

Correcting Cross-Cultural Bias in VLM Facial Emotion Recognition

Yuina Iseki
Stanford University
yuina@stanford.edu

Charlotte Zhu
Stanford University
xuyang1@stanford.edu

Abstract

Facial emotion recognition (FER) systems based on Vision-Language Models (VLMs) exhibit severe cross-cultural failure modes that are poorly understood and rarely corrected. We identify one such failure in LLaVA-1.5-7B: near-total collapse on neutral expression recognition under zero-shot inference (0.001 accuracy), with 81.5% of neutral faces predicted as sad across all racial groups. We investigate whether LoRA finetuning can correct this failure through a 2×2 ablation varying finetuned components (LLM-only vs. LLM+projector) and training data composition (unbalanced RAF-DB vs. demographically balanced RAF-DB+EAFE), evaluated across Caucasian and East Asian faces on two datasets alongside a second model, Qwen-2.5-VL-7B. LLM-only finetuning largely resolves LLaVA’s neutral collapse, and balanced training improves East Asian performance without sacrificing Caucasian accuracy. In doing so, we uncover several consequential findings: imbalanced finetuning on Western-skewed data actively widens cross-cultural performance gaps and reverses Qwen’s zero-shot cross-cultural advantage; projector finetuning generally underperforms LLM-only adaptation despite more trainable parameters; and greater face-region attention concentration does not predict accuracy and shows no systematic differences across racial groups. Our results suggest that training data composition is a more critical lever than projector finetuning for cross-cultural fairness in VLM-based FER.

1. Introduction

Facial Emotion Recognition (FER) is a fundamental task in computer vision that affects billions of people in areas such as human-computer interaction, affective computing, and clinical monitoring. Although Vision-Language Models (VLMs) grow more capable, models continue to systematically fail for specific racial/cultural groups, posing a fairness question with real-world consequences [2, 8].

An example of such a failure mode is the collapse of neutral expression recognition for LLaVA-1.5-7B [21], a state-

of-the-art(SOTA) VLM: in zero-shot inference on the RAF-DB benchmark, 81.5% of neutral faces are misclassified as sad, across all races. However, other SOTA VLMs such as Qwen-2.5-VL-7B [3] exhibit the complete opposite behavior: it correctly classifies 97% of neutral faces. However, Qwen shows very low accuracy in zero-shot performance for disgusted (0.131) and fearful (0.216). This project compares VLM performance with various fine-tuning configurations on different distributions of racial facial emotion data and evaluates its performance across race and emotion. We hope to uncover reasons for model calibration failure and point to ways of reducing any cross-cultural bias across both emotion categories and racial groups.

In this study we systematically evaluate two SOTA VLMs, LLaVA-1.5-7B and Qwen-2.5-VL-7B, through zero-shot and fine-tuned evaluation across two races (Caucasian and East-Asian) and seven basic emotions. The input to our models are two datasets representing two distinct demographic distributions (RAF-DB [15], predominantly Caucasian; EAFE [17], only East Asian). The output is Grad-CAM attention maps and FER accuracy rates across race and emotions, evaluated for race- and emotion-specific failure modes observed across all fine-tuning configurations.

We investigate whether LoRA finetuning can correct these failures through a 2×2 ablation varying two factors: which components are finetuned (LLM-only vs. LLM+projector) and what training data is used (RAF-DB only vs. demographically balanced RAF-DB+EAFE). We evaluate across two races (Caucasian, East Asian) and seven emotions(happy, sad, neutral, angry, fearful, disgusted, surprised) on two test sets, and use GradCAM region analysis to ask whether face attention concentration predicts accuracy. Specifically, we test four hypotheses:

- H1** Finetuning corrects LLaVA’s neutral collapse (H1a) and improves Qwen’s low-accuracy emotions (H1b).
- H2** Finetuning LLM+projector outperforms LLM-only.
- H3** Balanced training set improves East Asian performance relative to unbalanced training set.
- H4** Greater face-region attention concentration correlates with higher accuracy in emotion prediction.

H1a and H3 are confirmed; H1b, H2, and H4 are refuted,

and we discuss implications and recommendations for future directions.

2. Related Work

2.1. Racial Bias in FER

Despite strong benchmark performance across a range of architectures [16, 14], FER systems exhibit persistent racial bias rooted in how training data is collected and labeled. At the dataset level, Hosseini et al. (2025) analyzed four major FER datasets and found that White samples constitute over 50% of total images, while fear and disgust are simultaneously the least represented emotion categories — concentrating data sparsity precisely at the emotion categories where cross-cultural expression differences are theoretically predicted to be largest [8]. A structurally distinct problem arises from expression type: Khan and Stinson (2025) found that over 35% of RAF-DB images are posed rather than spontaneous, and because posed expressions tend to reflect mouth-centered Western display norms, models trained on these datasets learn a prototype of emotional expression that is systematically misaligned with the lower-intensity expressions characteristic of collectivist cultures [13]. At the labeling stage, Fan et al. (2023) showed that non-Asian annotators introduce systematically skewed labels for Asian faces, and that compositional rebalancing alone yields little fairness improvement when expression prototypes and annotation biases remain [7].

The downstream consequence is captured by Li et al. (2023), who found that FER performance degrades substantially for Eastern faces on negative emotions while remaining stable for happiness — a gap that constitutes a violation of measurement invariance with serious implications for cross-cultural deployment [18]. Together these studies establish racial bias in FER as a structural property of the field. Yet two gaps remain: prior work diagnoses bias only at the output level without examining where it originates inside the model pipeline, and intervention attempts have focused on data rebalancing alone, which Fan et al. themselves show is insufficient in traditional FER systems. Whether this finding extends to VLMs under lightweight adaptation remains untested — a gap our finetuning ablation directly addresses.

2.2. VLMs for Emotion Recognition

VLMs represent a fundamental shift in emotion recognition. Unlike traditional FER classifiers constrained to fixed taxonomies, VLMs process image and text jointly, enabling flexible querying and common-sense contextual reasoning that purely visual models lack. Lian et al. (2024) provided the first large-scale evaluation of this capacity, testing GPT-4V across 21 benchmarks and finding that VLMs are competitive with, and on some tasks superior to, supervised sys-

tems [19]. However, the same evaluation revealed a systematic failure mode: performance degrades sharply on subtle and low-intensity expressions. Wang et al. (2025) find that LLaMA-3.2 and Qwen2.5-VL reliably identify happiness under zero-shot prompting but fail on subtler academic emotional states such as distraction [24]. Agarwal et al. (2026) offer an architectural account of this limitation, arguing that VLMs’ sparse temporal sampling is fundamentally misaligned with the fleeting nature of micro-expressions, introducing systematic blind spots in the perception of low-intensity signals including those characteristic of East Asian expressive norms [1].

Beyond perceptual limitations, VLMs also exhibit measurable racial bias. FOCUS (2025) [5] and GRAS (2025) [22] demonstrate that facial demographic cues systematically shape model inferences even when race is not mentioned in the prompt, and that models such as CLIP misclassify Black individuals at disproportionately high rates.

What existing work does not address is the mechanism behind these disparities, nor whether they are correctable through lightweight adaptation. The studies reviewed here diagnose VLM bias at the output level. Recent work has shown that the vision-language projector is a bottleneck for fine-grained visual feature capture, and that projector-level intervention can recover local visual detail lost in translation to the language model [10]; whether this bottleneck also mediates racial bias in VLM emotion recognition remains untested. We address both gaps: GradCAM analysis of the vision encoder to locate where bias originates, and a projector finetuning ablation to test whether it is correctable at that stage.

3. Data

We work with two main datasets that exhibit distinct demographic distributions. Both share the same seven basic emotion taxonomy (happy, sad, neutral, angry, fearful, disgusted, surprised), originally proposed by Ekman and Friesen [6], enabling direct cross-dataset comparison without label remapping.

RAF-DB [15]. The Real-world Affective Faces Database is a facial expression dataset comprising 15,339 images labeled across seven basic emotion categories. Race labels for RAF-DB were estimated using the FairFace classifier [11] applied to all training and test images as a pre-processing step, as the original dataset does not provide demographic annotations. We use RAF-DB as our primary training and test dataset.

EAFE [17]. The Extended Asian Face Expression dataset provides an independent held-out test set for cross-dataset validation, enabling us to assess whether findings on RAF-DB generalize beyond a single data source.

Table 1 reports the number of images per race $\times$ emotion cell in the RAF-DB and EAFE test sets. The data exhibit

unbalanced representation across both emotion and race: happy and sad are heavily represented while fearful and disgusted are sparse, and East Asian faces are underrepresented relative to Caucasian faces in several emotion categories. We use the full RAF-DB test set rather than restricting to Caucasian faces due to the low count of fearful examples in that subset ($n = 57$). RAF-DB can thus be considered a mixed-race dataset dominated by Caucasian faces, while EAFE serves as a purely East Asian evaluation counterpart.

4. Method Experiments

4.1. Model Architecture

LLaVA-1.5-7B. LLaVA-1.5-7B consists of three components: a CLIP ViT-L/14 vision encoder that produces patch-level visual features, a two-layer MLP projector (7M parameters) that translates vision encoder outputs into the LLM’s input space, and a Vicuna-7B language model that generates the final prediction.

Qwen-2.5-VL-7B. Qwen2.5-VL-7B also consists of three components: a native ViT-based vision encoder with dynamic resolution support that produces patch-level visual features at variable spatial scales, a compression merger layer that pools and projects vision encoder outputs into the language model’s input space, and a Qwen-2.5-7B language model that generates the final prediction. Unlike LLaVA’s fixed-resolution CLIP encoder, Qwen2.5-VL’s vision encoder processes images at their native resolution, producing variable-length token sequences that are compressed by the merger before being passed to the LLM. Visual information enters the language model exclusively through the projector, making it the critical junction between visual representation and language processing. Therefore, we focus our architectural ablation on this component.

4.2. Zero-shot Inference

Both models are evaluated under zero-shot conditions using a structured prompt requesting a single-word emotion label from the list of seven basic emotions (randomized list of angry, disgusted, fearful, happy, neutral, sad, surprised). Emotion labels are presented in randomized order across inference runs to control for order effects. The prompt is:

“What emotion is shown? Answer with one word: (randomized list seven emotions)”

Model outputs are parsed by extracting the first matching emotion word appearing after the assistant delimiter. Responses not matching any label are coded as unknown and excluded from accuracy calculations.

4.3. LoRA Finetuning

We fine-tune both models using Low-Rank Adaptation (LoRA) [9], which inserts trainable low-rank decomposi-

tions into frozen pretrained weight matrices. Specifically, for a weight matrix $W \in \mathbb{R}^{d \times k}$, LoRA learns an additive update $\Delta W = BA$ where $B \in \mathbb{R}^{d \times r}$ and $A \in \mathbb{R}^{r \times k}$ with rank $r \ll \min(d, k)$, keeping W frozen throughout training. This reduces the number of trainable parameters by orders of magnitude relative to full fine-tuning, which is critical for adapting 7B-parameter models on limited FER training data where full fine-tuning would risk overfitting and is computationally prohibitive on a single A100. LoRA’s modular design also allows us to isolate specific architectural components for ablation without destabilizing pretrained representations.

We fine-tune both models under a 2×2 ablation design, varying two factors independently: which layers are fine-tuned (LLM only vs. LLM + projector) and what training data is used (unbalanced RAF-DB vs. balanced RAF-DB + EAFE). This yields 5 total configurations including the zero-shot baseline.

LLM-Only (Conditions 1A, 2A). LoRA adapters are applied to all query and value projection layers (`q_proj`, `v_proj`) across all transformer layers of the language model backbone. The vision encoder and projector/merger are kept frozen. This condition tests whether model calibration failure can be corrected through language-side adaptation alone, without modifying the vision-language interface.

LLM + Projector (Conditions 1B, 2B). LoRA adapters are applied identically to the LLM-only conditions, with the multimodal projector additionally made trainable via full weight updates. These layers are `multimodal_projector.linear_1` and `linear_2` for LLaVA; and `visual.merger.mlp.0` and `visual.merger.mlp.2` for Qwen. The vision encoder remains frozen in both models. This condition tests whether adapting the vision-language interface produces qualitatively different behavior compared to language-side adaptation alone.

Hyperparameters. LoRA hyperparameters are held constant across all conditions.

Qwen additionally uses gradient checkpointing due to higher memory requirements from dynamic resolution inputs. All runs used single NVIDIA A100-80GB GPUs. These values were selected based on original LoRA fine-tuning literature for similar use-cases, and hyperparameter tuning remains a direction for future work [4, 9, 20].

4.4. Training Set

Full RAF-DB (Conditions 1A, 1B). The complete RAF-DB training split of 11,021 images without modification, approximately 76% Caucasian, 16% Asian, and 8% African-American. The combined emotion distribution is: happy 4,268, neutral 2,280, sad 1,783, surprised 1,149, angry 637, disgusted 654, fearful 250.

Balanced RAF-DB + EAFE (Conditions 2A, 2B). We

	Happy	Sad	Neutral	Angry	Fearful	Disgusted	Surprised	Total
<i>RAF-DB test set</i>								
Asian	196	91	126	9	12	27	22	483
Caucasian	888	349	505	143	57	126	283	2351
African American	487	162	331	41	17	49	115	1202
<i>EAFE test set</i>								
East Asian	2050	1409	2578	1519	400	473	1570	9999

Table 1. Number of images per race×emotion cell in the RAF-DB test set and EAFE test set. Cells with $n < 100$ should be interpreted with caution.

Hyperparameter	Value
Optimizer	AdamW
Learning rate	2×10^{-4}
Epochs	2
LR schedule	Constant, no warmup
Precision	bf16
Batch size (LLaVA)	8 per GPU, grad accum 4
Batch size (Qwen)	4 per GPU, grad accum 8
Gradient checkpointing	Qwen only
LoRA rank r	16
LoRA α	32
LoRA dropout	0.05

Table 2. Training hyperparameters

constructed a combined training set to achieve demographic balance between Caucasian and East Asian faces. From RAF-DB, the Caucasian group is subsampled to 6,000 images while all Asian (1,721) and African-American (865) images are retained, yielding 8,308 RAF-DB training images. These are combined with 6,001 East Asian images from the EAFE training split stratified by emotion, for a total of 16,309 training images with exactly equal Caucasian and East Asian representation (7,722 each). The combined emotion distribution is: happy 5,238, neutral 3,675, sad 2,523, surprised 1,993, angry 1,506, disgusted 902, fearful 472.

4.5. GradCAM

We use GradCAM [23] to localize which spatial regions of the vision encoder most influenced each emotion prediction. Unlike accuracy-only evaluation, GradCAM provides a mechanistic lens: it lets us ask not just *what* the model predicts, but *where* it looks when making that prediction — a critical question for a bias study where the hypothesis is that models may attend to culturally non-diagnostic facial regions. Concretely, GradCAM computes a spatial importance map by gradient-weighting the feature map activations: for a target class c , the importance of spatial location (i, j) in feature map k is weighted by $\frac{\partial y_c}{\partial A_{ij}^k}$, producing a map that highlights which patch regions most drove the prediction.

We apply GradCAM at LLaVA’s multimodal projector

rather than at the vision encoder directly. The projector is the vision-language boundary, the sole pathway through which visual features are translated into the LLM’s token space, making it the most interpretable chokepoint for understanding which visual information survives into language processing. Gradients are backpropagated from the predicted emotion token through the language model to this layer, producing 24×24 spatial attention maps over the visual patch grid.

To quantitatively measure where each model directs its visual attention, we used dlib’s 68-point facial landmark predictor [12] to segment face images into four regions: eyes (landmarks 36–47), nose (landmarks 27–35), mouth (landmarks 48–59), and the full face (convex hull of all 68 points). Landmark coordinates were scaled from the original 100×100 image space to the 24×24 GradCAM grid, and a binary mask was constructed for each region using the convex hull of the corresponding landmarks. For each image, we computed the attention ratio—the fraction of total GradCAM activation energy falling within a given region—as:

$$r = \frac{\sum \text{CAM}[\text{mask}]}{\sum \text{CAM}} \quad (1)$$

These per-image ratios were then averaged across all test images within each demographic and emotion group to produce model-level summaries of regional attention distribution.

We evaluate each configuration against the four hypotheses outlined in Introduction, reporting per-emotion accuracy and GradCAM attention ratios across both test sets.

5. Results & Evaluation

Tables 3–6 report per-emotion accuracy across all conditions for both models on RAF-DB and EAFE test sets. ZS denotes zero-shot inference; 1A and 1B denote fine-tuning on unbalanced RAF-DB with LLM-only and LLM+Projector adaptation respectively; 2A denotes fine-tuning on the balanced RAF-DB+EAFE training set with LLM-only adaptation.

5.1. Zero-Shot Baseline

LLaVA zero-shot achieves 63.4% overall accuracy on RAF-DB but collapses near-totally on neutral faces, cor-

Emotion	ZS	1A	1B	2A	2B
angry	0.778	0.802	0.864	0.815	0.833
disgusted	0.544	0.800	0.331	0.756	0.431
fearful	0.189	0.486	0.041	0.581	0.081
happy	0.915	0.964	0.851	0.946	0.840
neutral	0.001	0.868	0.712	0.826	0.532
sad	0.883	0.837	0.837	0.877	0.816
surprised	0.638	0.875	0.878	0.924	0.875
Overall	0.634	0.885	0.775	0.881	0.732

Table 3. LLaVA accuracy on RAF-DB.

Emotion	ZS	1A	1B	2A	2B
angry	0.498	0.508	0.531	0.722	0.532
disgusted	0.307	0.575	0.277	0.730	0.360
fearful	0.062	0.438	0.025	0.425	0.044
happy	0.954	0.979	0.932	0.963	0.913
neutral	0.000	0.759	0.539	0.772	0.366
sad	0.969	0.698	0.706	0.764	0.709
surprised	0.868	0.898	0.987	0.939	0.976
Overall	0.561	0.758	0.679	0.813	0.634

Table 4. LLaVA accuracy on EAFE.

Emotion	ZS	1A	1B	2A	2B
angry	0.784	0.809	0.883	0.660	0.889
disgusted	0.131	0.306	0.081	0.269	0.081
fearful	0.216	0.243	0.203	0.270	0.203
happy	0.428	0.916	0.637	0.936	0.625
neutral	0.969	0.860	0.632	0.846	0.768
sad	0.446	0.674	0.749	0.797	0.686
surprised	0.249	0.796	0.371	0.739	0.207
Overall	0.530	0.799	0.598	0.808	0.597

Table 5. Qwen accuracy on RAF-DB.

Emotion	ZS	1A	1B	2A	2B
angry	0.592	0.476	0.731	0.423	0.758
disgusted	0.057	0.118	0.015	0.132	0.026
fearful	0.152	0.075	0.130	0.069	0.194
happy	0.498	0.933	0.766	0.962	0.796
neutral	0.894	0.404	0.483	0.629	0.561
sad	0.586	0.633	0.799	0.730	0.803
surprised	0.781	0.987	0.846	0.960	0.756
Overall	0.636	0.621	0.644	0.687	0.664

Table 6. Qwen accuracy on EAFE.

rectly classifying just 1 of 680 neutral images (recall 0.001). Of the misclassified neutral faces, 81.5% are predicted as sad. This collapse replicates on EAFE (recall 0.000), confirming that the failure is a systematic calibration artifact of LLaVA rather than a dataset-specific anomaly, and that it is not corrected by cross-dataset exposure alone. Qwen zero-shot, however, exhibits the opposite pattern. Qwen achieves 0.969 neutral recall on RAF-DB and 0.894 on EAFE, which suggests that the failure is specific to LLaVA’s architecture or pretraining rather than a universal VLM property. Notably, Qwen zero-shot performs better on EAFE (0.636) than RAF-DB (0.530), which is the reverse of LLaVA’s pat-

tern, suggesting that perhaps its pretraining distribution included more East Asian content.

5.2. Finetuning Results

We report four key findings from the finetuning results on the unbalanced RAF-DB dataset and the balanced RAF-DB + EAFE dataset.

H1: LoRA Finetuning corrected Neutral collapse in LLaVA but had minimal impact on disgusted and fearful failure in Qwen. Condition 1A (LLM-only, RAF-only) largely resolves LLaVA’s neutral collapse, raising recall from 0.001 to 0.868 on RAF-DB and from 0.000 to 0.759 on EAFE, which confirms that the failure is a correctable calibration problem rather than a fundamental architectural limitation. This confirms **H1a**. However, despite a significant increase from 0.131 for disgusted in Qwen zeroshot to 0.306 in 1A condition, none of the accuracy in Qwen exceeded 0.4 for disgusted and fearful. This is probably due to the lack of training images in disgusted and fearful because they have the lowest image count across both RAF train set and the curated balanced training set. We therefore conclude that the improvement is insufficient to support **H1b**.

H2: Finetuning LLM-only outperforms LLM+Projector. For LLaVA, condition 1B, 2B generally underperforms 1A, 2A in overall accuracy across both datasets. Exceptions exist for two emotions: angry on RAF-DB, where 1B outperforms 1A (0.864 vs. 0.802), and surprised on EAFE, where 1B substantially outperforms 1A (0.987 vs. 0.898). Both suggest that projector adaptation may provide localized benefit for specific emotions — though the pattern does not generalize across emotions or datasets, making it difficult to draw strong conclusions without further targeted experiments.

For Qwen, we see a similar pattern when the finetuned models are tested on RAF-DB. 1B and 2B yield about 0.6 accuracy which is significant lower than the around 0.8 overall accuracy for 1A and 2A. For Qwen evaluated on EAFE, however, we observe marginally better overall performance of 1B relative to 1A, and slightly better performance for 2A compared to 2B. All overall accuracy for Qwen evaluated against EAFE is between 0.6-0.7 so the difference is relatively minimal. The results shown overall refutes **H2**.

A notable failure mode emerges from LLM+Projector finetuning (conditions 1B, 2B): LLaVA 1B drops fearful accuracy to 0.041 on RAF-DB, substantially below even zero-shot performance (0.189), suggesting that projector adaptation can actively destabilize recognition of low-frequency emotion categories. This instability may reflect the projector’s sensitivity to distribution shift when trained on sparse emotion classes.

H3: Balanced training improves East Asian performance. We measure the cross-cultural gap as the dif-

ference in overall accuracy between RAF-DB and EAFE for the same condition (RAF-DB – EAFE); a larger gap indicates greater disparity in performance across populations. For LLaVA, balanced fine-tuning (2A) improves EAFE accuracy substantially relative to imbalanced fine-tuning (1A): 0.813 vs. 0.758, a gain of +0.055 overall, with pronounced per-emotion gains (angry +0.214, disgusted +0.155). Crucially, this improvement comes without sacrificing RAF-DB performance (0.881 vs. 0.885), confirming H3 for LLaVA. For Qwen, 2A similarly improves EAFE relative to 1A (0.687 vs. 0.621, +0.066), though balanced training is less effective at closing the disparity: a cross-cultural gap of 0.121 remains for Qwen compared to 0.068 for LLaVA, suggesting a partial confirmation of H3.

Imbalanced fine-tuning amplifies and can reverse cross-cultural gaps. Beyond confirming H3, the results reveal a more concerning pattern: imbalanced fine-tuning actively widens the cross-cultural gap relative to zero-shot. For LLaVA, the gap grows from 0.073 (ZS) to 0.127 (1A), with per-emotion degradation visible in sad recognition on EAFE (0.969 to 0.698). For Qwen, the effect is more dramatic: at zero-shot, Qwen holds a cross-cultural *advantage* favoring EAFE (−0.106), but imbalanced fine-tuning eliminates this entirely, flipping the gap to +0.178. This suggests that standard fine-tuning on Western-skewed datasets does not produce culturally neutral improvements — it encodes the demographic skew of the training data into the model’s predictions.

5.3. GradCAM Results

H4: No clear correlation is observed between face-region attention concentration and higher accuracy in emotion prediction. Figure 1 shows mean GradCAM maps across the five conditions. Fine-tuning consistently shifts attention toward the face region: the face attention ratio increases from 0.625 under zero-shot to above 0.75 for all fine-tuned conditions (Table 7), with attention on bounded subregions rising correspondingly. However, these results do not support H4: greater face attention concentration does not predict better accuracy. Condition 1B achieves the highest face attention ratio (0.845) yet ranks fourth in overall accuracy (0.775), while condition 1A achieves the best accuracy (0.885) with a substantially lower ratio (0.753).

Figure 2 reinforces this: 1A performs best for neutral expressions overall, and 2A performs best on neutral faces in EAFE, yet neither achieves the highest face attention ratio (0.710 for 1A, compared to 0.841 for 2B and 0.829 for 2A). Taken together, these results suggest that greater attention concentration on the face is not sufficient and could even be counterproductive for improving emotion recognition accuracy. We tentatively speculate that overly concentrated spatial attention may discard contextual cues that contribute to accurate prediction, though this remains unverified and war-

rants direct investigation in future work.

We also find minimal differences in regional attention ratios across racial groups as shown in Figure 2 and in Table 8. This does not support the hypothesis that display rule theory drives fundamentally different attention patterns for Asian versus Caucasian faces; the models appear to apply similar spatial attention strategies regardless of the subject’s race.

6. Limitations & Future Work

Demographic scope. Our evaluation is limited to two main demographic groups: RAF-DB is mostly Caucasian and EAFE is only East Asian. Our analysis does not cover African-American, South Asian, Latino, or other racial groups. Extending evaluation to broader demographics requires benchmark datasets with the same ground-truth emotion labels, which remain scarce for non-Western populations. The next step would be to start by constructing such datasets.

Dataset distribution mismatch. RAF-DB images are web-scraped from real-world sources while EAFE images are sourced from Chinese, Japanese, and Korean film and web-scraping [?]. The underlying sampling distributions differ in ways that are difficult to characterize, and factors such as lighting conditions, image quality, and expression intensity may all vary systematically between the two sources, independently of any racial or cultural difference. Cross-dataset accuracy gaps therefore reflect a combination of demographic generalization as well as distribution shift, and the two cannot be fully disentangled in our current evaluation.

Emotion class imbalance. Both RAF-DB and our balanced training set contain very few negative emotions, especially fearful and disgusted images, relative to other emotion categories. The consistently low accuracy on these emotions across all conditions makes it difficult to disentangle model limitations from data limitations. Future work could include a training set that is balanced across both race and emotion, with equal representation of fearful and disgusted across all demographic groups. This design could isolate the contribution of each variable in the observed accuracy patterns.

Hyperparameter tuning. LoRA hyperparameters were selected based on prior literature rather than systematic search. Systematic tuning over rank, learning rate, and training epochs could improve accuracy in fine-tuning approaches and clarify whether results are due to an inherent structural property of the architecture or due to suboptimal hyperparameter settings.

Projector finetuning configuration. Finally, we were surprised to see conditions 1B and 2B (LLM+Projector) fine-tuning underperforming compared to LLM-only fine-tuning in overall accuracy, despite having more trainable

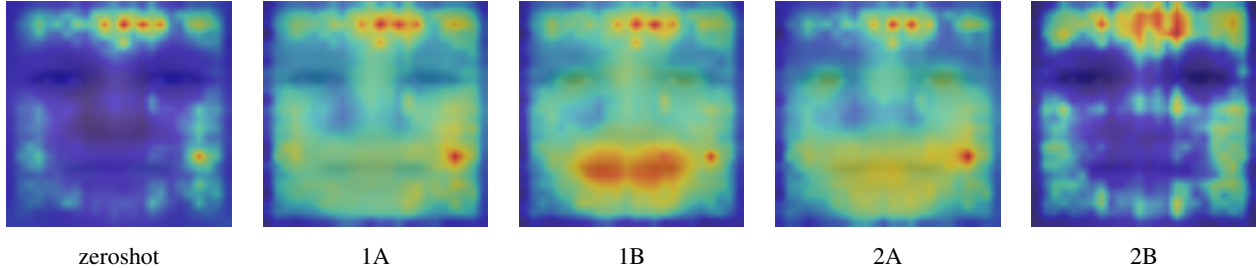


Figure 1. Mean GradCAM attention on LLaVA’s projector across all test images for each model variant on RAF-DB Test Set ($n=3068$).

Region	Zero-shot	1A	1B	2A	2B
Face	0.625	0.753	0.845	0.789	0.774
Eyes	0.039	0.059	0.093	0.070	0.045
Nose	0.025	0.051	0.063	0.054	0.059
Mouth	0.058	0.091	0.186	0.124	0.104

Table 7. Mean GradCAM attention ratios by face region (overall).

parameters. Future work could explore alternative architectural modifications to the projector itself. We believe it may be possible that the projector requires more training data, more epochs, or a different learning parameter to converge stably, given that it serves as the sole vision-language interface. Hence, we hypothesize that it may be more sensitive to distribution shift than the LLM attention layers.

7. Conclusion

This work identifies a severe calibration failure in LLaVA-1.5-7B: near-total collapse on neutral expression recognition, correctable through LLM-only LoRA finetuning. More consequentially, we find that imbalanced finetuning on Western-skewed data actively widens cross-cultural performance gaps and reverses Qwen-2.5-VL-7B’s inherent zero-shot cross-cultural advantage, while balanced training improves East Asian accuracy without sacrificing Caucasian performance. Projector finetuning generally underperforms LLM-only adaptation, and GradCAM reveals no clear relationship between face-region attention concentration and accuracy. Together, these results suggest that training data composition is a more critical lever than projector finetuning for cross-cultural fairness in VLM-based FER, and that standard finetuning on demographically skewed data can silently introduce or amplify cross-cultural bias.

Contributions & Acknowledgements

Both authors contributed equally to this project. Yuina Iseki led finetuning and inference for LLaVA and Qwen across zero-shot and all four finetuned conditions. Charlotte Zhu led the literature review, data analysis, and GradCAM setup. Both authors contributed to experimental design, analysis, and writing. A third teammate, America George, identified the two datasets used in this project and

contributed to the milestone 2 presentation slides, but did not participate in subsequent research, implementation, or writing.

Claude and ChatGPT were used for consulting on implementation strategies, debugging, proofreading, and grammar correction. All project ideas, experimental results, data analysis, and writing are the authors’ own work.

References

- [1] M. Agarwal, S. A. Tsafaris, L. Sevilla-Lara, and S. McDonagh. Why do vision language models struggle to recognize human emotions? *arXiv preprint arXiv:2604.15280*, 2026. [2](#)
- [2] N. AlDahoul, H. Ibrahim, M. Park, T. Rahwan, and Y. Zaki. Inclusive content reduces racial and gender biases, yet non-inclusive content dominates popular culture. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024. [1](#)
- [3] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang, H. Zhong, Y. Zhu, M. Yang, Z. Li, J. Wan, P. Wang, W. Ding, Z. Fu, Y. Xu, J. Ye, X. Zhang, T. Xie, Z. Cheng, H. Zhang, Z. Yang, H. Xu, and J. Lin. Qwen2.5-vl technical report, 2025. [1](#)
- [4] D. Biderman, J. Portes, J. J. G. Ortiz, M. Paul, P. Green-gard, C. Jennings, D. King, S. Havens, V. Chiley, J. Frankle, C. Blakeney, and J. P. Cunningham. Lora learns less and forgets less. *Transactions on Machine Learning Research*, 2024. [3](#)
- [5] H. Chen, Q. Huang, J. Zhao, Q. Jiang, X. Chang, and J. Yu. Measuring social bias in vision-language models with face-only counterfactuals from real photos. *arXiv preprint arXiv:2601.06931*, 2026. [2](#)
- [6] P. Ekman, W. V. Friesen, et al. *The repertoire of nonverbal behavior: Categories, origins, usage and coding*, volume 1. Mouton de Gruyter Berlin, 1969. [2](#)

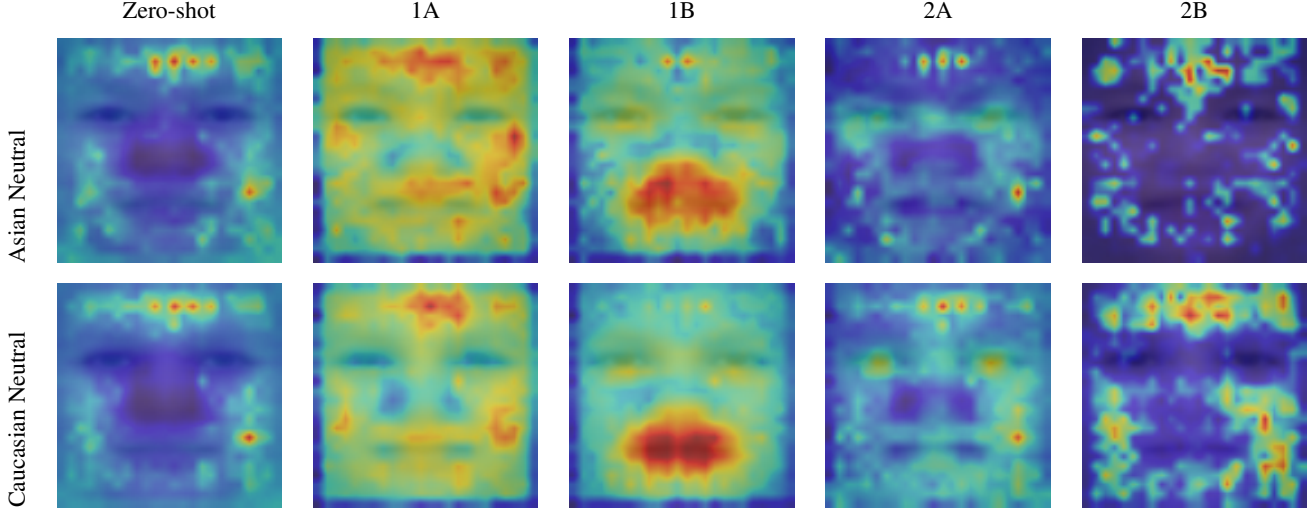


Figure 2. Mean GradCAM attention on LLAVA’s projector for neutral faces by race across all model conditions on RAF-DB Test Set. Asian Neutral ($n=126$), Caucasian Neutral ($n=505$).

Condition	Region	Zero-shot	1A	1B	2A	2B
Asian, Neutral	Face	0.617	0.705	0.840	0.726	0.884
	Eyes	0.045	0.055	0.072	0.074	0.034
	Nose	0.022	0.048	0.081	0.047	0.098
	Mouth	0.038	0.054	0.117	0.051	0.026
Caucasian, Neutral	Face	0.619	0.708	0.824	0.708	0.834
	Eyes	0.043	0.056	0.081	0.074	0.054
	Nose	0.021	0.046	0.072	0.043	0.075
	Mouth	0.040	0.054	0.125	0.052	0.050
All, Neutral	Face	0.619	0.710	0.829	0.710	0.841
	Eyes	0.044	0.056	0.078	0.074	0.048
	Nose	0.021	0.047	0.075	0.044	0.078
	Mouth	0.040	0.055	0.124	0.053	0.049

Table 8. Mean GradCAM attention ratios by face region for neutral expression (Asian, Caucasian, and all subsets).

- [7] A. Fan, X. Xiao, and P. Washington. Addressing racial bias in facial emotion recognition. *arXiv preprint arXiv:2308.04674*, 2023. 2
- [8] M. M. Hosseini, A. P. Fard, and M. H. Mahoor. Faces of fairness: Examining bias in facial expression recognition datasets and models. *arXiv preprint arXiv:2502.11049*, 2025. 1, 2
- [9] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. Lora: Low-rank adaptation of large language models. *International Conference on Learning Representations*, 2022. 3
- [10] P. Jiao, B. Zhu, J. Chen, C.-W. Ngo, and Y.-G. Jiang. From holistic to localized: local enhanced adapters for efficient visual instruction fine-tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2728–2737, 2025. 2
- [11] K. Kärkkäinen and J. Joo. Fairface: Face attribute dataset for balanced race, gender, and age. *arXiv preprint arXiv:1908.04913*, 2019. 2
- [12] V. Kazemi and J. Sullivan. One millisecond face alignment with an ensemble of regression trees. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1867–1874, 2014. 4
- [13] R. Khan and C. Stinson. Auditing facial emotion recognition datasets for posed expressions and racial bias. *arXiv preprint arXiv:2507.10755*, 2025. 2
- [14] D. Kollias, P. Tzirakis, A. Cowen, S. Zafeiriou, I. Kotsia, A. Baird, C. Gagne, C. Shao, and G. Hu. The 6th affective behavior analysis in-the-wild (abaw) competition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4587–4598, 2024. 2
- [15] S. Li and W. Deng. Blended emotion in-the-wild: Multi-label facial expression recognition using crowdsourced annotations and deep locality feature learning. *International Journal of Computer Vision*, 2019. 1, 2
- [16] S. Li and W. Deng. Deep facial expression recognition: A survey. *IEEE transactions on affective computing*, 13(3):1195–1215, 2020. 2
- [17] S. Li, L. Guo, and J. Liu. Towards east asian facial expression recognition in the real world: A new database and deep

- recognition baseline. *Sensors*, 22(21):8089, 2022. 1, 2
- [18] Y.-T. Li, S.-L. Yeh, and T.-R. Huang. The cross-race effect in automatic facial expression recognition violates measurement invariance. *Frontiers in Psychology*, 14:1201145, 2023. 2
 - [19] Z. Lian, L. Sun, H. Sun, K. Chen, Z. Wen, H. Gu, B. Liu, and J. Tao. Gpt-4v with emotion: A zero-shot benchmark for generalized emotion recognition. *Information Fusion*, 108:102367, 2024. 2
 - [20] H. Liu, C. Li, Y. Li, and Y. J. Lee. Improved baselines with visual instruction tuning. *NeurIPS*, 2023. 3
 - [21] H. Liu, C. Li, Y. Li, and Y. J. Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26296–26306, 2024. 1
 - [22] S. Malik, H. M. Abdullah, S. Saha, and A. P. Sheth. Ask me again differently: Gras for measuring bias in vision language models on gender, race, age, and skin tone. In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 2327–2388, 2026. 2
 - [23] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 4
 - [24] D. Wang, C. Yang, and G. Chen. Using vision language models to detect students’ academic emotion through facial expressions. *arXiv preprint arXiv:2506.10334*, 2025. 2