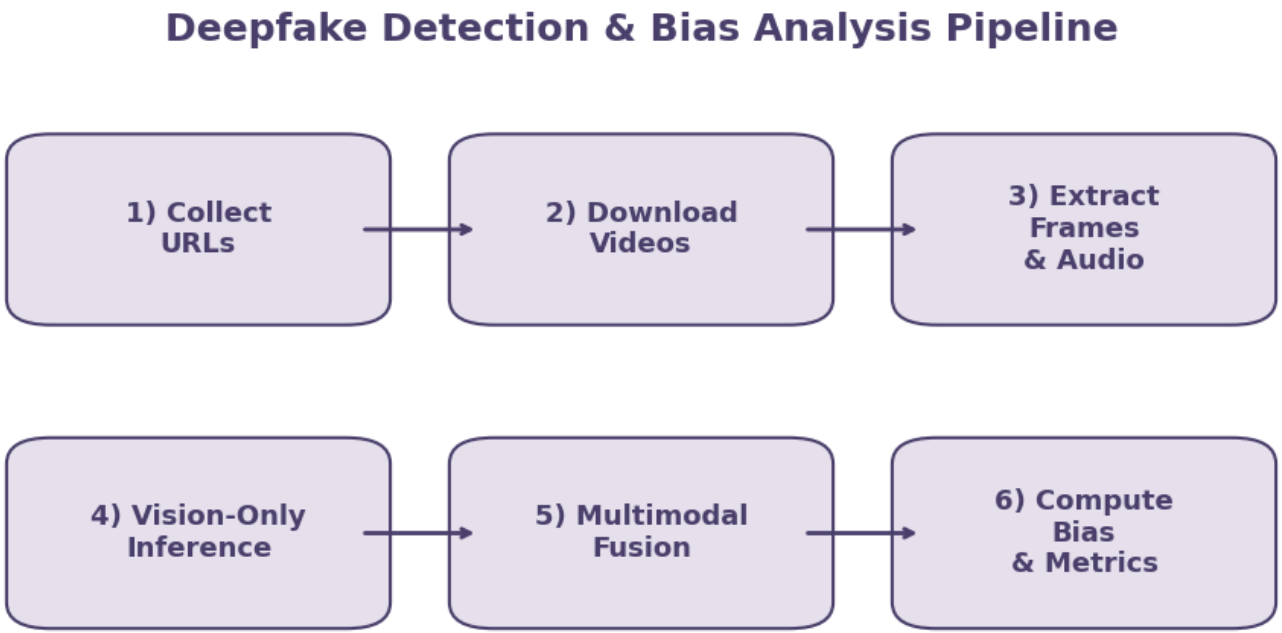
Detecting Political Deepfakes: Vision, Audio & Political Bias

Jonathan Monroe

MACSS 30200 Final Presentation

Pipeline Overview



Interpreting the Pipeline

- **Data Collection:** Scrape real/fake YouTube clips for Biden & Trump.
- **Preprocessing:** Crop face-centered frames and extract 2 s audio snippets.
- **Modeling:** (1) Vision-only transformer, (2) Multimodal fusion via logistic regression.
- **Evaluation:** ROC AUC, false-positive rate gap, Grad-CAM, coefficient analysis.

Testing Output Summary

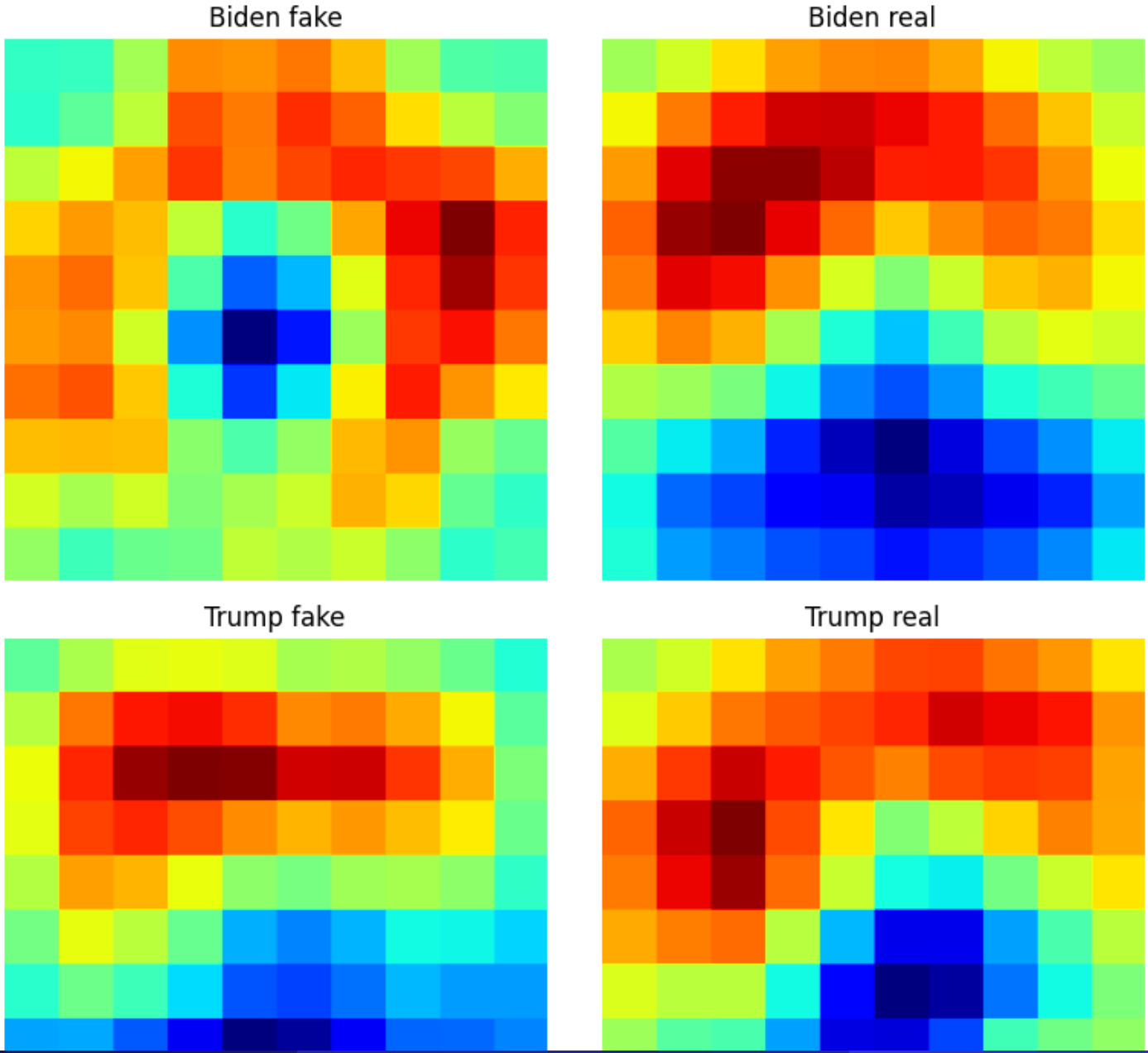
Key Test Metrics

Model	ROC AUC	FPR (Biden–Trump)
Vision-Only	0.461	0.031
Multimodal	0.990	0.056

Interpreting Test Results

- Vision-only vs. multimodal AUC shows a dramatic performance lift.
- FPR line tracks change in false-positive bias gap by candidate.
- Audio + vision boosts accuracy but increases Biden's false-positive rate.

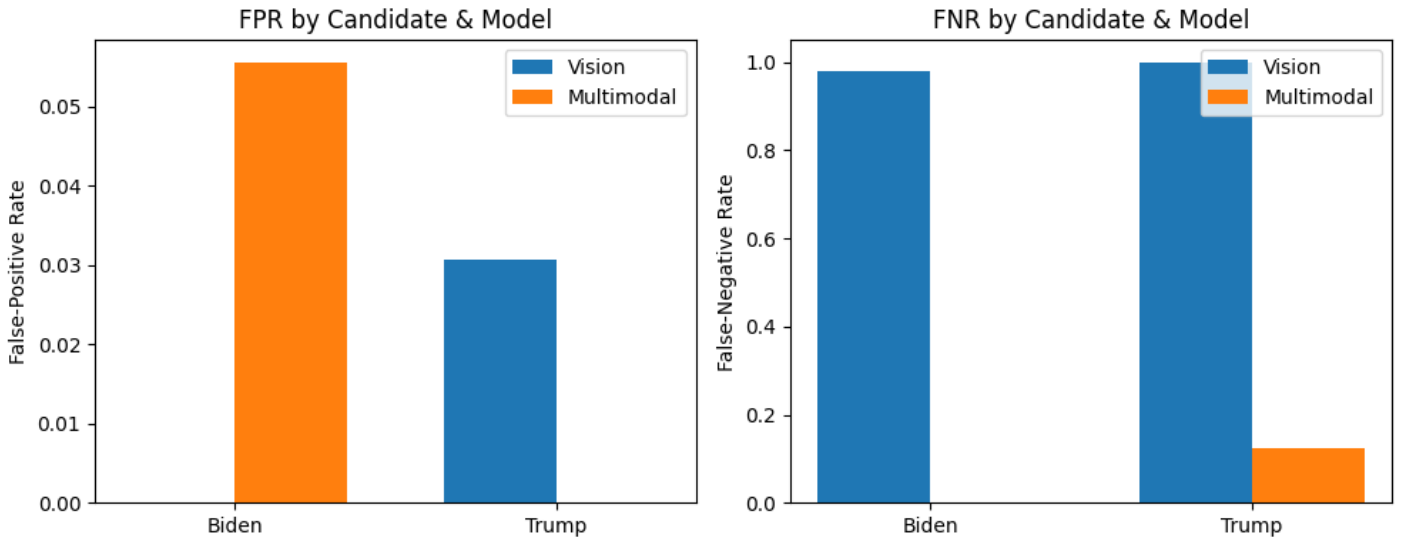
Center-of-Mass Heatmaps



Interpreting COM Heatmaps

- Shows average Grad-CAM attention location per group.
- Fake frames focus higher on the face; real frames focus lower.
- Biden vs. Trump reveals distinct spatial "attention" patterns.

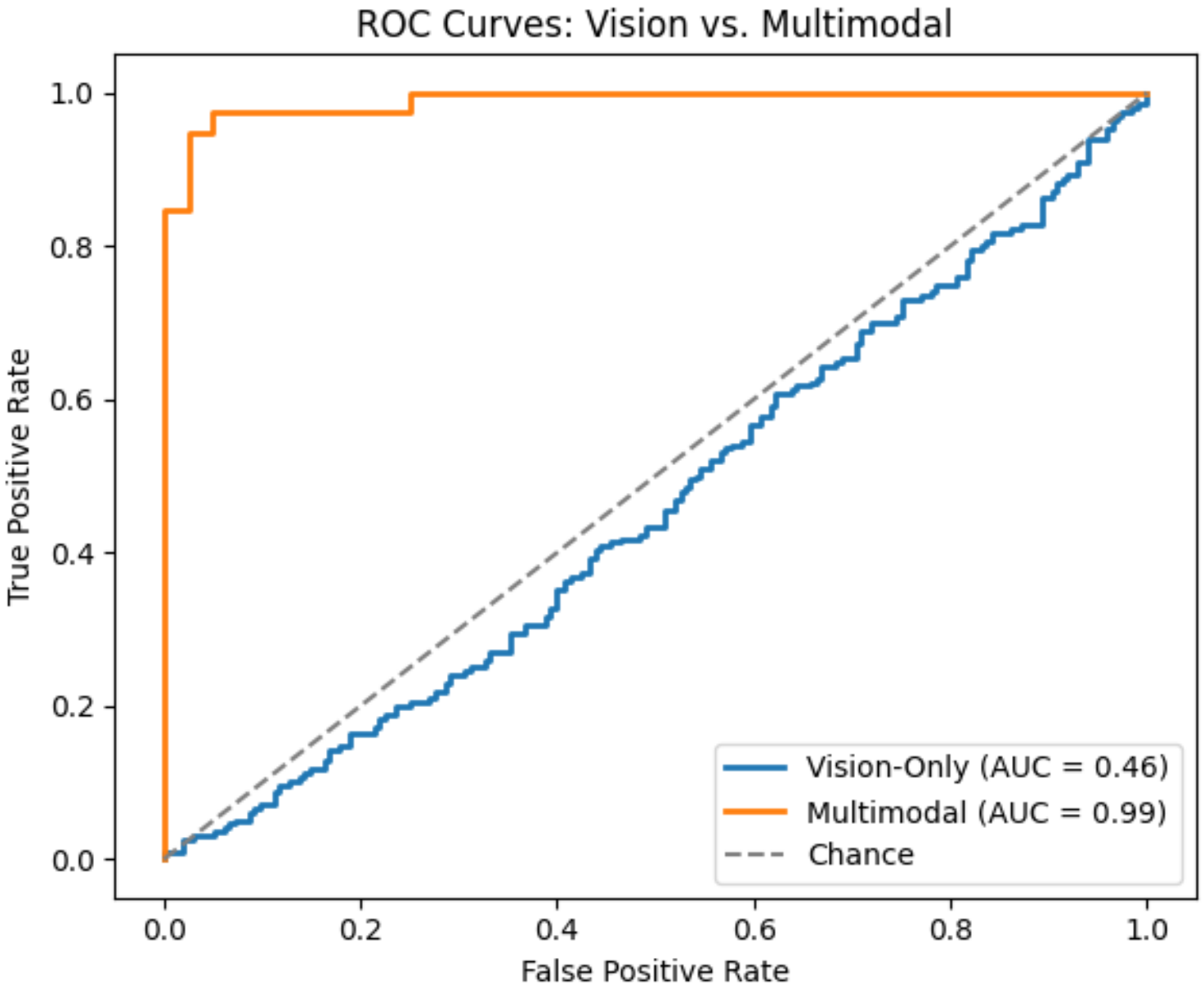
Error Rates by Candidate



Interpreting Error Rates

- **FPR:** Real flagged as fake; Trump higher under vision-only.
- **FPR shift:** Biden's false positives increase when audio is added.
- **FNR:** Fake missed as real; audio reduces false negatives overall.

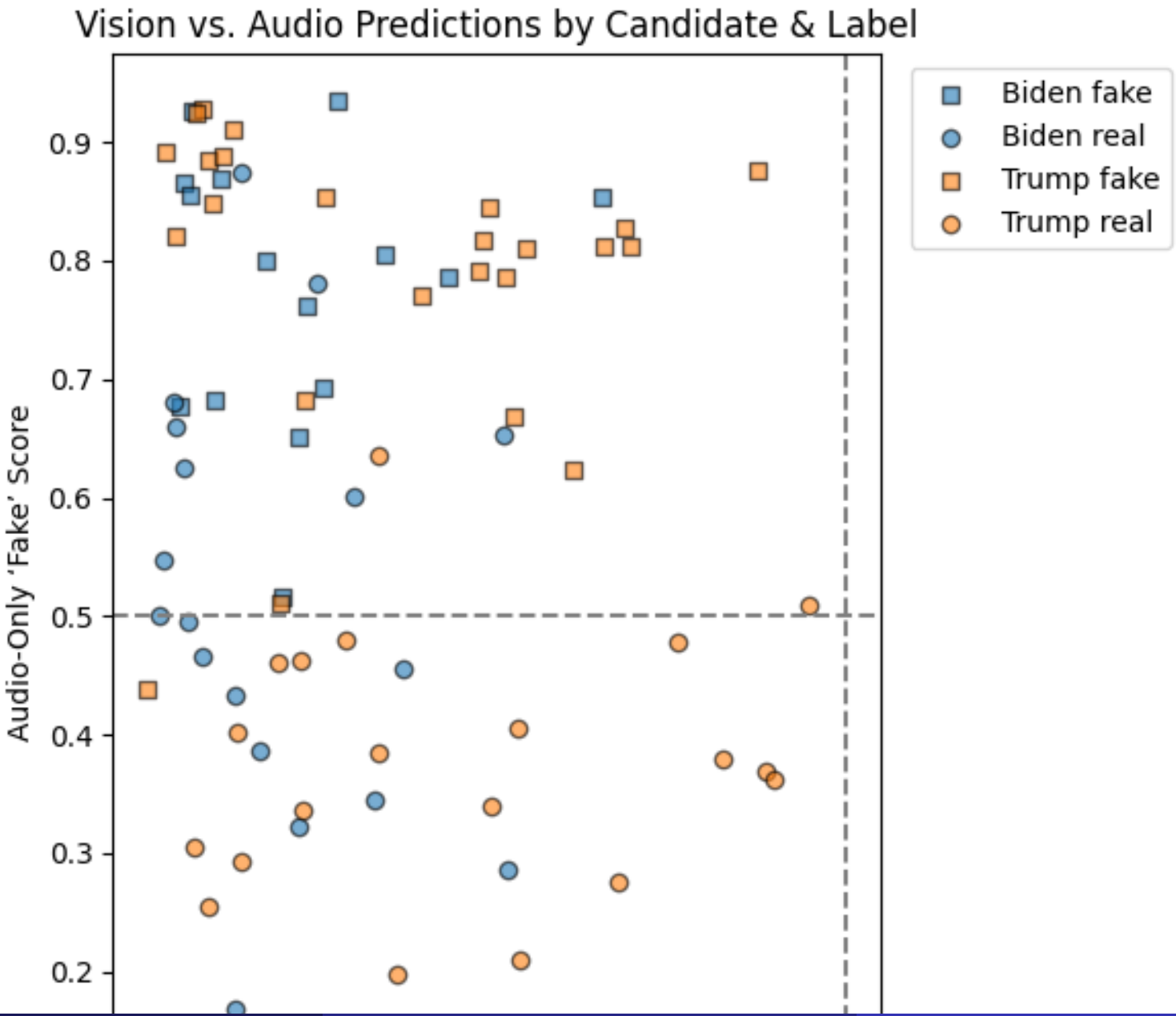
ROC AUC Curve



Interpreting ROC AUC

- Plots true positive rate vs. false positive rate at various thresholds.
- Area under the curve (AUC) quantifies overall classification performance.
- Multimodal model's curve approaches the top-left corner (near-perfect).
- Vision-only model tracks closer to the diagonal (near-random).
- The gap at low FPR highlights bias shifts between candidates.

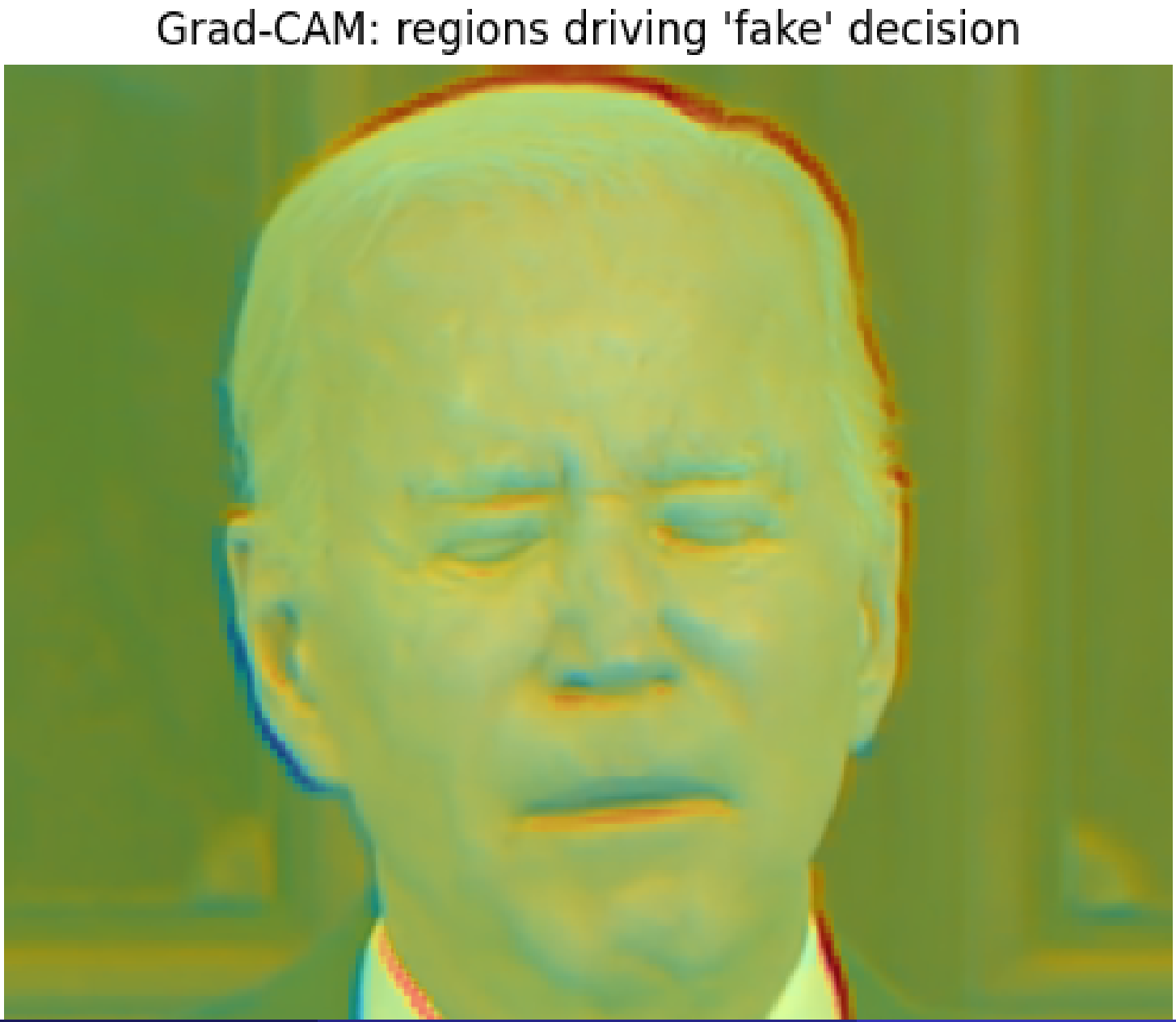
Fusion Decision Surface



Interpreting Decision Surface

- X-axis: vision-only fake score; Y-axis: audio-only fake score.
- Dashed line: logistic-reg real→fake decision boundary.
- Clustering by candidate shows group-specific score distributions.

Grad-CAM Example



Interpreting Grad-CAM

- Warmer regions show pixels that most influenced the model's decision.
- Highlights where the vision model "looks" in fake vs. real frames.
- Candidate-specific maps reveal systematic attention differences.

Why Logistic Regression?

- Simple, interpretable decision boundary in 2D score space.
- Coefficients directly measure modality importance.
- Enables quantitative comparison of audio vs. vision influence.

Conclusion

- Multimodal fusion drastically improves detection (AUC ↑) while shifting bias.
- Grad-CAM COM reveal where vision models focus by candidate.
- ROC AUC curve clearly distinguishes model performance levels.

Contact & Repo

GitHub: `https://github.com/JonathanPMonroe/MACSS-30200`
Email: jonathanmonroe@uchicago.edu