DivFusion: Divergence-Based Multimodal Fusion
for Ambivalence and Hesitancy Recognition

Aislan Gabriel O. Souza · Bruno Fernandes · Leandro Honorato Silva · Agostinho Freire
Universidade de Pernambuco (UPE) — PPGEC, Recife, Brazil

CVPR JUNE 3-7, 2026 DENVER COLORADO

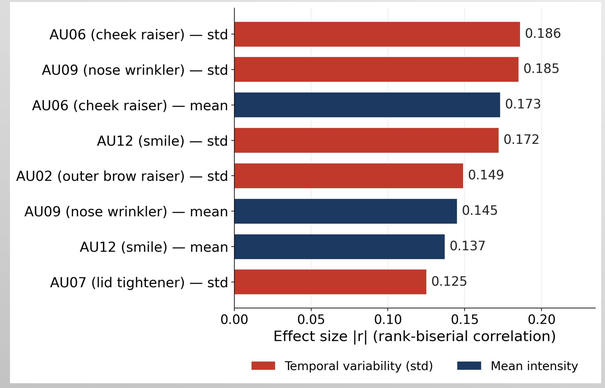
Abstract

Problem
- Ambivalence/hesitancy (A/H) is a behavioural state defined by conflicting signals across face, voice and words — not by a single distinctive expression.
- Prior fusions (late blending, early concatenation) treat modalities as complementary and never measure cross-modal disagreement explicitly.

Hypothesis & approach
- The divergence between modalities — not their content — is the defining cue for A/H. DivFusion computes pairwise absolute differences between projected modality embeddings.

Result on BAH (10th ABAW)
- Macro F1 = 0.6808 on the public test set.

Contributions
- (1) Divergence-based fusion that explicitly models cross-modal conflict.
- (2) Interpretable AU features with temporal-window stats capturing facial instability.
- (3) Full ablation: 3 modalities × 3 fusion strategies on BAH.



Methods and Materials

Dataset
- BAH (Behavioural Ambivalence/Hesitancy): {598 train / 107 val / 427 test} videos with synchronised face video, audio, and transcripts.
- Binary annotation per video: A/H vs. non-A/H.
- Released for the 10th ABAW Challenge (CVPR 2026), A/H Recognition track.

Per-modality features
- Visual: 20 continuous Action Units / frame (Py-Feat); ~10 fps; optional windowed stats (μ, σ, slope, range), W=16, S=8.
- Audio: Wav2Vec 2.0 (base-960h), 768-d @ 50 Hz.
- Text: transcripts → BERT-base [CLS]; last 2 layers fine-tuned at 1/10 LR.

Temporal modelling
- 2-layer BiLSTM (hidden 64 → output 128) per temporal modality with masked attention pooling.
- Linear projection into shared D = 128 space → h'_v, h'_a, $h'_□$.

Fusion strategies (compared)
- A — implicit: concat(h'_v, h'_a, $h'_□$)
- B — divergence (ours): pairwise $|h'_i - h'_□|$
- C — combined: concat(A, B)
- Classifier: 3-layer MLP (3D → 128 → 64 → 1), ReLU, dropout 0.3.

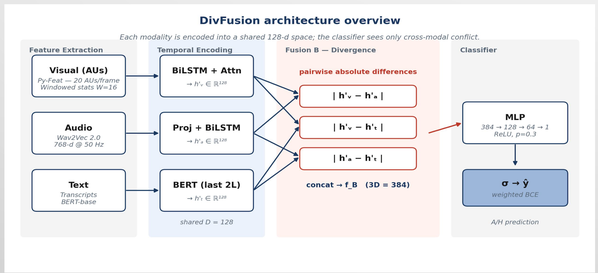
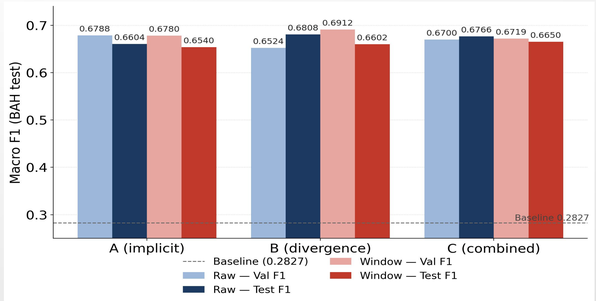


Figure 1. DivFusion overview — pairwise divergences between projected modality embeddings.

Results

Unimodal Macro F1 (BAH test)
- Visual (AUs): XGBoost 160-d → 0.5642 · BiLSTM raw → 0.5490
- Audio (Wav2Vec 2.0): BiLSTM + proj → 0.6141 · XGBoost+PCA → 0.6021
- Text (BERT, partial fine-tune): 0.5904

Multimodal fusion (test, raw AUs)
- A (implicit): 0.6604
- B (divergence — ours): 0.6808 ← best
- C (combined): 0.6766
- Divergence-based fusion beats both standard concatenation strategies — supporting the hypothesis that explicit cross-modal conflict is the key signal for A/H.



0.6808
Macro F1 on BAH (10th ABAW) test
Official 10th ABAW challenge submission (private test): 0.5362.

Discussion

Why divergence works
- Aligns with the theoretical definition of A/H as coexisting conflicting signals — when face/voice/words disagree, the $|h'_v - h'_□|$ and $|h'_v - h'_a|$ terms become large.

Divergence is the load-bearing component
- Strategy C (concat A + B) reaches 0.6766 — slightly below B alone (0.6808).
- Adding the raw concatenated embeddings on top of the divergence terms doesn't help and may dilute the conflict signal.
- Implication: the lift comes specifically from explicit cross-modal disagreement, not from feeding the classifier more features.

The visual ceiling
- Visual-only models plateau at Macro F1 ≈ 0.56 across architectures; AU effect sizes are small.
- Multimodal fusion lifts performance to 0.68 by combining the weak visual signal with stronger audio (0.61) and text (0.59).

Conclusions
- DivFusion explicitly models inter-modality conflict and reaches Macro F1 = 0.6808 on BAH (10th ABAW)
- Divergence > concatenation: measuring how modalities disagree is more informative than combining their content for A/H.
- Temporal AU variability — not mean intensity — is the dominant visual discriminator of A/H.

Future work
- Contrastive alignment regulariser to enforce semantic correspondence in the shared space.
- Utilize more robust feature extraction encoders.
- Temporal cross-modal alignment to localise specific conflict moments.

Contact

Aislan Gabriel O. Souza · UPE — PPGEC, Recife, Brazil · aislan.gabriel@upe.br
Advisor: Prof. Bruno Fernandes (UPE/PPGEC) · Co-authors: L. Honorato Silva, A. Freire

References

1. Gonzalez-Gonzalez et al. BAH dataset for ambivalence/hesitancy recognition in videos for digital behavioural change. ICLR 2026.
2. Kollias et al. From emotions to violence: Multimodal fine-grained behaviour analysis at the 9th ABAW. ICCV 2025.
3. Kollias et al. Advancements in affective and behaviour analysis: 8th ABAW workshop. CVPR 2025.
4. Baevski et al. wav2vec 2.0: A framework for self-supervised learning of speech representations. NeurIPS 2020.
5. Devlin et al. BERT: Pre-training of deep bidirectional transformers for language understanding. NAACL 2019.
6. Cheong et al. Py-Feat: Python facial expression analysis toolbox. Affective Science 4, 2023.
7. Savchenko. HSEmotion team at ABAW-8 competition. arXiv:2503.10399, 2025.
8. Hallmen et al. Semantic matters: Multimodal features for affective analysis. arXiv:2504.11460, 2025.