and subsequently fine-tuning it on the smaller, human-annotated EmoNetFace Binary dataset.

Our final big model was trained using this two-stage approach. First, it was pre-trained for 5 epochs on the binarized and balanced EMoNet-Face Big dataset. For this stage, we used a learning rate of 5e-5 and a ReduceLROnPlateau learning rate scheduler (monitoring validation loss, factor=0.2, patience=10). After pre-training, the training was restarted with the same initial learning rate of 5e-5 on the binarized human-annotated EmoNetFace Binary dataset. Fine-tuning on this dataset continued for 200 epochs. Throughout both training stages, the checkpoint that achieved the best performance on the corresponding validation set was saved for later use.
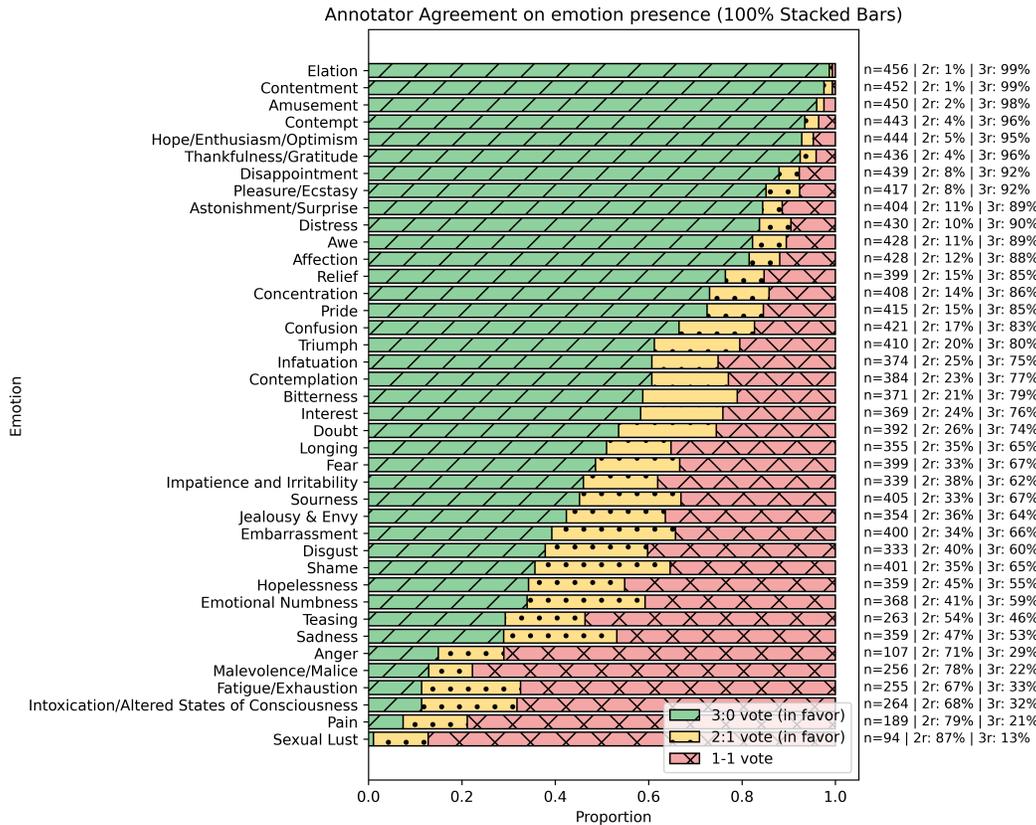

Figure 8: Annotator agreement for Human Annotators (n=4) in EMONET-FACE BINARY. Stacked bars show the proportion of image-emotion pairs with full agreement (all annotators in favor of emotion presence), partial agreement (2:1 split in favor), and disagreement (1:1 split) for each emotion. The numbers to the right of each bar indicate the total number of samples (n) and the percentage of cases rated by 2 or 3 annotators. Note that our annotation process was designed to only annotate positively rated image a second time, and double-positively rated images a third time.

B.3   Preference Study: Model vs. Human Perception Alignment

To evaluate how well our model (`Empathic Insight Face Large`) aligns with human perception, we conducted a blind preference study using 999 of 2,500 total images of EMONET-FACE HQ (see Appendix C for UI details). Two annotators, not involved in initial labeling, each reviewed 499 to 500 images and compared two anonymized emotion label sets: (1) the human median of the five highest-rated emotions and (2) the model's five top-confidence predictions. Annotators selected the set that best matched the depicted expression. Of 999 collected judgments, 707 (70.77%)
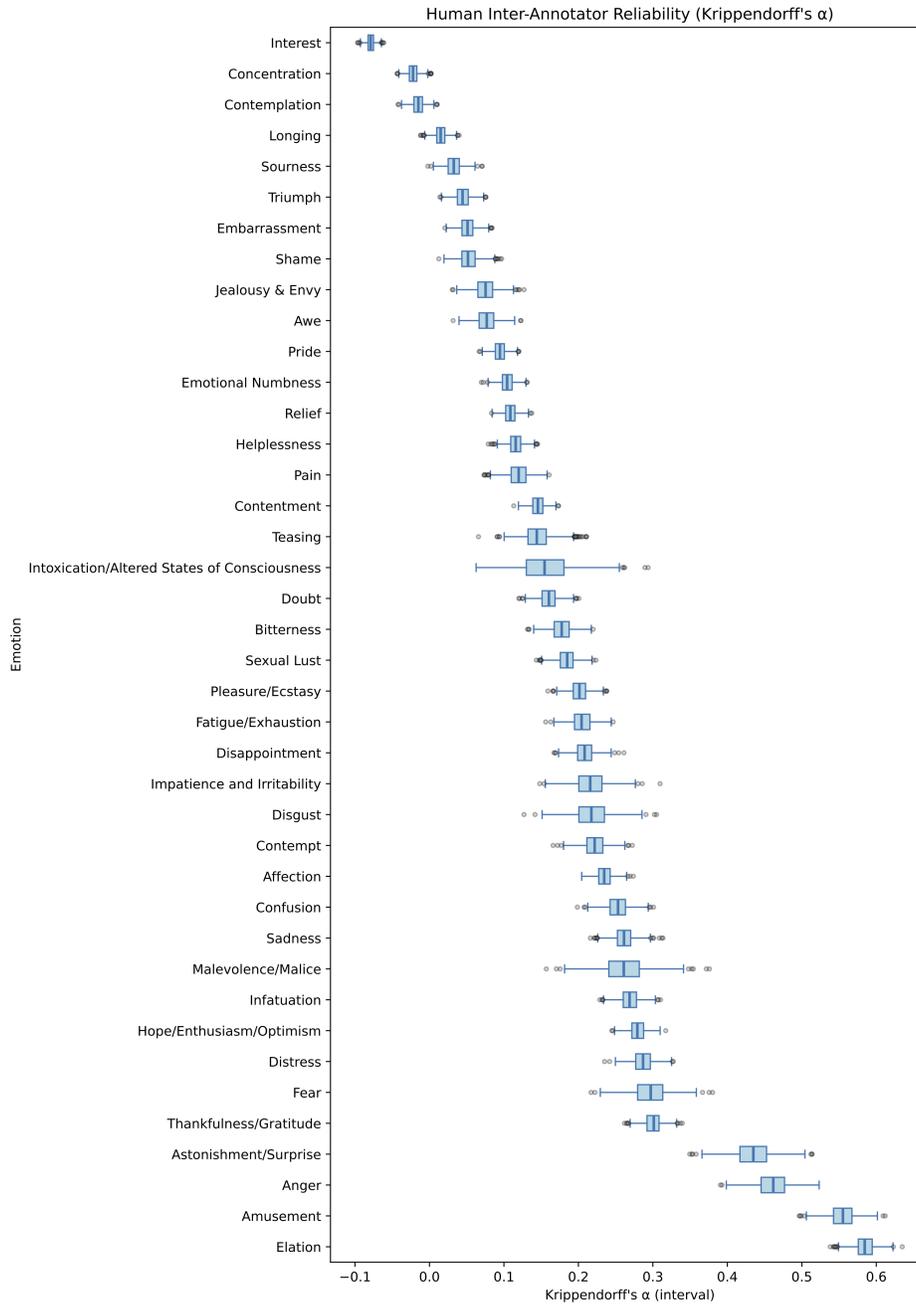
Figure 9: Krippendorff's Alpha (α, interval level) for Human Annotators (n=8) in EMONET-FACE HQ. Boxplot depicts median, IQR, and outliers across all 40 Emotions, with higher bars indicate better agreement.

favored human labels, while 292 (29.23%) preferred the model's output. Though human experts were preferred overall, our model's top 5 predictions had been preferred to the average of 4 human expert annotations in almost 30% of the samples. These results suggest the model can remarkably well approximate human judgment, especially in contexts where human annotation is impractical due to scale or cost. Future work may focus on fine-tuning to address specific expression types where model performance lags.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction accurately state the paper's core contributions: the introduction of EmoNet-Face, a comprehensive suite comprising a novel 40-category emotion taxonomy, two expert-annotated synthetic image datasets, one qualitative, one binary, and a high-performing emotion estimation model that beats SOTA generative foundation models in human-like emotion estimation. The scope is clearly defined as advancing synthetic emotion recognition from facial expressions to foster more empathetic and equitable AI.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses limitations inherent in emotion recognition from facial cues alone (see intra-rater agreement analysis of the binary task), especially for fine-grained categories, aligning with the Theory of Constructed Emotion (Section 3.1). It also explicitly states the scope is currently limited to facial expressions, deferring speech and reasoning to future work (Section 1).

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best

judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification:This paper is primarily empirical, focusing on dataset creation, annotation, and model evaluation. It discusses established psychological theories (Section 2) to inform the taxonomy development but does not present novel theoretical results requiring formal proofs.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper describes the synthetic image generation process (Section 3.2), the detailed multi-phase human annotation methodology (Section 3.3), and the statistical analysis approach for inter-annotator agreement (Section 4). Details for reproducing the evaluation of models against human judgments and training the provided model will be included with the public code release.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.

(b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.

(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The entire EmoNet-Face suite, including the taxonomy, datasets (Qualitative, Binary, Big), the custom annotation platform code, the analysis and prompting scripts, and the EmoNetFace Model training code and model weights, will be made publicly available upon publication (Footnote 1, Section 1, Section 3.4). Instructions for use and reproduction will be provided under a permissive license.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Yes, all the training and test details (e.g., data splits, hyperparameters, etc.) necessary to understand the results are added in the paper (also in the supplementary materials) and in the appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.

- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The paper reports inter-annotator agreement using metrics like Cohen's Weighted Kappa, Spearman's Rank Correlation, and Krippendorff's Alpha (Section 4.2), which quantify the variability and agreement among human raters. Performance evaluation of models against human judgments (Section 3.3, Section 4) will include appropriate statistical measures.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Information on the compute resources used for synthetic image generation, annotation processing, and model training/evaluation will be provided in the supplementary material to aid reproducibility.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

Answer: [Yes]

Justification: The research adheres to the Code of Ethics as detailed in the dedicated "Ethics Discussion" section. It addresses responsible AI development, discusses risks of misuse, details safeguards against bias and harmful content in synthetic data, and confirms no personally identifiable information was used.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: The paper discusses potential positive impacts, such as enabling empathetic AI, in the Introduction (Section 1) and Conclusion. Potential negative impacts, such as misuse and bias, are discussed in the dedicated "Ethics Discussion" section.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [Yes]

Justification: As described in Section 3.4 and the "Ethics Discussion", safeguards like automated filtering, manual review, and detailed tagging were implemented during dataset creation to mitigate risks associated with potential harmful stereotypes or inappropriate content in the synthetic images.

Guidelines:

- The answer NA means that the paper poses no such risks.

- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The paper cites external assets used, including psychological literature and text-to-image generative models (Sections 2, 3.1, 3.2). Compliance with their respective terms and licenses will be ensured, and licenses for the released assets will be provided (Section 13).

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: The paper introduces a novel taxonomy, three datasets, an annotation platform, analysis code, and a model as new assets. These are documented within the paper (Sections 3, 4, Technical Limitations, Ethics Discussion), and detailed documentation, including a datasheet for the datasets, will be provided alongside the public release.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [Yes]

Justification: Yes, all annotation instructions are available in the appendix section as screenshots. All annotators were compensated (see methodology section).

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [Yes]

Justification: The human annotation process involved minimal risk and was conducted in accordance with institutional ethical guidelines. Internal, informal ethical review, participant consent was obtained.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: The paper describes the use of a large language model (GPT-4) for taxonomy development (Section 3.1) and text-to-image generative models (Flux-Dev, Midjourney 6) as important and non-standard components of the synthetic image generation methodology (Section 3.2), which are central to the dataset creation.

Image: flux_prompt_9-image_246.png
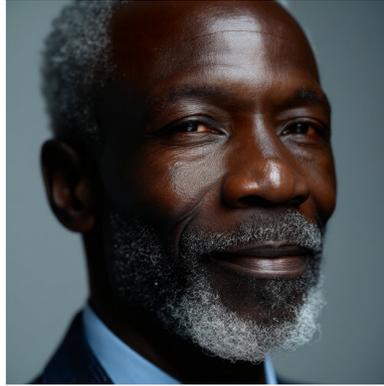
annotator	human-1	human-2	human-3	human-7	gpt-4o	nova-pro	pixtral-large-latest	qwen-plus
emotion								
Cognitive States and Processes **Concentration**	5	2	0	0	5	5	4	4
Cognitive States and Processes **Confusion**	0	0	0	0	4	3	2	3
Longing & Lust **Infatuation**	4	0	0	0	0	0	0	0
Longing & Lust **Sexual Lust**	1	0	0	0	0	0	0	0
Negative High-Energy Emotions **Anger**	0	0	0	0	2	6	1	1
Negative High-Energy Emotions **Disgust**	0	0	0	0	2	3	2	1
Negative High-Energy Emotions **Distress**	0	0	0	0	3	3	3	3
Negative High-Energy Emotions **Fear**	0	0	0	0	1	4	2	1
Negative High-Energy Emotions **Impatience and Irritability**	0	0	0	0	1	3	2	2
Negative High-Energy Emotions **Malevolence/Malice**	0	0	0	0	3	2	1	1
Negative Low-Energy Emotions **Bitterness**	0	0	0	0	1	2	2	1
Negative Low-Energy Emotions **Contempt**	0	0	0	0	1	3	1	1
Negative Low-Energy Emotions **Disappointment**	0	0	0	0	2	4	3	2
Negative Low-Energy Emotions **Doubt**	0	0	0	0	2	3	3	2
Negative Low-Energy Emotions **Emotional Numbness**	0	0	0	0	3	1	1	1
Negative Low-Energy Emotions **Jealousy & Envy**	0	0	0	0	1	3	2	1
Negative Low-Energy Emotions **Sadness**	0	0	0	0	4	4	4	1
Negative Low-Energy Emotions **Shame**	0	0	0	0	1	2	2	1
Physical and Exhaustive States **Fatigue/Exhaustion**	0	0	0	0	2	2	3	2
Physical and Exhaustive States **Intoxication/Altered States of Consciousness**	0	0	0	0	1	1	1	1
Positive High-Energy Emotions **Amusement**	2	3	0	0	0	0	0	0
Positive High-Energy Emotions **Astonishment/Surprise**	0	0	0	0	5	5	4	6
Positive High-Energy Emotions **Awe**	0	0	0	0	4	2	1	4
Positive High-Energy Emotions **Elation**	4	0	3	2	0	0	0	0
Positive High-Energy Emotions **Hope/Enthusiasm/Optimism**	4	0	4	0	0	0	0	0
Positive High-Energy Emotions **Interest**	5	2	0	0	2	6	3	5
Positive High-Energy Emotions **Pleasure/Ecstasy**	1	1	4	0	0	0	0	0
Positive High-Energy Emotions **Teasing**	4	0	0	0	0	0	0	0
Positive High-Energy Emotions **Triumph**	2	3	0	3	3	2	2	3
Positive Low-Energy Emotions **Affection**	5	2	5	0	0	0	0	0
Positive Low-Energy Emotions **Contemplation**	5	3	0	0	1	4	5	7
Positive Low-Energy Emotions **Contentment**	4	2	5	2	5	1	3	3
Positive Low-Energy Emotions **Pride**	3	4	7	4	4	5	2	2
Positive Low-Energy Emotions **Relief**	5	0	5	0	0	0	0	0
Positive Low-Energy Emotions **Thankfulness/Gratitude**	5	0	6	0	0	0	0	0

Figure 10: Discrepancies between human annotators and zero-shot prompted VLMs.

C Annotation Platform Instructions and UI

Instructions

This project is about assessing the emotions that a person, shown in an image, appears to be feeling, according to **your opinion**. All faces that you find in this project are generated by AI and do not resemble any real existing person. This study aims to improvethe abilities of AI assistants to empathetically react to their user's feelings in contexts like counseling, therapy, and social support.

To help these systems to make a realistic assessment of their users' feelings, it is important that you fulfill these tasks with your honest **uncensored** and **spontaneous intuitions**.

Do not hesitate to omit certain categories you consider irrelevant for a given image or rate other categories highly if you feel this is appropriate.

Emotion recognition is always influenced by cultural and social factors and is therefore never 100% objective. Therefore, **don't worry too much about whether your evaluation is right or wrong**, but listen to your intuition and first impressions. After chosing the categories, you will adjust the slider from 0 (emotions not present) over 1 (emotions are weakly present) to 7 (the emotions are very strong). If you are unsure, take another look at the face and ask yourself how much this person seems to be feeling for the given emotions right now, and then adjust the slider to the number that comes to your mind.

Whatever comes to your mind: **It is okay**.

Recommendations
1) Close your eyes for a moment when you start annotating a new sample and focus for ~10 seconds on your breath, until you feel calm & relaxed. Feeling relaxed will make you more perceptive of small nuances in the face and make it easier for you to listen to your intuition.
2) Then look at the face & ask yourself for each category: **How much does the category seem to describe what this person seems to be feeling**?
3) Repeat the process for the emotions: **How much do the given emotions seem to describe what this person seems to be feeling**? Instead of if a category is present or not (clicked or not), you now adjust the slider accordingly (leave it as it is for not present, or slide it on a value between 1 to 7 if present). Trust your gut feeling, your subjective intuition. It is fine.
4) After you have processed all emotions move to the next sample starting with step 1) by clicking Submit and Proceed .

Do the annotations in a calm environment in a relaxed state of mind. A tea or coffee (with or without cookies) and listening to relaxing music might help.

Listen to your heart, **trust your intuition & try to have fun**. 🙂

Figure 11: Instructions given to the human annotator for the expert annotation of EMONET-FACE HQ.
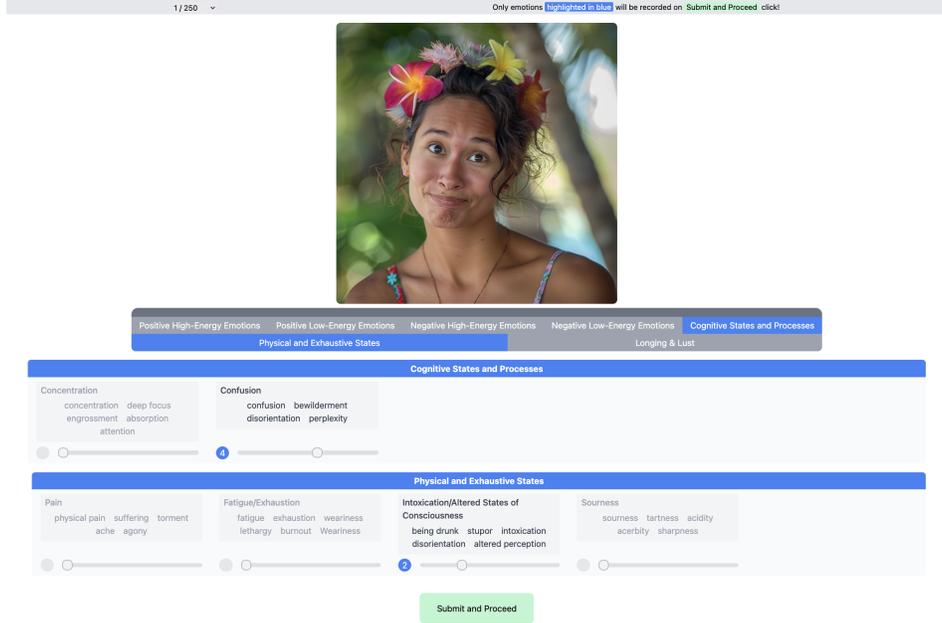
Figure 12: UI of our expert annotation tool for EMONET-FACE HQ.

D   Taxonomy Mapping for Real-World Generalization

To assess the generalization capability of our EMPATHICINSIGHT-FACE model on established real-world emotion recognition datasets such as FERD and AffectNet (see Section 4.3), we mapped our fine-grained 40-category EMONET-FACE taxonomy to the 8 basic emotion categories used in these benchmarks. Because FERD and AffectNet differ slightly in their label definitions, we derived separate mappings for each dataset, as detailed in Table 8. This mapping process necessarily involves some loss of granularity, as nuanced affective states are compressed into broader emotion classes. For instance, both *Elation* and *Amusement* are mapped to *Happy*, while *Awe* corresponds to *Surprise* on FERD but to *Fear* on AffectNet. The resulting target categories—*Anger*, *Contempt*, *Disgust*, *Fear*, *Happy*, *Neutral*, *Sad*, and *Surprise*—align with the shared emotion taxonomy across these benchmarks.

Emotion Classification Tutorial

Instructions

In this task, you'll be assessing whether a specific emotion appears to be present in a face. Each face you see is generated by AI and does not resemble any real person. The goal of this study is to improve AI's ability to empathize by recognizing emotions in counseling, therapy, and social support settings.

To help make these systems better at detecting emotions, your honest, uncensored, and spontaneous responses are essential.

Task Overview

You will be shown a face along with a list of emotions. For each emotion, your task is simply to decide Yes (emotion is present) or No (emotion is not present). This is not a precise science, so rely on your intuition and first impressions to make your decisions.

Remember, emotion recognition is influenced by cultural and social factors and is therefore subjective. There are no "right" or "wrong" answers here — just your honest responses based on how you perceive the expressions in the faces.

Recommendations

- **Relax First:** Close your eyes for a moment before starting a new image and take a few deep breaths. This can help you feel calm and improve your awareness of subtle facial cues.

- **Observe & Decide:** Look at the face and consider each emotion in turn. Ask yourself, "Does this person seem to be feeling this emotion right now?" Then answer Yes or No based on your gut feeling.

- **Repeat & Submit:** After answering for each emotion, click Submit and Proceed to move to the next image. Repeat the process with the same relaxed focus.

- **Create a Calm Environment:** Working in a calm space with minimal distractions, possibly with a tea or coffee, can make the task more enjoyable and enhance your intuition.

Figure 13: Instructions given to the human annotator for the expert annotation of EMONET-FACE BINARY.

resentment, acrimony, bitterness, cynicism, rancor

Are the emotions present in the image?
(You can also click "y" for Yes and "n" for No on the keyboard.)

Yes (y)   No (n)

Progress: 2 / 2000 | Last annotation: 57486ms

Figure 14: UI of the self-developed expert annotation tool for EMONET-FACE BINARY.

Preference Annotation Task

Please follow these steps for the following annotation task

1. Examine the person in the image carefully.

2. Choose the emotion group (left or right) that most accurately reflects the person's emotions and facial expression. Each group displays up to the top five emotions, ranked in descending order by annotator or model score.

3. Submit your choice by either:
   - Pressing the ← or → arrow key
   - Clicking the corresponding on-screen arrow button

Please work systematically and trust your first impression. Thank you for your careful annotations.

Figure 15: Instructions given to the human annotators for the preference annotation of of EMONET-FACE HQ.

4 / 250

Use the **left** and **right** arrow keys, or click the **arrow buttons**, to choose the emotions that best describe the person.

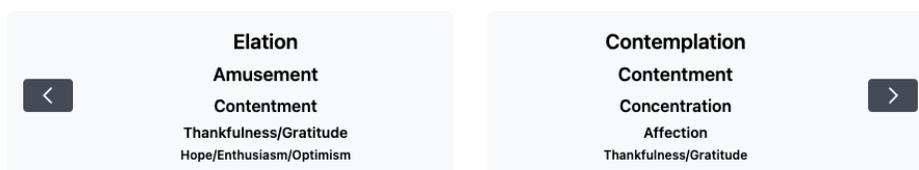| Elation	Contemplation
Amusement	Contentment
Contentment	Concentration
Thankfulness/Gratitude	Affection
Hope/Enthusiasm/Optimism	Thankfulness/Gratitude

Figure 16: UI of our expert preference annotation tool for EMONET-FACE HQ.

35

Table 8: Mapping from our 40-category EMONET-FACE taxonomy to the 8 basic emotion categories used for evaluation on the FERD and AffectNet datasets.

EMONET-FACE Category	FERD Mapping	AffectNet Mapping
Affection	Happy	Neutral
Anger	Anger	Anger
Astonishment / Surprise	Fear	Surprise
Awe	Surprise	Fear
Bitterness	Sad	Sad
Concentration	Neutral	Neutral
Contemplation	Neutral	Neutral
Contempt	Contempt	Contempt
Contentment	Neutral	Neutral
Confusion	Surprise	Neutral
Disappointment	Sad	Anger
Disgust	Disgust	Disgust
Distress	Fear	Sad
Doubt	Fear	Fear
Elation	Happy	Happy
Embarrassment	Anger	Anger
Emotional Numbness	Neutral	Neutral
Fatigue / Exhaustion	Neutral	Neutral
Fear	Surprise	Fear
Helplessness	Fear	Sad
Hope / Enthusiasm / Optimism	Happy	Neutral
Impatience and Irritability	Anger	Anger
Infatuation	Happy	Sad
Interest	Neutral	Neutral
Intoxication / Altered States of Consciousness	Neutral	Neutral
Jealousy & Envy	Anger	Disgust
Lighthearted Fun / Amusement	Happy	Happy
Longing / Yearning	Sad	Sad
Malevolence / Malice	Contempt	Anger
Pain	Sad	Sad
Pleasure / Ecstasy	Surprise	Surprise
Pride	Happy	Happy
Relief	Neutral	Neutral
Sadness	Sad	Sad
Sexual Lust	Neutral	Neutral
Shame	Sad	Disgust
Sourness	Disgust	Disgust
Teasing	Contempt	Contempt
Thankfulness / Gratitude	Happy	Neutral
Triumph	Happy	Surprise