CASSANDRA : Few-Shot Multimodal Learning via CAuSality-Based DiSentanglement AND CRoss-Modal FlAttening

Utkarsh Byahut
Mentor: Prof. Huan Liu
Ira A. Fulton Schools of Engineering

The Causal Bottleneck in Multimodal Learning

The Spurious Correlation Trap

Current multimodal models are "lazy." When trained on limited data (Few-Shot), they converge on low-level statistical shortcuts rather than semantic truth. For example, a model classifies an object based on background texture (water) rather than the object's morphology (the bird). This results in high training accuracy but catastrophic failure in out-of-distribution (OOD) real-world environments.

The Objective: Causal Invariance

CASSANDRA moves beyond simple pattern matching. By treating data as a product of disentangled causal factors, we force the architecture to isolate:

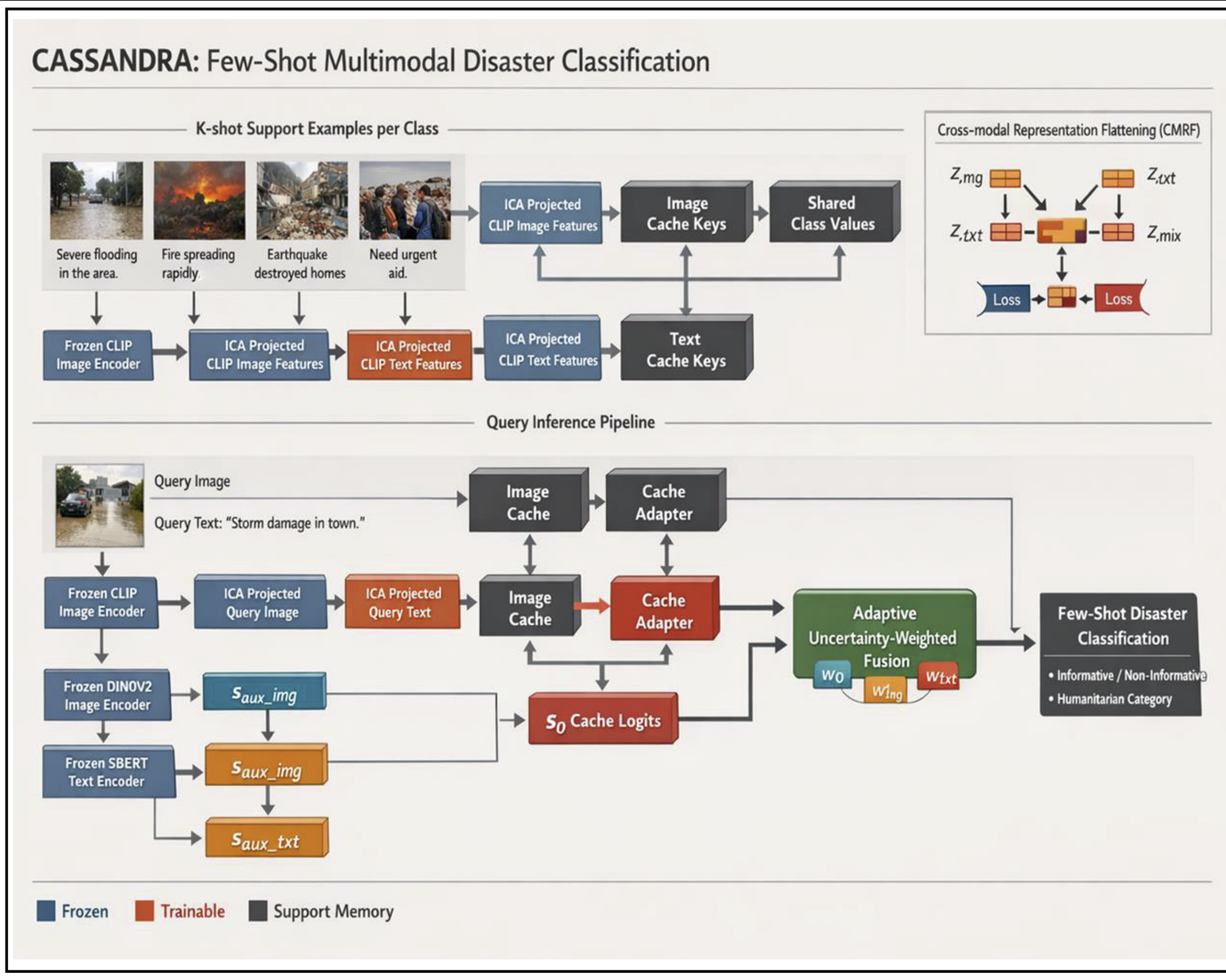
- **Semantic Content (S):** Stable features invariant across domains.
- **Environmental Noise (N):** Transient style factors that pollute the signal.

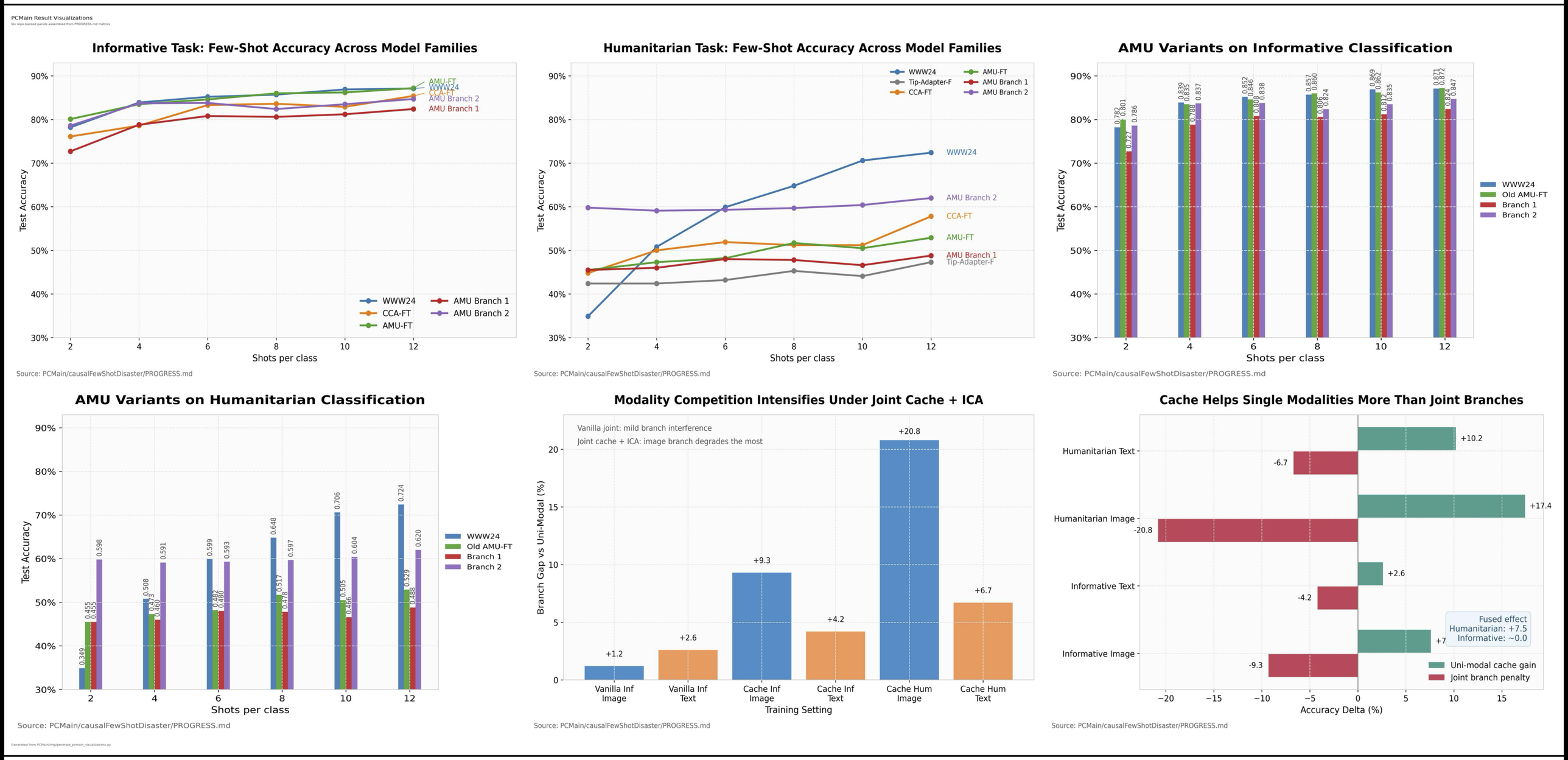
Cross-Modal Flattening

Standard multimodal alignment often suffers from "modality gap," where image and text embeddings remain in isolated clusters.

The Mechanism: We implement a Cross-Modal Flattening layer that projects $\mathbf{Z_c}$ from disparate manifolds onto a shared, low-rank subspace.

The Result: By "flattening" the distance between visual and textual causal features, the model achieves semantic parity. This allows the system to generalize a class concept from a single text description to a visual instance (and vice versa) without the need for dense co-occurrence data.





The CASSANDRA Framework

Structural Disentanglement

CASSANDRA architecture explicitly partitions the latent space to resolve the confounding influence of modality-specific noise. We decompose the input S into a dual-stream representation:

Invariant Causal Features (Z_c): High-level semantic properties (geometry, texture, anatomical markers) that share a direct causal link to the ground-truth label Y.

Variant Style Latents (Z_s): Nuisance variables like lighting, sensor noise, or linguistic syntax that vary across modalities but hold zero predictive value for OOD generalization.

Method

We benchmark few-shot disaster classifiers across informative and humanitarian tasks, from frozen CLIP baselines to AMU-style multimodal variants. Our main model builds class caches from CLIP support features, adds complementary DINOv2 image and SBERT text branches, and fuses them with uncertainty-aware weighting.

Results

On the informative task, the best methods improve from roughly 78 to 80% at 2 shots to about 87% at 12 shots, with WWW24 and AMU-FT leading overall. On the humanitarian task, AMU Branch 2 is strongest in the extreme low-shot regime at 2 shots at about 60%, but WWW24 scales better and finishes highest at about 72% by 12 shots.

Conclusion

The main takeaway is that complementary image-text information matters most when labels are scarce, especially for humanitarian categories. However, cache-based adaptation helps single modalities more than joint branches, revealing modality competition as the core bottleneck for further gains.

Arizona State University

FURI

PMML

Ira A. Fulton Schools of Engineering
Arizona State University