Modeling Cultural Bias in Facial Expression Recognition with Adaptive Agents

Authors: David Freire-Obregón, José Salas-Cáceres, Javier Lorenzo-Navarro, Oliverio J. Santana, Daniel Hernández-Sosa, Modesto Castrillón-Santana

Affiliation: SIANI, ULPGC, Spain

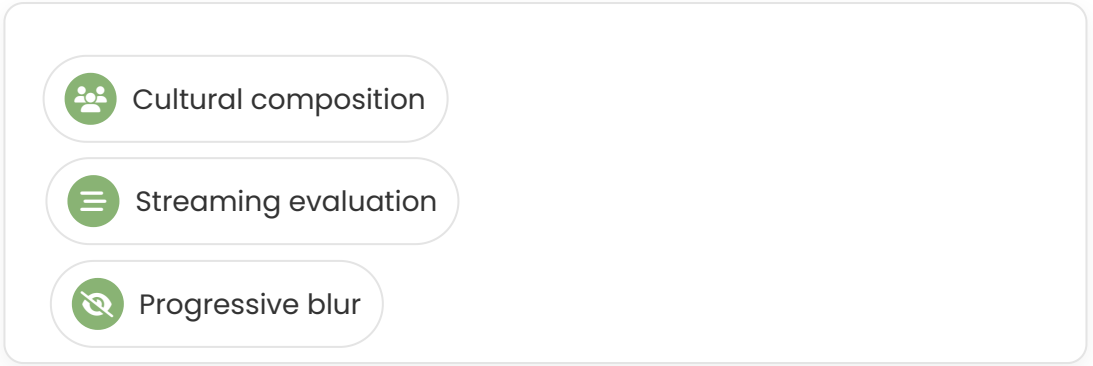
The International Symposium on Agentic Artificial Intelligence Systems AAIS 2025 - Part of FLLM 2025

Outline

Problem Statement

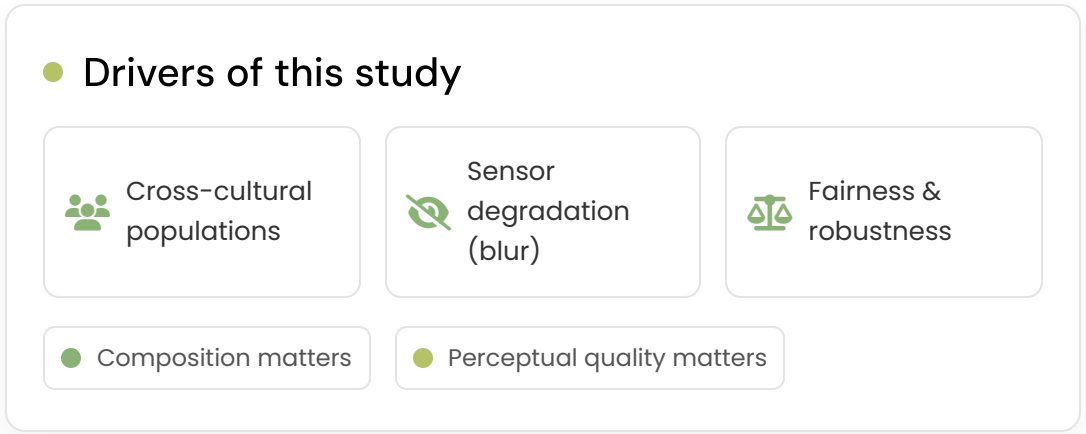
- Facial expression recognition must remain robust under cultural variation and degraded visual conditions.
- Most evaluations assume homogeneous data and high-quality imagery, overlooking real-world diversity.
- Gap: Lack of streaming benchmarks that jointly probe cultural composition and progressive blur.

Robustness Factors



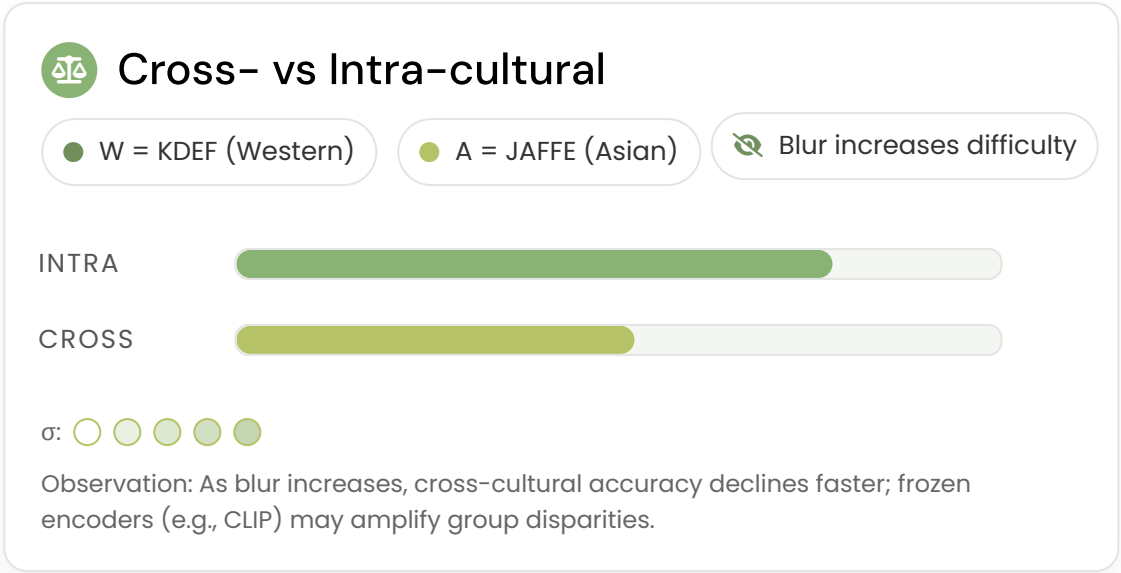
Research Motivation

- Real-world deployments face cross-cultural populations and sensor degradation (e.g., blur).
- Fairness and robustness: average scores can mask group-level disparities.
- Need for a dynamic evaluation capturing how population composition and perceptual quality interact over time.



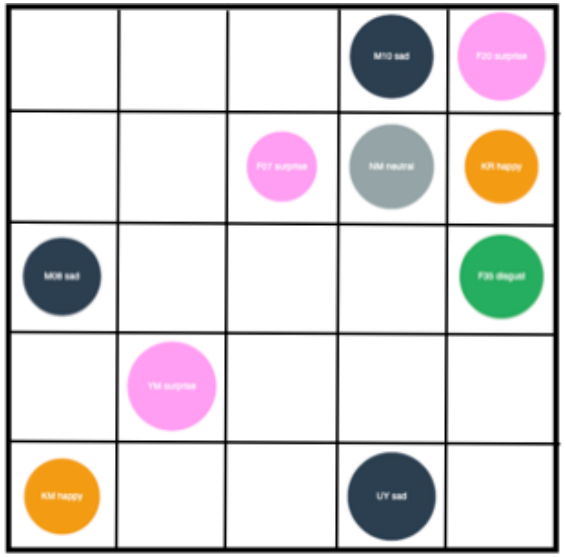
FER Cultural Bias

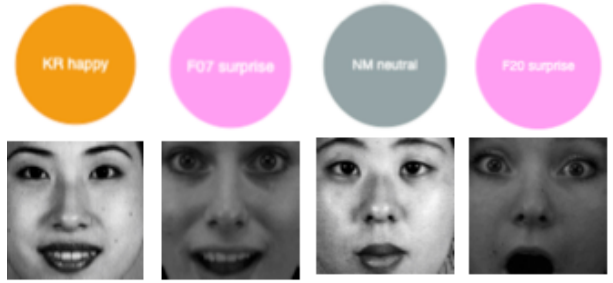
- Only 7 basic emotions are considered.
- Cross-cultural recognition is systematically harder than intra-cultural.
- Representation bias in frozen feature encoders can favor Western faces.
- Progressive blur exacerbates cross-cultural gaps in mixed and imbalanced populations.



Agent-Based Benchmark

Dynamic evaluation of cultural composition × blur





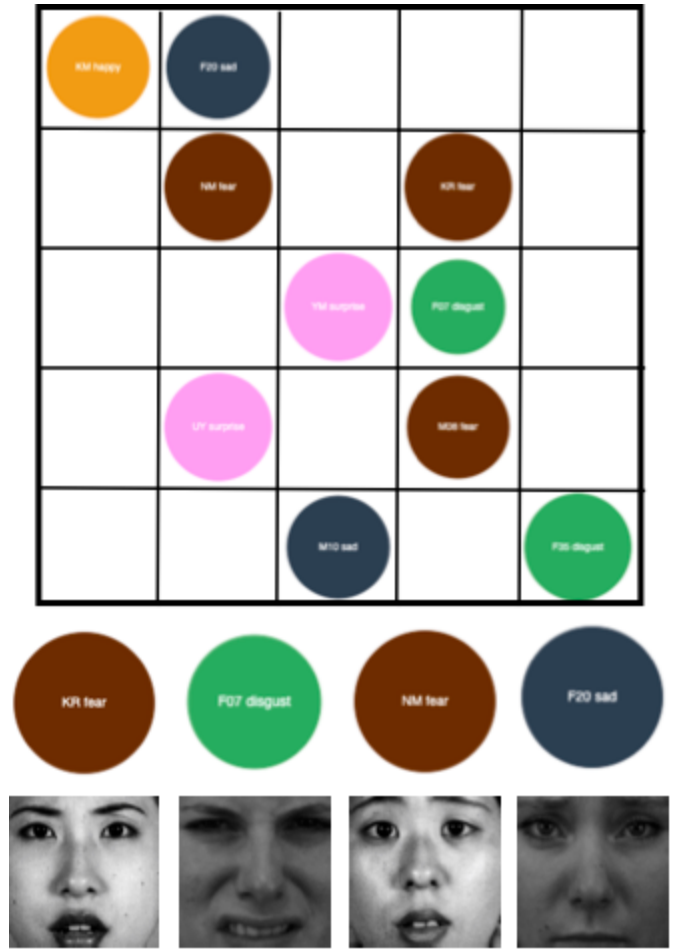
- Agents represent individuals from distinct cultures: KDEF (Western) and JAFFE (Asian).
- Streaming protocol: at each tick, agents display an expression; neighbors perceive and classify.
- Progressive Gaussian blur increases over blocks ($\sigma = 0 \rightarrow 4$), stressing perception.
- Measures robustness as a function of cultural mix and degradation over time.

System Architecture

Agents interact locally while the grid wraps around at edges (toroidal).

- Spatial setup: 25 agents on a 5×5 lattice with wrap-around boundaries.
- Neighbor interaction: agents classify neighbors' displayed expressions; confidence is visualized as outer rings.
- Dynamics: optional movement by local valence and peer-learning during the clean training phase.

KDEF agent (Western) JAFFE agent (Asian) Bigger ring (prediction strength)

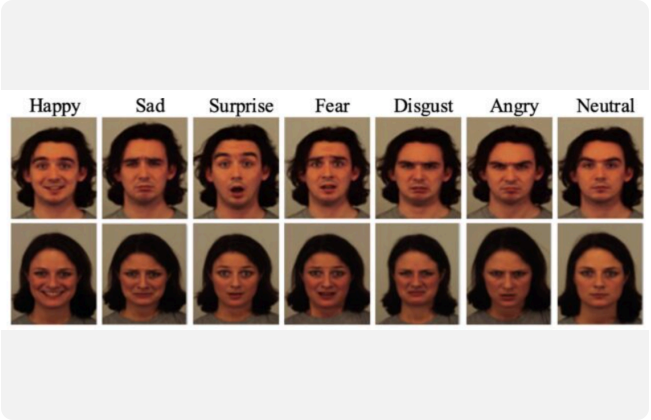


Datasets

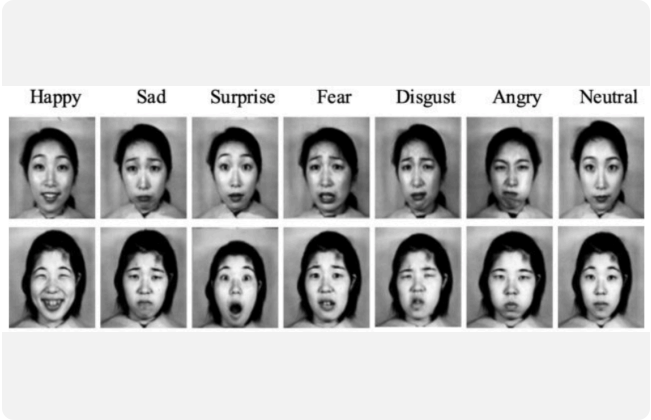
KDEF: 70 Western actors; 7 emotions; controlled captures.

JAFFE: 10 Japanese female subjects; 7 emotions; controlled captures.

Role in this work: define agents' cultural identity and enable intra- vs. cross-cultural evaluation within the streaming benchmark.



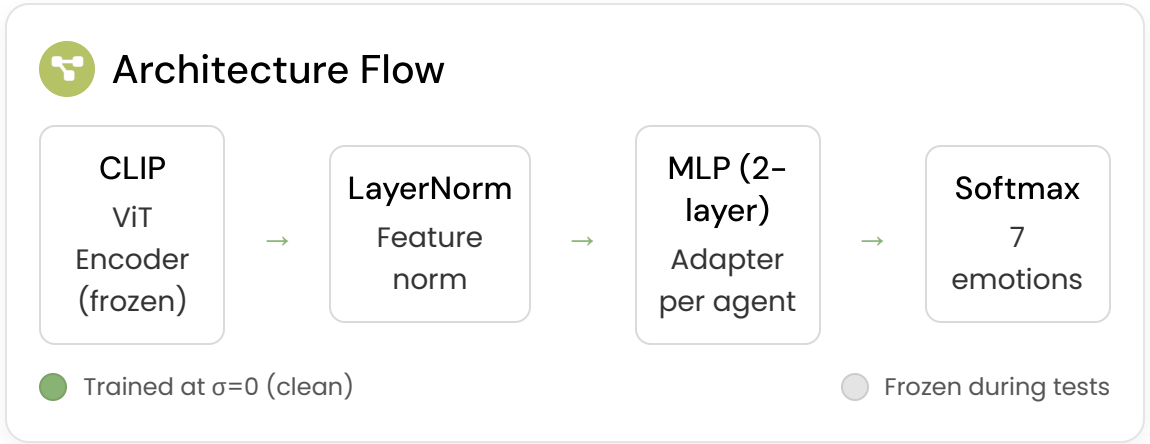
KDEF (WESTERN)
Actors: 70 • Emotions: 7 • Controlled studio.



JAFFE (JAPANESE FEMALE)
Subjects: 10 • Emotions: 7 • Controlled studio.

Agent Design

- Frozen CLIP embeddings as vision backbone (shared feature space).
- Per-agent 2-layer MLP adapter; trained at σ = 0 and frozen for tests.
- Training: cross-entropy with label smoothing; optimizer AdamW.
- Optional peer-learning: learn from high-confidence neighbors.

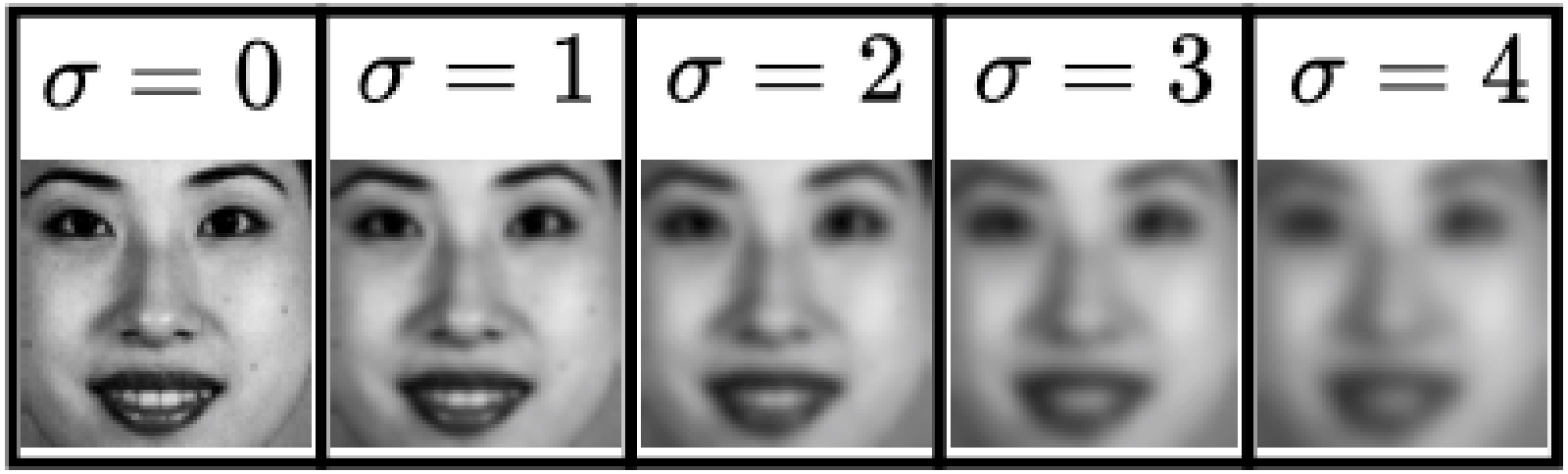


Protocol: train at σ=0 → evaluate as blur increases (σ=0–4)

Progressive Blur Degradation (σ = 0 → 4)

Clean learning at σ = 0, followed by evaluation blocks with increasing Gaussian blur to probe robustness as perceptual quality worsens.

Visual example of the same face under progressively stronger blur levels used in the benchmark protocol:

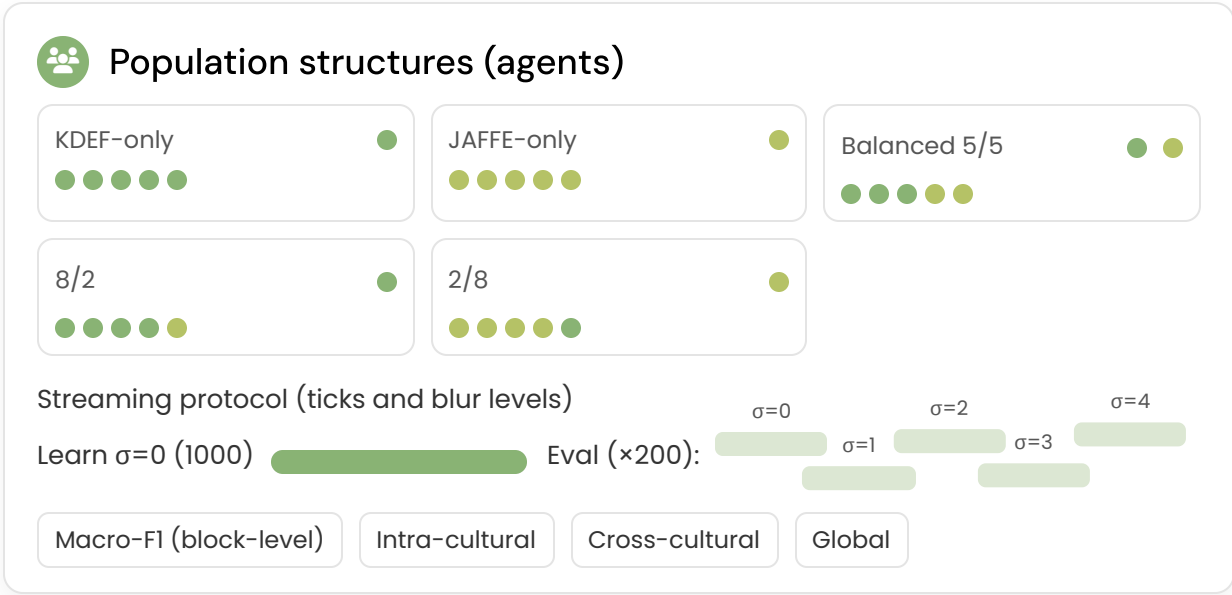


• Training phase: agents learn on clean images (σ = 0) before testing.

• Testing phase: performance is measured block-wise as blur increases (σ = 0 → 4).

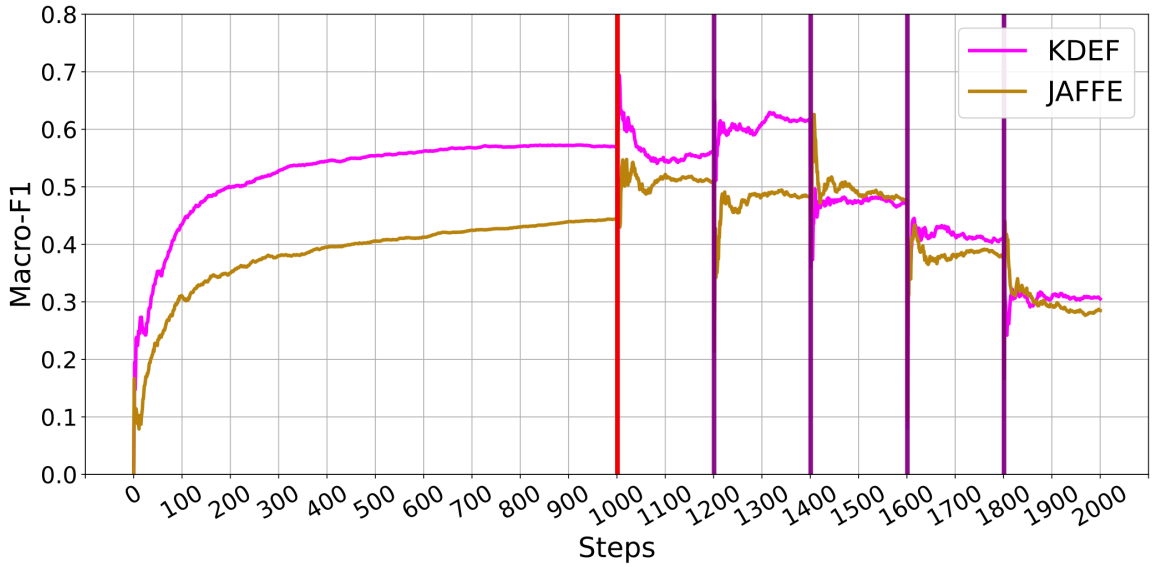
Experimental Setup

- Timeline: 1000 ticks learning at σ = 0; then evaluation blocks of 200 ticks across σ = 0, 1, 2, 3, 4.
- Metrics: Macro-F1 (block-level) for intra-cultural, cross-cultural, and global comparisons.



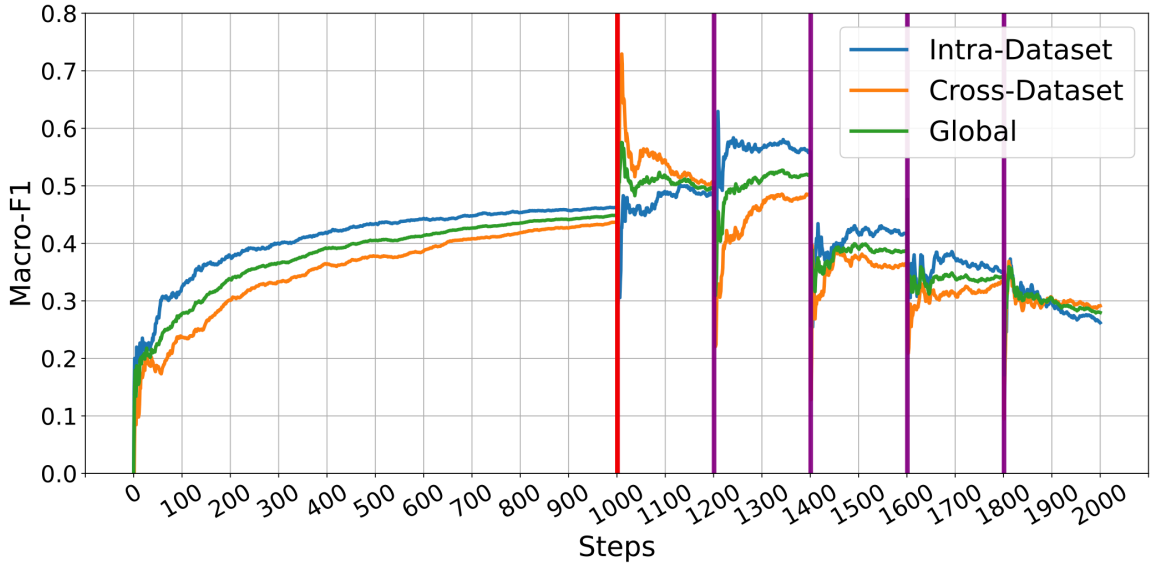
Results: Monocultural Populations

- Clean data (σ=0): KDEF converges faster and higher (~0.70) vs JAFFE (~0.58).
- Under blur: JAFFE holds at low blur (σ≈1) but drops sharply at intermediate σ; KDEF degrades more uniformly.
- Implication: asymmetric robustness across cultural groups.



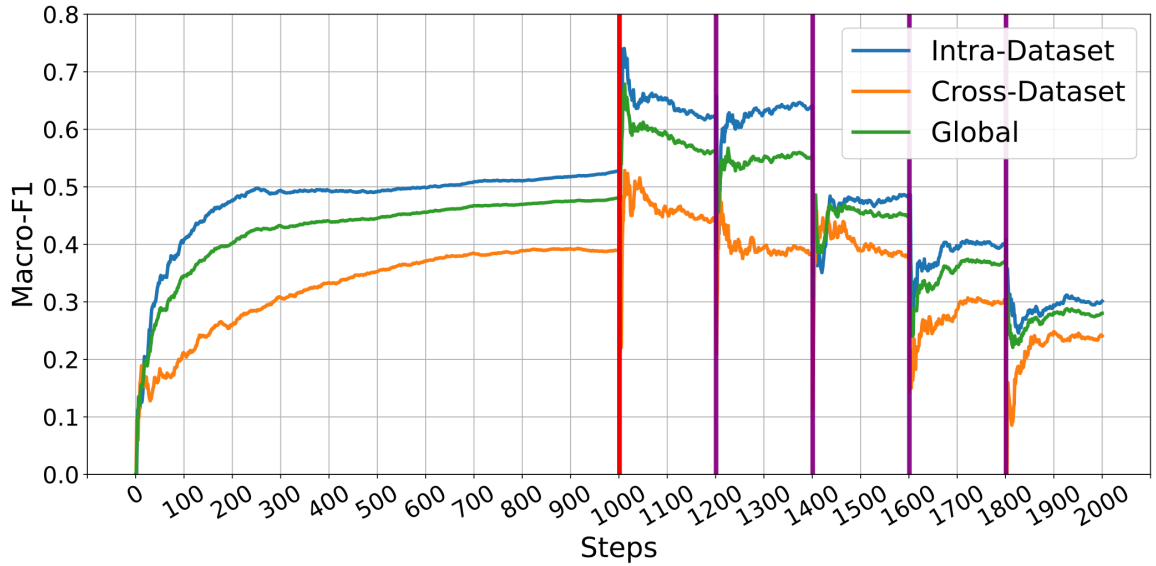
Results: Balanced Mixed Populations

- Intra-cultural accuracy is higher than cross-cultural even on clean images (~0.70–0.75 vs ~0.50).
- Increasing blur widens the intra vs cross-cultural gap across σ levels.
- Balanced mixtures mitigate early-stage degradation compared to monocultures.



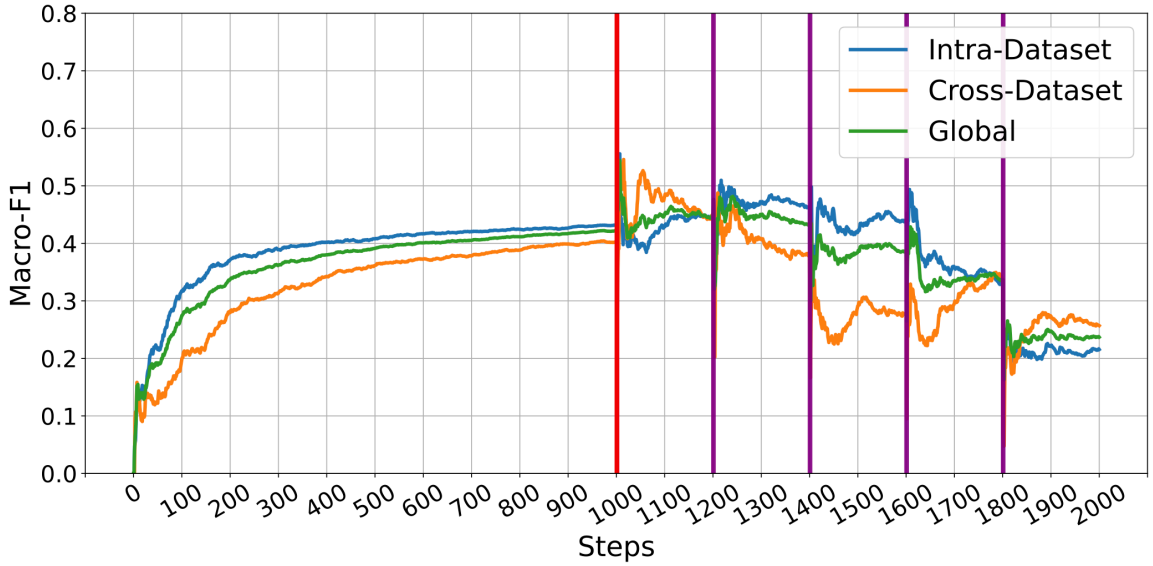
Results: Imbalanced Populations

- Global averages can hide minority disadvantages in mixed settings.
- KDEF-majority: strong intra-cultural stability; cross-cultural performance weakens at high blur.
- Takeaway: composition and blur jointly shape robustness and fairness.



Results: Imbalanced Populations

- Global averages can hide minority disadvantages in mixed settings.
- JAFFE-majority: lower baseline and earlier degradation, consistent with embedding bias effects.



Cultural Bias Analysis

- Asymmetric robustness: KDEF (Western) degrades more uniformly; JAFFE (Asian) shows sharper drops at mid blur.
- Frozen CLIP embeddings exacerbate bias—trained on Western-centric data, benefiting Western facial appearance.
- Cross-cultural recognition is consistently weaker than intra-cultural and worsens as perceptual quality deteriorates.

Performance Summary

- All scenarios show Macro-F1 degradation as blur increases.
- Balanced mixtures mitigate early-stage degradation; imbalanced mixes amplify group disparities at high blur.
- Key takeaway: cultural composition and interaction structure critically shape robustness under worsening conditions.

Δσ by population setting

Scenario	Blur notes
Monocultural — KDEF	uniform degradation
Monocultural — JAFFE	sharper mid-blur drop
Balanced mixed (5/5)	mitigates early loss
KDEF-majority (8/2)	largest high-blur drop
JAFFE-majority (2/8)	early degradation

Conclusions & Future Work

CONTRIBUTIONS

- A agent-based streaming FER benchmark combining cultural composition and progressive blur.
- Quantifies asymmetric degradation and persistent cross-cultural gaps under worsening perceptual conditions.
- Reveals the impact of frozen embeddings (e.g., CLIP) on cultural robustness and fairness.

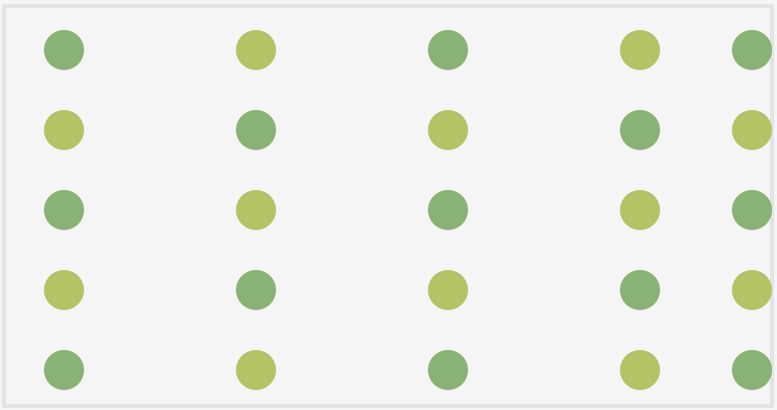
RECOMMENDATIONS / FUTURE WORK

- Pursue balanced/neutral feature encoders.
- Apply dynamic debiasing strategies.
- Expand to more cultures and interaction topologies.
- Evaluate additional degradation types (e.g., noise, compression).
- Explore online/adaptive encoders within the stream.

Thank you / Q&A

Paper: Modeling Cultural Bias in Facial Expression Recognition with Adaptive Agents

Presenter: Freire-Obregón et al., SIANI, ULPGC, Spain



Agent-based benchmark context for FER cultural bias