# Ambivalence/Hesitancy (AH) Video Recognition Challenge, 3rd edition, ABAW11th - ECCV 2026

**Task: Video classification**

**Jul-Aug 2026**

LIVIA, ETS Montreal, Canada
Dept. of Health, Kinesiology & Applied Physiology, Concordia University, Montreal, Canada

ÉTS Le génie pour l'industrie — ÉCOLE DE TECHNOLOGIE SUPÉRIEURE, Université du Québec

ABAW

ECCV 2026 | MALMÖ | SEPT 8–12

UNIVERSITÉ Concordia UNIVERSITY

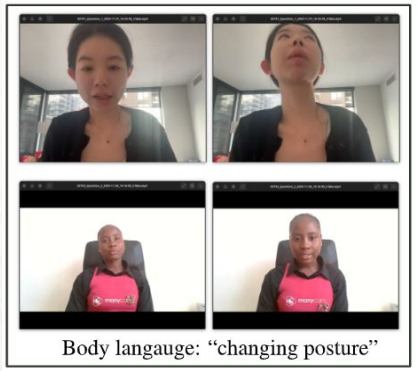
LIVIA — LABORATOIRE D'IMAGERIE, DE VISION ET D'INTELLIGENCE ARTIFICIELLE

CHAIRE DOUBLE FRSQ EN IA ET EN SANTÉ NUMÉRIQUE — FRSQ DOUBLE CHAIR IN AI AND DIGITAL HEALTH

# Outline

- **Ambivalence/Hesitancy (AH) Video Recognition Challenge**
- **Proposed Methods**
- **Leaderboard**

# Task Ambivalence/Hesitancy (AH) Video Recognition Challenge


Body langauge: “changing posture”


Body langauge: “looking away”

Examples of videos from BAH[1].

- This is the second edition of AH Challenge. It is held within ABAW 10th Workshop - CVPR 2026
- AH Challenge consists in predicting whether there is ambivalence/hesitancy in a video
- Teams are provided access to BAH dataset [1] of annotated videos for training. A private test set is held to be released at the end of the challenge
- Call for the challenge: link

**Important Dates**:

| | |
|---|---|
| Call for participation announced, team registration begins, data available: | May 25, 2026 |
| Test set release: | July 10, 2026 |
| Final submission deadline (Predictions, Code and ArXiv paper): | July 16, 2026 |
| Winners Announcement: | July 18, 2026 |
| Final Paper Submission Deadline: | July 20, 2026 |
| Review decisions sent to authors; Notification of acceptance: | August 10, 2026 |
| Camera ready version: | August 15, 2026 |

[1]: “BAH Dataset for Ambivalence/Hesitancy Recognition in Videos for Digital Behavioural Change”, González et al., ICLR 2026.

# Task Ambivalence/Hesitancy (AH) Video Recognition Challenge

- 17 teams registered
- 13 teams submitted predictions
- Link to papers
- Link to leaderboard

| Rank | Teams | Method |
|---|---|---|
| 1 | AffectVision | Multimodal, Gated fusion (AMF), Textual markers, ASR-erased time |
| 2 | RAS | Multimodal, text-centric |
| 3 | PUCPR | Text-audio, handcrafted support psycholinguistic features |
| 4 | AIM for Emo | Multimodal with behavioural-statistics descriptors |
| 5 | AIWELL | Multimodal, pose, ensembles 3 models |
| 5 | CASIA-26 | Modality discrepancy |
| 6 | Tamimi | Text-only |
| 7 | CPR | Modality discrepancy |
| 8 | IUSD | Multimodal with gated fusion |
| 9 | MINDH Lab | N/A yet |
| 10 | HSEmotion | Multimodal with late fusion |
| 11 | GGL | Audio-text with bidirectional cross-attention |
| 12 | AVULAB | Multimodal with synchronized face and full frame |

# Teams Proposed Methods

# Team 1: AffectVision

[Paper][Code]

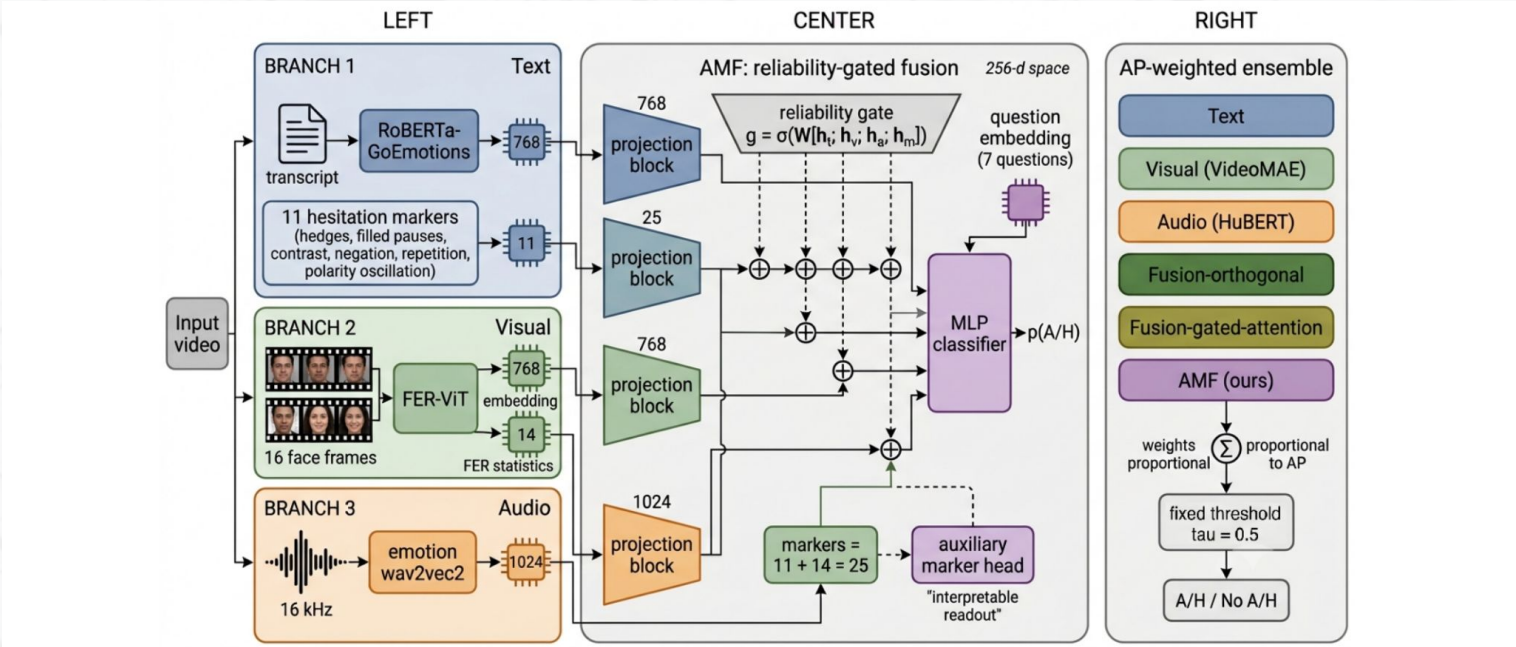



Figure 1: Overview. Text, video, and audio inputs are turned into affect-specialised features. AMF gates weak channels and predicts A/H. Six members are combined by AP-weighted averaging at a fixed threshold of 0.5. Features that hurt the model (gaze/brow dynamics, prosody, cue supervision) are excluded.

- Multimodal framework: visual, audio, and transcript
- Introduce and compute 11 interpretable text markers: *rates of hedges, filled pauses, contrastive ("but"-type) words, negations, first-person uncertainty, repeated words, and how much sentiment flips within the answer*
- Leverage the question in the video
- Introduces AMF for Affective Marker Fusion which assigns a weight to each modality stream among visual, autio, text, and markers.
- Dealt with fillers deletion in speech recognizers, and recover them using timestamps in transcripts.
- The training regurlarizes the classification loss of the fusion using the classification loss over the markers prediction.
- Claim: results show that AH is concentrated in language
- Use ensembles

# Team 1: AffectVision

[Paper][Code]

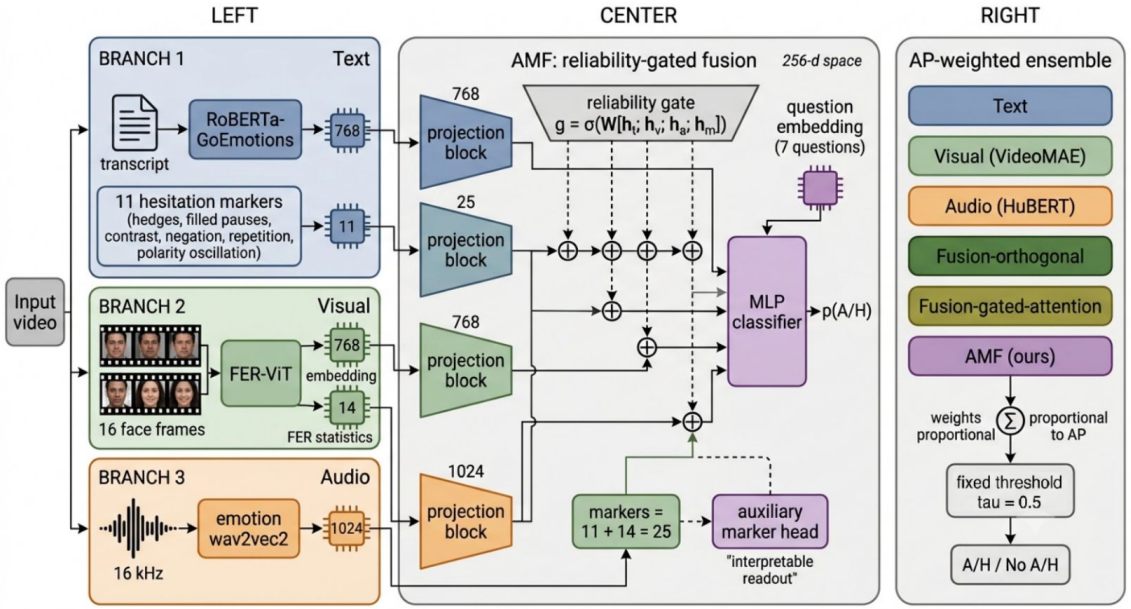



Figure 1: Overview. Text, video, and audio inputs are turned into affect-specialised features. AMF gates weak channels and predicts A/H. Six members are combined by AP-weighted averaging at a fixed threshold of 0.5. Features that hurt the model (gaze/brow dynamics, prosody, cue supervision) are excluded.

Table 4: Per-stream logistic probe, fixed threshold 0.5.

| Feature stream | Val F1 | Test F1 | Test AP |
|---|---|---|---|
| Text: RoBERTa-GoEmotions (768) | 0.632 | 0.646 | **0.811** |
| Hesitation markers (11) | 0.615 | 0.672 | 0.800 |
| Audio: emotion wav2vec2 (1024) | 0.647 | 0.610 | 0.764 |
| ASR-erased time (16) | 0.497 | 0.566 | 0.718 |
| FER statistics (14) | 0.529 | 0.551 | 0.669 |
| Video: AU/gaze (36) | 0.538 | 0.540 | 0.665 |
| Video: VideoMAE general (768) | 0.483 | 0.538 | 0.650 |
| Video: FER embedding (768) | 0.506 | 0.507 | 0.636 |
| Audio: HuBERT general (1024) | 0.283 | 0.283 | 0.606 |

Table 7: All-data members on the shared 113-video holdout; AP weights, fixed $\tau = 0.5$. The holdout serves both early stopping and weighting, so member numbers are mildly optimistic; the public test (Table 3) is the clean evidence.

| Member | Holdout F1 | Holdout AP |
|---|---|---|
| Text (RoBERTa-GoEmotions) | 0.788 | 0.906 |
| Video (VideoMAE) | 0.770 | 0.871 |
| Audio (HuBERT) | 0.726 | 0.831 |
| Fusion (orthogonal) | 0.814 | 0.895 |
| Fusion (gated) | 0.805 | 0.904 |
| AMF (ours) | 0.751 | 0.900 |
| Ensemble, no AMF (5) | 0.823 | 0.903 |
| **Ensemble, with AMF (6)** | 0.814 | **0.907** |

# Team 2: RAS

[Paper][Code]

- Multimodal (cropped face, full frame, audio, transcript), text-centric framework
- Text-residual fusion model

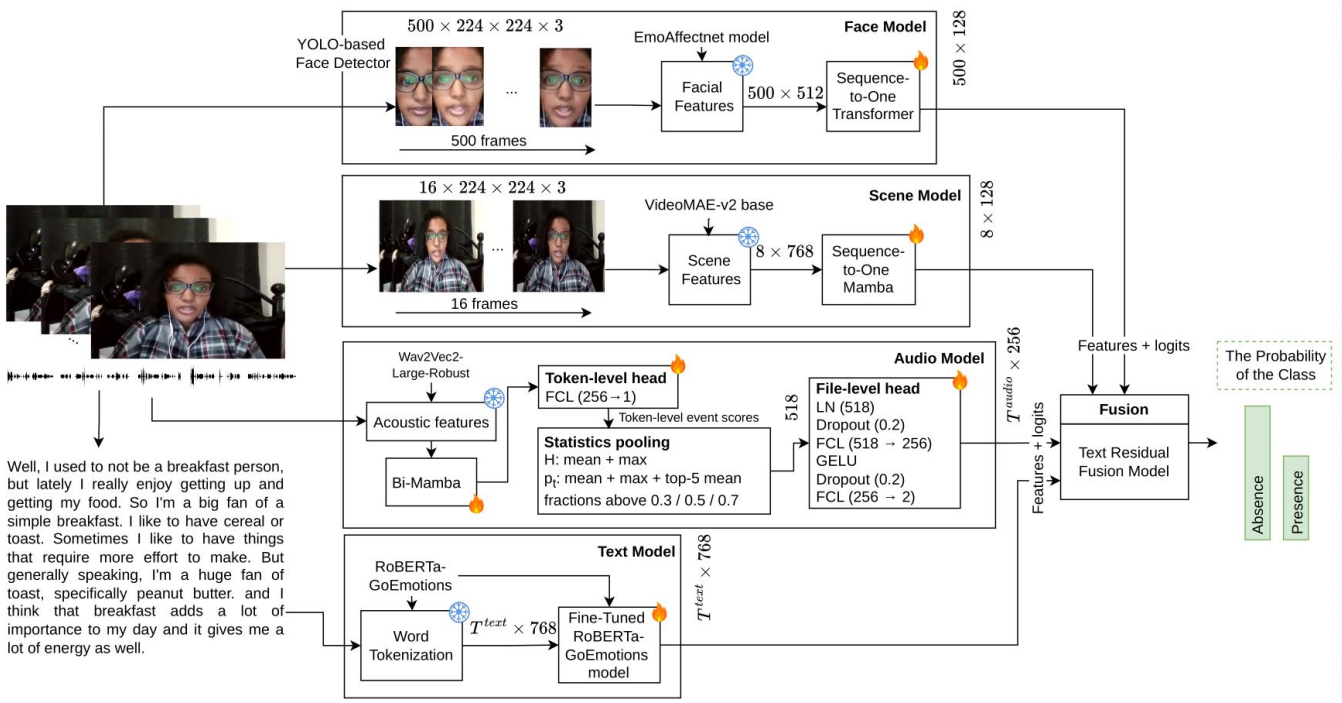



**Fig. 1:** Pipeline of text-centered multimodal approach.

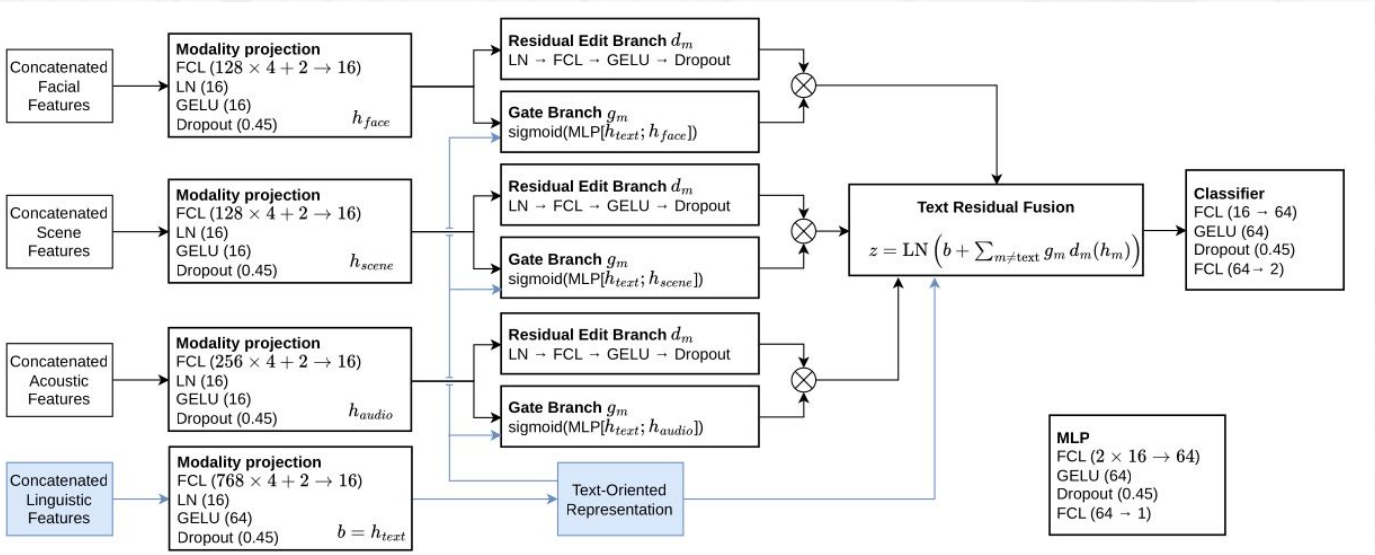



**Fig. 2:** Pipeline of the Text Residual Fusion model.

# Team 2: RAS

[Paper][Code]

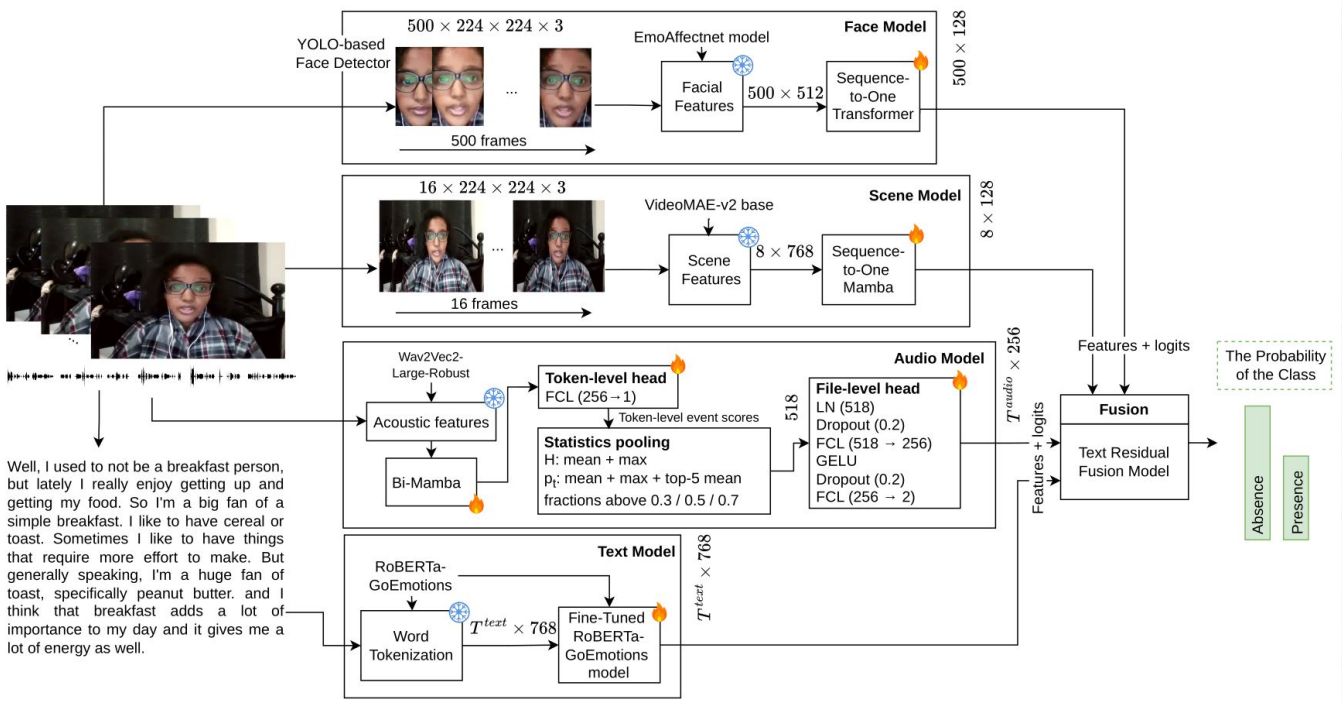



**Fig. 1:** Pipeline of text-centered multimodal approach.

**Table 1:** Experimental results (MF1, %) obtained by various configurations of the proposed approach. FM refers to flow matching. LS to label smoothing. TS to temporal smoothing.

| ID | Modality | Features | Temporal model | Regularization | Devel subset | Public Test subset | Average Devel/Public Test | Private Test subset |
|---|---|---|---|---|---|---|---|---|
| 1 | Text | RoBERTa-GoEmotions (freeze layers=4) | – | **72.55** | **72.10** | **72.33** | 74.21 |
| 2 | Audio | Wav2Vec2 [2] | Bi-Mamba | TS | 70.19 | 69.28 | 69.74 | – |
| 3 | Face | EmoAffectNet [24] | Transformer | FM | 64.07 | 61.27 | 62.67 | – |
| 4 | Scene | VideoMAE-v2 base [29] | MambaSSM | LS | 61.07 | 61.18 | 61.12 | – |
| 5 | Text, Audio, Face, Scene | IDs 1, 2, 3 and 4 | Text Residual Fusion | LS | 76.13 | 74.14 | 75.14 | 78.24 |

# Team 3: PUCPR

[Paper][Code]

**Table 1.** Composition of the support feature vector.

| Feature group | Dim. |
|---|---|
| Question type | 7 |
| Timing and pauses | 6 |
| Speech rate and articulation | 3 |
| Pitch and intonation | 4 |
| Loudness | 2 |
| Voice quality and uncertainty score | 3 |
| Ambivalence indices | 20 |
| Emotion dynamics | 6 |
| Contrast and polarity shift | 5 |
| Epistemic hedges | 18 |
| Total | 74 |

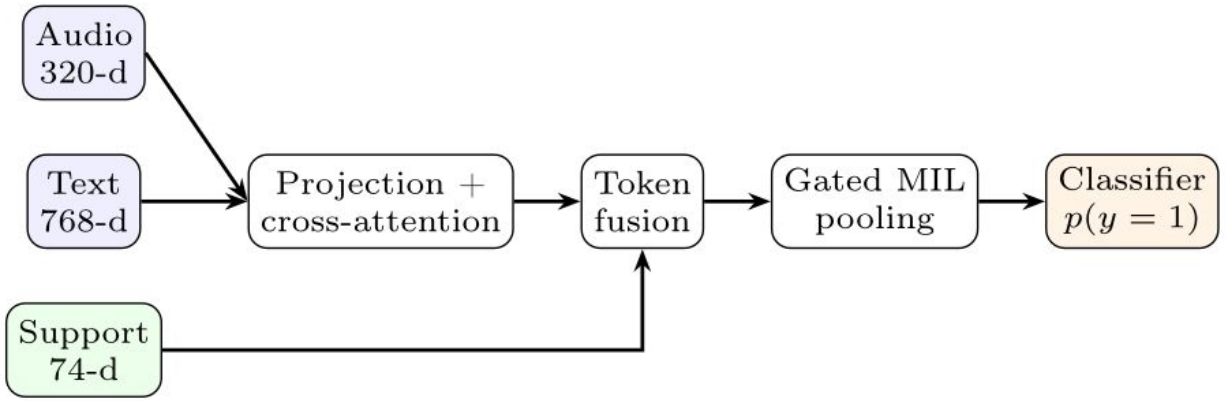



**Fig. 1.** Overview of the proposed audio–text architecture.

*Psycholinguistic support features.* Table 1 summarizes the 74-dimensional support vector. Seven dimensions encode the question type. Eighteen acoustic dimensions describe timing, pauses, speaking rate, pitch, loudness, jitter, shimmer, and a composite uncertainty score. Forty-nine textual dimensions encode positive/negative intensity, psychometric ambivalence indices, emotional dynamics, contrast and polarity shifts, and epistemic hedges. Response-level measures are broadcast to all windows of a video, while local hedge, contrast, entropy, and prosodic measures remain window-specific.

- Multimodal: audio, text, handcrafted support psycholinguistic features (uncertainty, hedging, and attitudinal conflict)
- Excludes visual modality
- Temporal fusion for audio and text. Support features are injected after fusion
- MIL is used over a video that is represented using overlapping 5-sec windows, aligned with transcript

# Team 3: PUCPR

[Paper][Code]

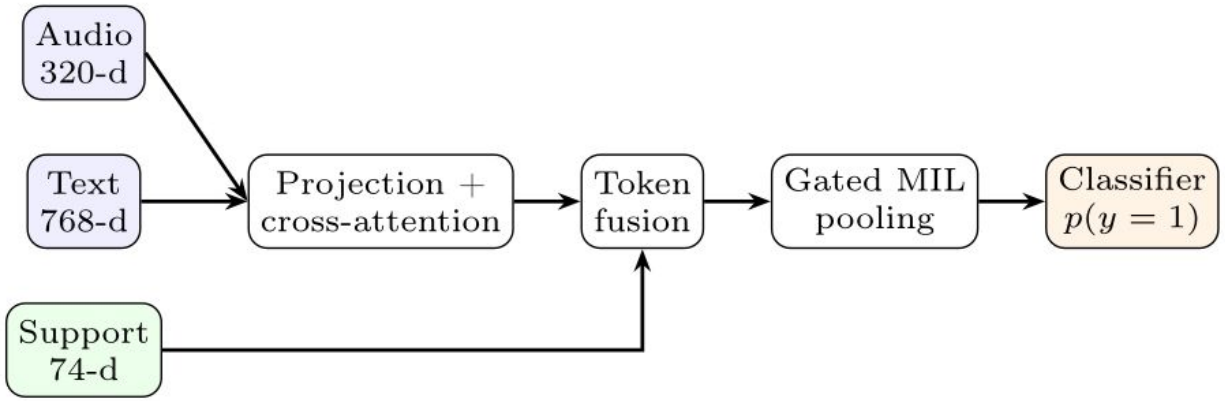



**Fig. 1.** Overview of the proposed audio–text architecture.

**Table 2.** Results on the 525-video labeled test split of BAH (ABAW10).

| System | Macro-F1 | AP |
|---|---|---|
| ConflictAwareAH [2] (Fennec) | 0.694 | — |
| Mean pooling baseline | 0.701 | 0.828 |
| MIL, audio + text | 0.705 | 0.835 |
| MIL + 74 support features, 5 seeds | **0.722** | **0.875** |

# Team 4: AIM for Emo

[Paper][Code]

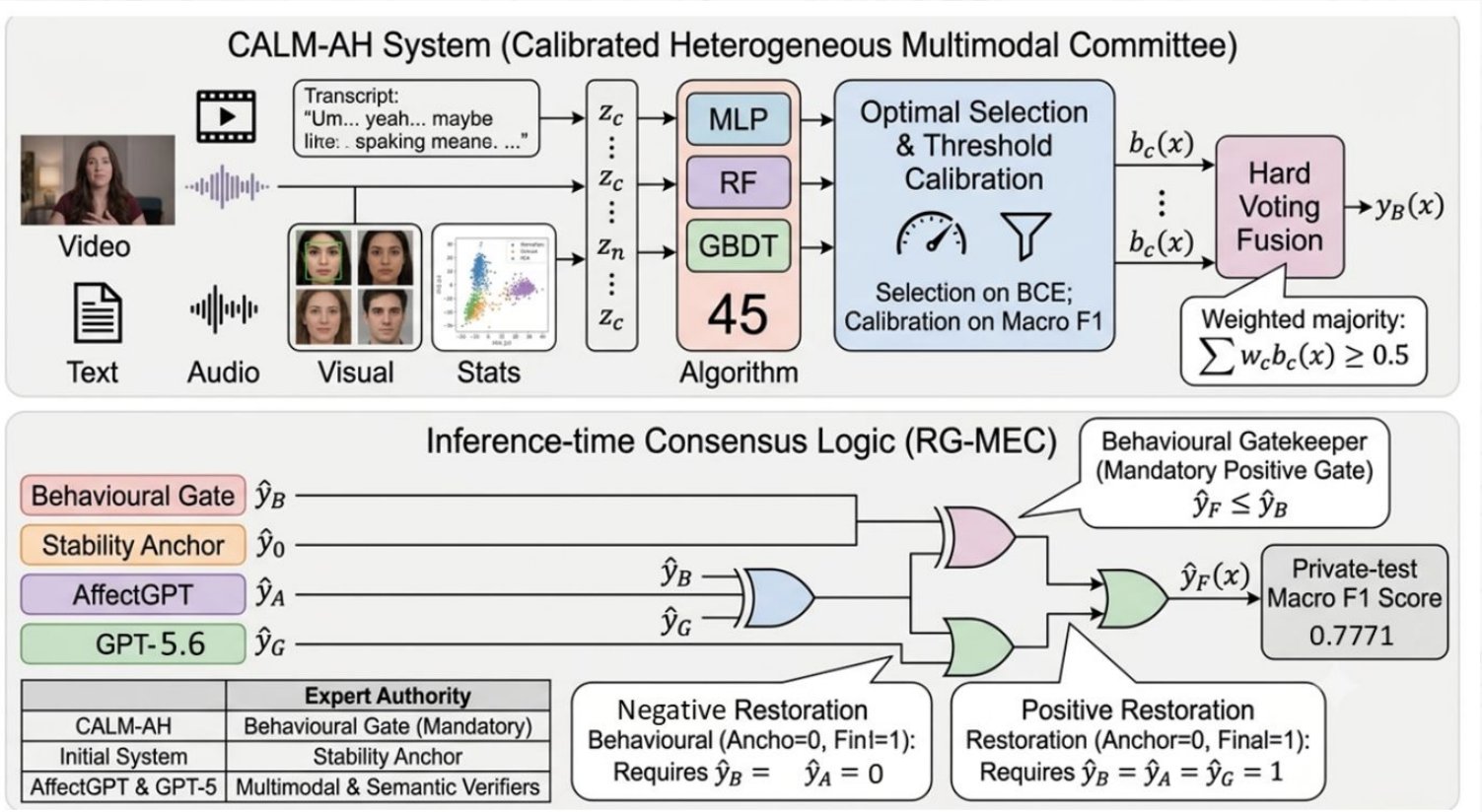



**Fig. 1:** The pipeline of our method.

- Multimodal framework (visual, audio, text, behavioural-statistics descriptors): CALM-AH
- Ensemble-based framework
- For each modality, MLP, random-forest, and gradient-boosting candidates are trained
- Reliability-gated multi-expert consensus

# Team 4: AIM for Emo

[Paper][Code]

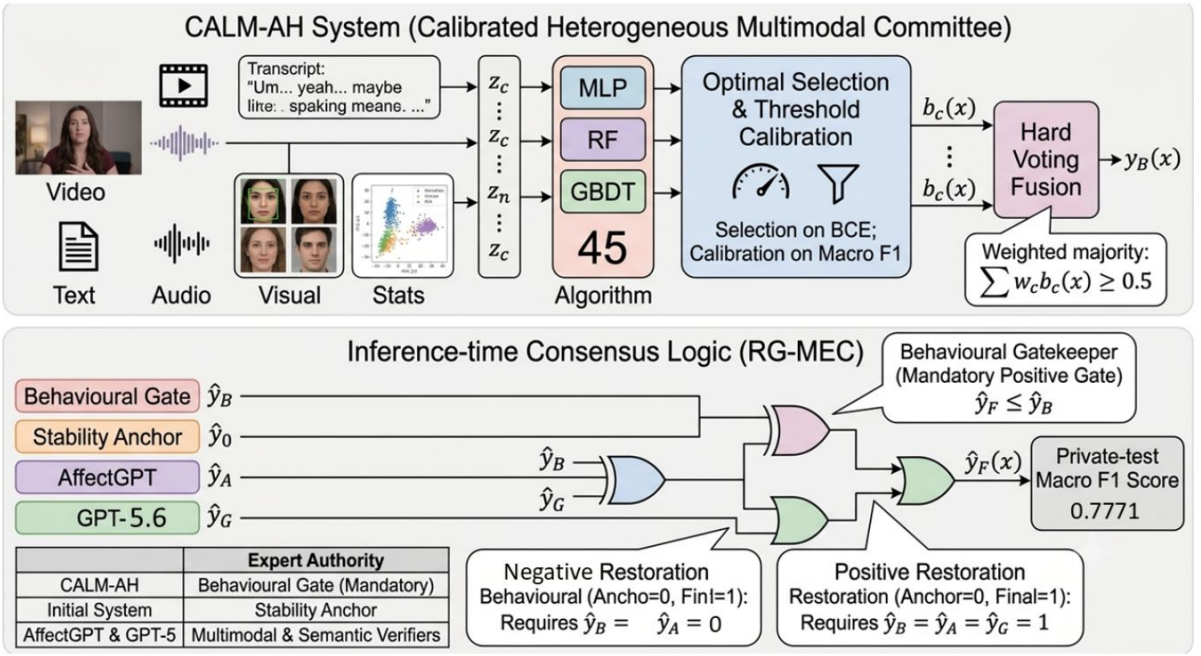



**Fig. 1:** The pipeline of our method.

**Table 2:** Public-test ablation of the asymmetric multi-expert decision rule. All systems are evaluated on the same participant-disjoint public-test partition.

| Configuration | Acc. | MF1 | No-A/H F1 | A/H F1 |
|---|---|---|---|---|
| Initial anchor $\hat{y}_0$ | 0.7790 | 0.7485 | 0.6608 | 0.8362 |
| CALM-AH $\hat{y}_B$ | 0.7752 | 0.7525 | 0.6776 | 0.8275 |
| Anchor gate $\hat{y}_B \wedge \hat{y}_0$ | 0.7829 | 0.7638 | 0.6968 | 0.8309 |
| Full RG-MEC $\hat{y}_F$ | 0.7981 | 0.7771 | 0.7088 | 0.8455 |

# Team 5: AIWELL

[Paper][Code]

- Multimodal framework: face, audio, transcript, pose.
- Ensembles 3 models trained with different architectural biases: CrossAttn, Reliability, Pose.

**Table 1:** Results on the BAH public test set, macro-F1 with the threshold selected on validation. The baseline score is the organizers' official number for this video pool [5].

| | System | Macro-F1 |
|---|---|---|
| | Baseline: zero-shot Video-LLaVA [10], vision only | 0.2827 |
| Our individual models | CrossAttn (face, audio, text) | 0.7178 |
| | Reliability (face, audio, text) | 0.7103 |
| | Pose (face, audio, text, pose) | 0.7235 |
| | CrossAttn, soft-F1 loss [2] | 0.733 |
| Our ensembles | Equal-weight (CrossAttn+Reliability+Pose) | 0.7336 |
| | + temperature calibration | **0.7358** |
| | + threshold 0.5 instead of 0.402 | 0.7066 |
| | + cache-personalization TTA [15], global | 0.725 |

**Table 2:** Submission scoreboard: macro-F1 on the public test set and on the private test set as each submission is scored. "–" marks not yet submitted.

| Sub.# | Configuration | Public | Private |
|---|---|---|---|
| 1 | Calibrated ensemble ($T$=1.10, thr. 0.402) | 0.7358 | **0.7361** |
| 2 | Calibrated ensemble, thr. 0.5 | 0.7066 | – |
| 3 | Soft-F1 individual model | 0.733 | – |
| 4 | Calibrated ensemble + TTA (global) | 0.725 | – |
| 5 | Pose member alone | 0.7235 | – |

# Team 5: CASIA-26

[Paper][Code]

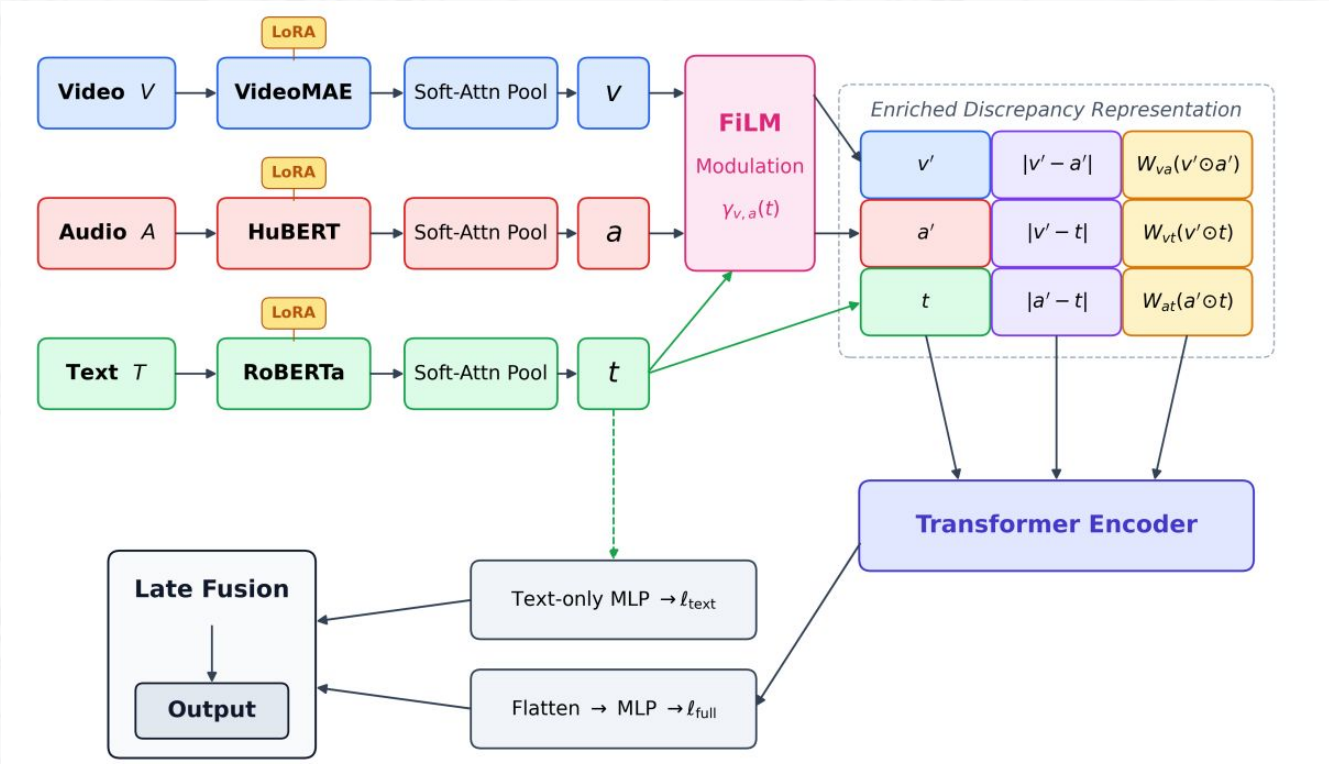



**Fig. 1:** MDT overview. Three frozen encoders produce modality embeddings, pooled via soft-attention and projected to $\mathbf{v}, \mathbf{a}, \mathbf{t}$. Six discrepancy features (three absolute-difference, three Hadamard-product) form a 9-token sequence processed by a 2-layer Transformer. Text-conditioned FiLM modulates video and audio before discrepancy computation. LoRA adapters fine-tune each encoder. A text-only auxiliary head blends with the multimodal output at inference.

- Modality Discrepancy Transformer (MDT)
- Based on work from AH 2nd edition challenge: '*Conflict-aware multimodal fusion for ambivalence and hesitancy recognition*'. It goes from 6 to 9 tokens to model discrepancy.

# Team 5: CASIA-26

[Paper][Code]

**Table 2:** A/H recognition performance on the BAH labelled test set (525 videos). Published baselines are from González *et al.* [5]; CA-AH refers to the conflict-aware architecture of Bekhouche *et al.* [3]. MDT denotes our full model.

| Model | Macro F1 |
|---|---|
| BAH: ZF M-LLM (vision only) [5] | 0.283 |
| BAH: Video-FocalNet [5] | 0.566 |
| BAH: LFAN (V+A+T) [5] | 0.593 |
| BAH: ZF M-LLM + transcript [5] | 0.634 |
| CA-AH [3] | 0.715 |
| **MDT (labelled test)** | **0.7408** |
| **MDT (private test, 151 unlabelled)** | **0.7368** |

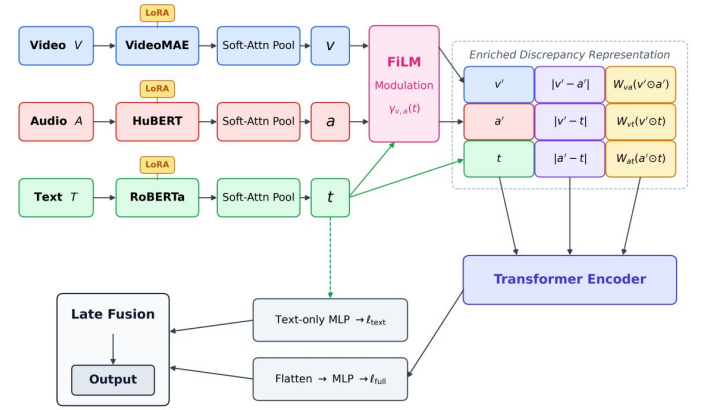



**Fig. 1:** MDT overview. Three frozen encoders produce modality embeddings, pooled via soft-attention and projected to $\mathbf{v}, \mathbf{a}, \mathbf{t}$. Six discrepancy features (three absolute-difference, three Hadamard-product) form a 9-token sequence processed by a 2-layer Transformer. Text-conditioned FiLM modulates video and audio before discrepancy computation. LoRA adapters fine-tune each encoder. A text-only auxiliary head blends with the multimodal output at inference.

**Table 5:** Incremental training strategy ablation. Each row adds one component to the CA-AH baseline (6-token abs, top-2 unfrozen, $\alpha$=0.6).

| Configuration | Macro F1 | F1-AH | F1-NoAH |
|---|---|---|---|
| CA-AH baseline | 0.7219 | 0.7982 | 0.6457 |
| + LoRA fine-tuning | 0.7366 | 0.8119 | 0.6613 |
| + Hadamard discrepancy (9-token) | 0.7322 | 0.7944 | 0.6700 |
| + Focal loss ($\gamma = 2.0$) | 0.7275 | 0.7836 | 0.6715 |
| + CutMix augmentation | 0.7300 | 0.8077 | 0.6524 |
| + LR warmup (5 epochs) | 0.7242 | 0.7883 | 0.6600 |
| + Multi-window training ($K = 3$) | 0.7354 | 0.8024 | 0.6684 |
| **MDT (all combined)** | **0.7408** | **0.7857** | **0.6959** |

# Team 6: Tamimi

[Paper][Code]

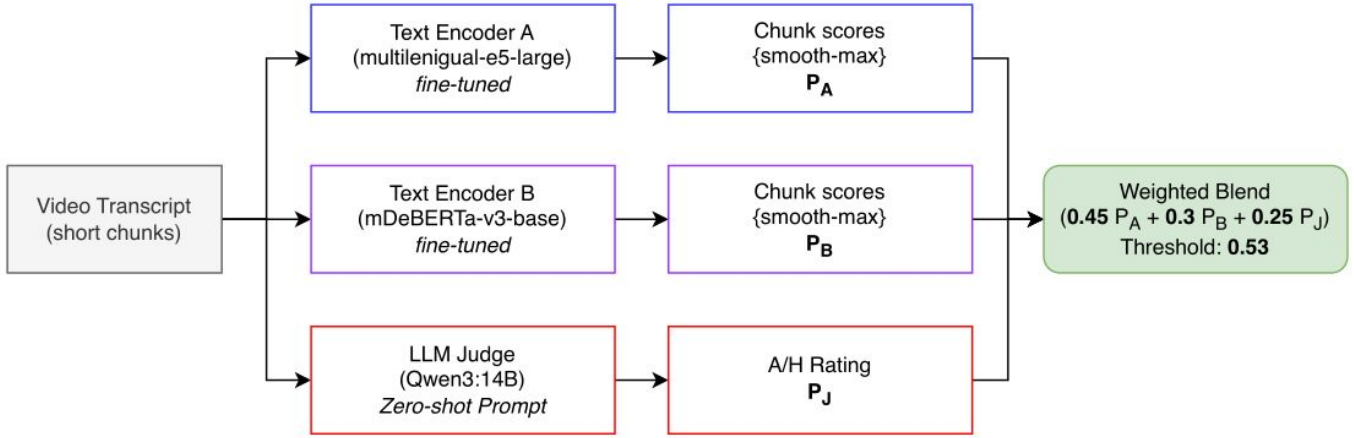



Figure 1: The TellTale system. The interview transcript feeds three streams: two fine-tuned text encoders whose per-chunk scores are pooled by a smooth maximum into video probabilities $p_{\mathrm{A}}$ and $p_{\mathrm{B}}$, and a zero-shot LLM judge whose 0–100 A/H rating becomes $p_{\mathrm{J}}$. A weighted average of the three probabilities, followed by one threshold, produces the decision. All weights and the threshold were chosen on participant-grouped cross-validated predictions only.

- Text-only framework
- Two text encoders are fientuned with LoRA adapters using MIL on chunks of transcripts, then pooled with smooth max-op
- A third stream is based on zero-shot LLM
- The three streams are weighted for a final prediction

# Team 6: Tamimi

[Paper][Code]

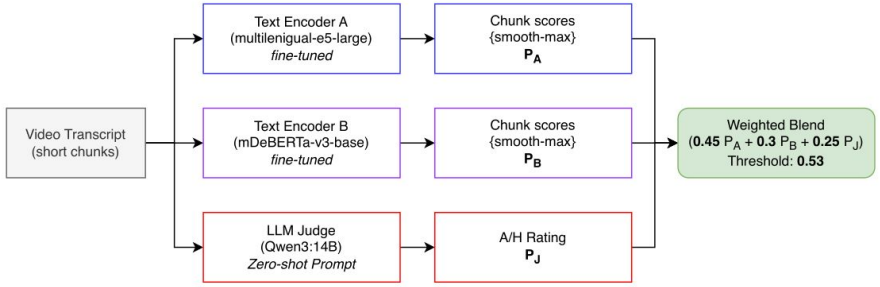



Figure 1: The TellTale system. The interview transcript feeds three streams: two fine-tuned text encoders whose per-chunk scores are pooled by a smooth maximum into video probabilities $p_A$ and $p_B$, and a zero-shot LLM judge whose 0–100 A/H rating becomes $p_J$. A weighted average of the three probabilities, followed by one threshold, produces the decision. All weights and the threshold were chosen on participant-grouped cross-validated predictions only.

Table 2: Private-test results (152 videos, 30 unseen participants) for the full system and its components. OOF is the pooled 5-fold out-of-fold estimate on the 1,427 labelled videos; blends show the conservative (leave-one-fold-out) estimate.

| System | OOF F1 | Private F1 | Private AP |
|---|---|---|---|
| Stream A alone | 0.7147 | 0.7165 | 0.8119 |
| Streams A+J | 0.7179 | 0.7168 | 0.7989 |
| Streams A+B | 0.7141 | 0.7235 | 0.8137 |
| Streams A+B+J (full) | 0.7213 | 0.7364 | 0.7940 |
| Official baseline [4] | — | 0.2827 | — |

# Team 7: CPR

[Paper][Code]

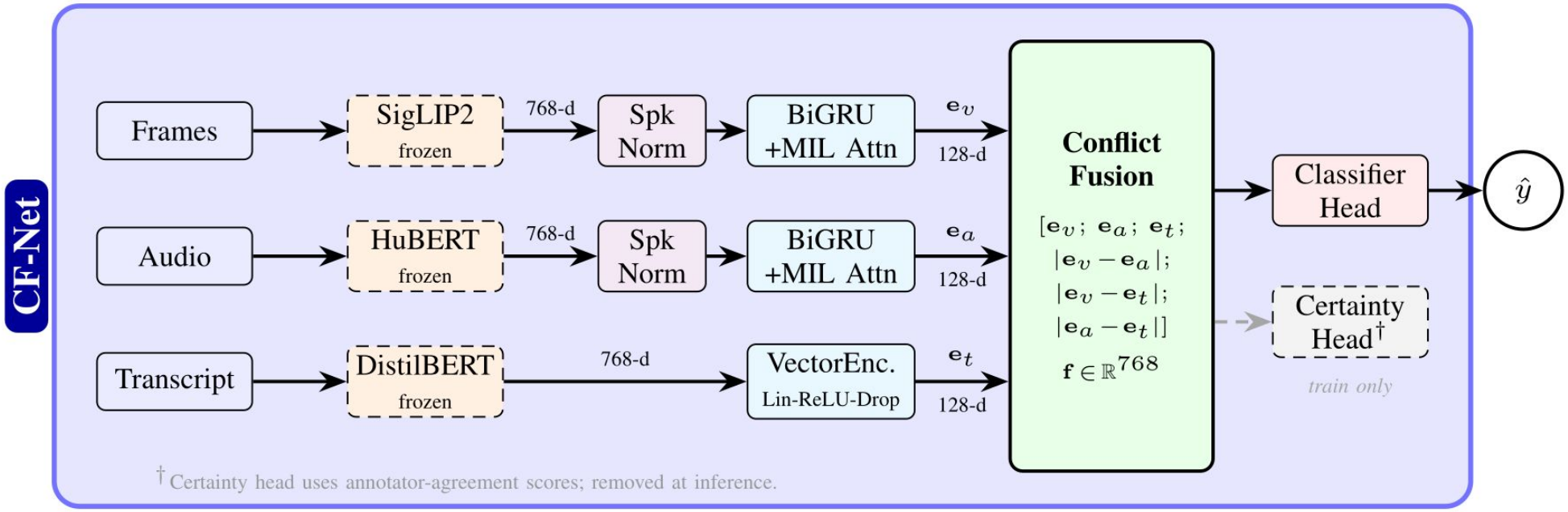



Fig. 1. Overview of CF-Net. Frozen SigLIP2, HuBERT, and DistilBERT backbones extract 768-d features per modality. Visual and audio features are speaker-normalised (Spk Norm) before being encoded by independent BiGRU encoders with MIL attention into 128-d embeddings $\mathbf{e}_v, \mathbf{e}_a$; the transcript embedding $\mathbf{e}_t$ is projected by a VectorEncoder. ConflictFusion concatenates the three embeddings with all pairwise absolute differences, producing $\mathbf{f} \in \mathbb{R}^{768}$. A classifier head produces the AH probability $\hat{y}$; an auxiliary certainty head (†, train only) regresses annotator-agreement scores to regularise learning on ambiguous samples.

- Multimodal frame (frame, audio, transcript), based on modeling discrepancy between modalities.
- Based on work from AH 2nd edition challenge: '*Conflict-aware multimodal fusion for ambivalence and hesitancy recognition*'.

# Team 7: CPR

[Paper][Code]

TABLE I

MACRO F1 AND AVERAGE PRECISION (AP) OF CF-NET ON THE BAH VALIDATION AND PRIVATE CHALLENGE TEST SETS. ↑: HIGHER IS BETTER. * PRIVATE TEST COMPRISES THE 152 HELD-OUT CLIPS SCORED BY THE CHALLENGE SERVER.

| | Validation (124) | | Private Test (152)* | |
|---|---|---|---|---|
| Method | F1↑ | AP↑ | F1↑ | AP↑ |
| CF-Net (seed 1234) | **0.7155** | **0.8236** | **0.7364** | **0.7439** |

TABLE III

COMPARISON ON THE BAH VALIDATION AND PRIVATE TEST SETS. ↑: HIGHER IS BETTER. —: NOT REPORTED. ‡: 10TH ABAW TEST (161 VIDEOS) FROM THE LEADERBOARD; CF-NET ON 11TH ABAW TEST (152 VIDEOS).

| Method | Val F1↑ | Test F1↑ |
|---|---|---|
| Organiser baseline [1] | – | 0.2827 |
| Team Time Visao [25] | 0.652 | 0.536‡ |
| Team Lenovo PCIE [26] | – | 0.675‡ |
| Team LEYA [5] | **0.830** | 0.714‡ |
| Team Fennec [4] | 0.746 | 0.715‡ |
| Team BROTHER [10] | 0.736 | 0.727‡ |
| CF-Net (ours) | 0.716 | **0.7364** |

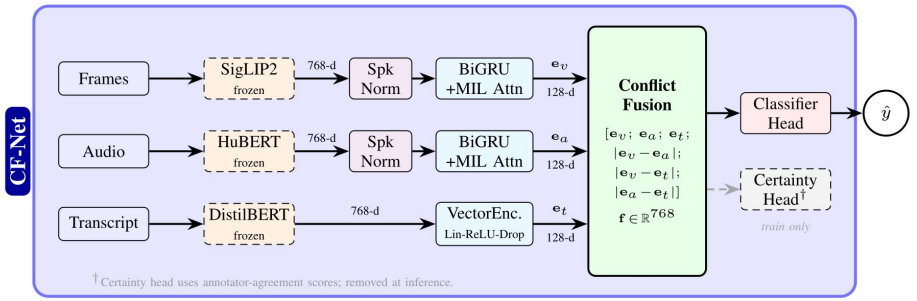



Fig. 1. Overview of CF-Net. Frozen SigLIP2, HuBERT, and DistilBERT backbones extract 768-d features per modality. Visual and audio features are speaker-normalised (Spk Norm) before being encoded by independent BiGRU encoders with MIL attention into 128-d embeddings $\mathbf{e}_v, \mathbf{e}_a$; the transcript embedding $\mathbf{e}_t$ is projected by a VectorEncoder. ConflictFusion concatenates the three embeddings with all pairwise absolute differences, producing $\mathbf{f}\in\mathbb{R}^{768}$. A classifier head produces the AH probability $\hat{y}$; an auxiliary certainty head (†, train only) regresses annotator-agreement scores to regularise learning on ambiguous samples.

TABLE II

INCREMENTAL ABLATION ON BAH VALIDATION SET, ADDING ONE COMPONENT AT A TIME FROM THE TOP DOWN. UNIMODAL BASELINES USE A BILSTM TEMPORAL ENCODER; THE SWITCH TO BIGRU IS ITS OWN ABLATION STEP. MACRO F1↑; BEST IN BOLD.

| Configuration | Val Macro F1↑ |
|---|---|
| Text only (DistilBERT + classifier) | 0.6505 |
| Audio only (HuBERT + BiLSTM) | 0.4672 |
| Visual only (SigLIP2 + BiLSTM) | 0.5888 |
| All modalities, mean fusion | 0.6428 |
| All modalities, ConflictFusion (no aux head) | 0.6612 |
| + Auxiliary certainty head | 0.6722 |
| + Focal loss + Manifold mixup ($\alpha = 0.2$) | 0.6870 |
| + Modality dropout ($p = 0.15$) | 0.6984 |
| + Speaker normalisation | 0.7110 |
| + BiGRU encoder | 0.7136 |
| + Certainty-weighted focal loss | **0.7155** |

# Team 8: IUSD

[Paper][Code]

- Multimodal framework (visual, audio, transcript) on top of frozen encoders
- Classical classifiers are considered: MLP, Random, Forest, GBDT
- Cross-attention and gated-fusion is used

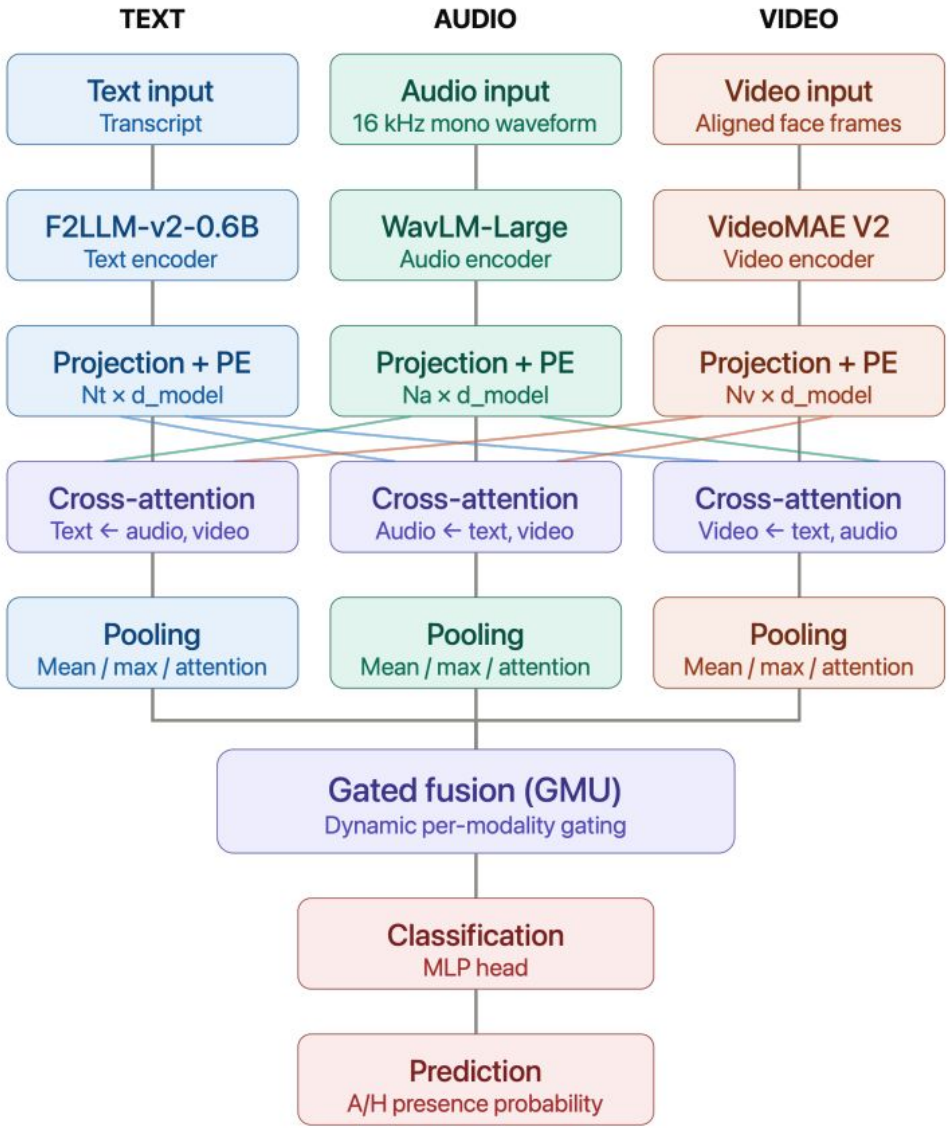



**Fig. 1.** Overview of the proposed multimodal recognition pipeline, integrating modality-specific feature projection and positional encoding, bidirectional interactions through cross-attention across all modality pairs, followed by pooling, gated multimodal fusion, and classification using an MLP head

# Team 8: IUSD

[Paper][Code]

**Table 1.** Unimodal classification results on the BAH test set (525 videos). Best per modality in **bold**. Macro F1 is the official challenge metric.

| Modality | Classifier | Macro F1 | $F1_0$ | $F1_1$ |
|---|---|---|---|---|
| Text (F2LLM-v2-0.6B) | MLP | 0.6550 | 0.6096 | 0.7003 |
| | RF | **0.6659** | 0.6360 | 0.6958 |
| | GBDT | 0.6505 | 0.5959 | 0.7051 |
| Audio (WavLM-Large) | MLP | 0.5946 | 0.5336 | 0.6556 |
| | RF | **0.6275** | 0.5494 | 0.7055 |
| | GBDT | 0.5717 | 0.5089 | 0.6346 |
| Visual (VideoMAE V2) | MLP | **0.5727** | 0.5206 | 0.6248 |
| | RF | 0.5364 | 0.4905 | 0.5823 |
| | GBDT | 0.5368 | 0.4926 | 0.5809 |
| Zero-shot | Video-LLaVA [1] | 0.2827 | – | – |

**Table 2.** Multimodal fusion results on the BAH validation set (124 videos), compared to the best unimodal baseline and the zero-shot baseline.

| Model | Macro F1 | Accuracy |
|---|---|---|
| Zero-shot Video-LLaVA [1] | 0.2827 | – |
| Best Unimodal (Text, RF) | 0.6659 | – |
| **Cross-Attention + GMU (Ours)** | **0.7394** | **0.7500** |

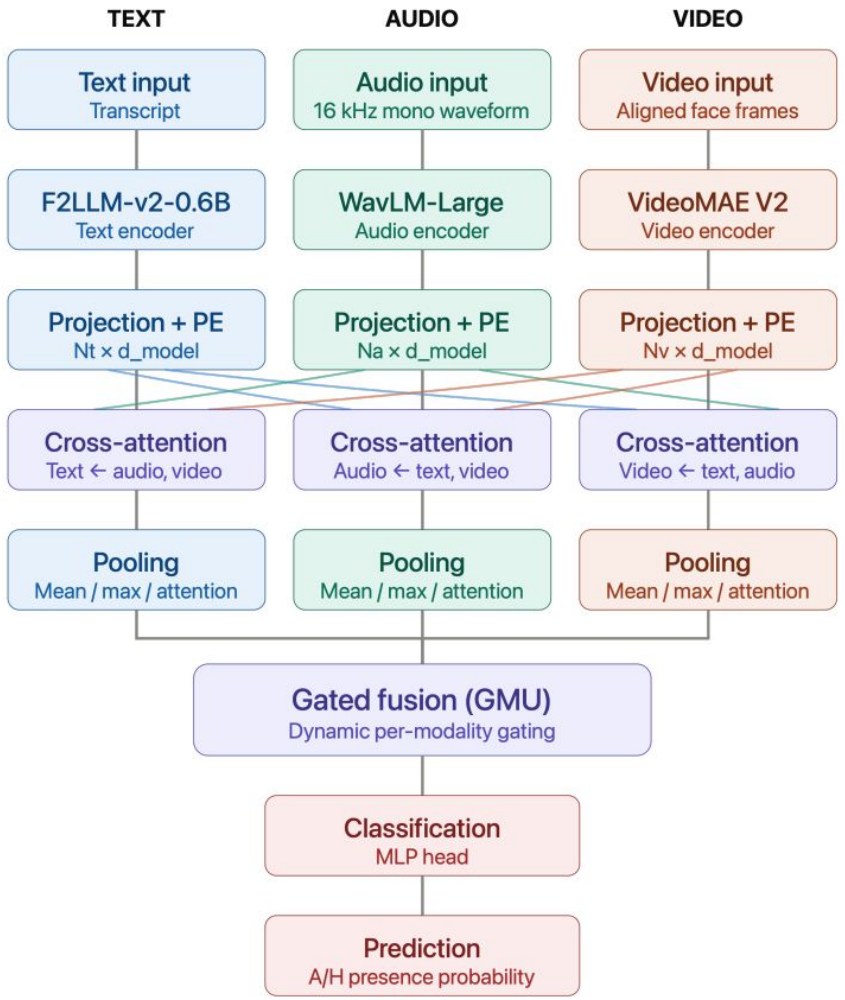



**Fig. 1.** Overview of the proposed multimodal recognition pipeline, integrating modality-specific feature projection and positional encoding, bidirectional interactions through cross-attention across all modality pairs, followed by pooling, gated multimodal fusion, and classification using an MLP head

# Team 9: MINDH Lab

[Paper][Code]

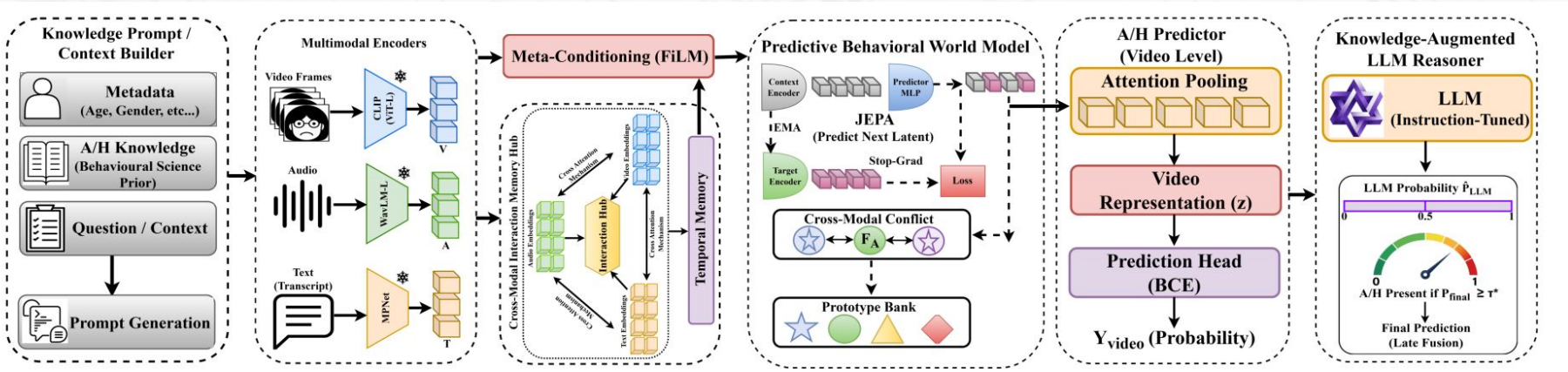



**Fig. 1:** Overview of PRISM-AH. Frozen encoders produce time-aligned window features that are conditioned on participant metadata through feature-wise modulation. A cross-modal interaction hub with temporal memory couples the channels, a predictive component exposes a hesitation surprise signal, an explicit conflict measure scores cross-channel disagreement, and a prototype bank matches recurring patterns. An attention-pooled head produces the calibrated video-level probability, and the same evidence is distilled into a textual summary. A knowledge-guided language model reasons over the summary and contributes a second opinion through late fusion.

- Propose: PRISM-AH (Predictive Reasoning over Interacting Streams for Multimodal Ambivalence / Hesitancy recognition)
- Use late fusion between a classifier and an LLM which reasons over textual summary of compact representation of a video
- Cross-modal interaction, temporal memory, and JEPA are used to construct video parts representations

# Team 9: MINDH Lab

[Paper][Code]

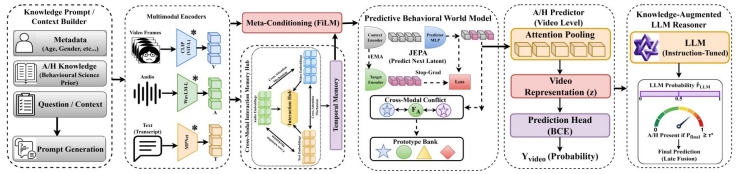



**Fig. 1:** Overview of PRISM-AH. Frozen encoders produce time-aligned window features that are conditioned on participant metadata through feature-wise modulation. A cross-modal interaction hub with temporal memory couples the channels, a predictive component exposes a hesitation surprise signal, an explicit conflict measure scores cross-channel disagreement, and a prototype bank matches recurring patterns. An attention-pooled head produces the calibrated video-level probability, and the same evidence is distilled into a textual summary. A knowledge-guided language model reasons over the summary and contributes a second opinion through late fusion.

**Table 2:** Main results. Macro F1 is the primary metric. The baseline is the reported zero-shot multimodal LLM [3]. The private test value is the organiser-reported leaderboard score of the submitted approach.

| Method | Public test mF1 | Private test mF1 |
|---|---|---|
| Zero-shot baseline [3] | 0.2827 | n/a |
| PRISM-AH (discriminative) | 0.5790 | n/a |
| **PRISM-AH + knowledge-guided reasoning** | **0.6133** | **0.6324** |

**Table 3:** Reasoning model sweep on the public test partition, on a fixed discriminative checkpoint whose base macro F1 is 0.5825. The fusion weight is selected on the validation partition, and a weight of zero indicates that fusion was not adopted.

| Reasoning model | Public test macro F1 | Fusion weight |
|---|---|---|
| None (discriminative only) | 0.5825 | n/a |
| Qwen2.5-7B [12] | 0.5825 | 0.0 |
| Mistral-7B [4] | 0.5934 | 0.1 |
| **Qwen2.5-14B [12]** | **0.6070** | 0.3 |

**Table 4:** Prompt ablation on the public test partition, with the strongest reasoner. Removal of the encoded expert cue taxonomy eliminates the gain, which identifies the domain knowledge as the essential ingredient.

| Prompt configuration | Public test macro F1 |
|---|---|
| Discriminative only | 0.5825 |
| Codebook knowledge and structured evidence | 0.6070 |
| Without codebook knowledge | 0.5825 |
| Without structured evidence | 0.6147 |

**Table 6:** Modality contribution over three seeds on the public test partition. Single-modality and modality-pair models isolate each contribution. Values are mean ± standard deviation.

| Configuration | Val macro F1 | Public test macro F1 |
|---|---|---|
| Vision only | 0.5850 ± 0.0069 | 0.5570 ± 0.0232 |
| Audio only | 0.5763 ± 0.0183 | 0.5563 ± 0.0285 |
| Text only | 0.5906 ± 0.0101 | 0.5514 ± 0.0479 |
| Vision and audio | 0.5902 ± 0.0178 | 0.5477 ± 0.0257 |
| Vision and text | 0.5917 ± 0.0056 | 0.5593 ± 0.0154 |
| Audio and text | 0.5891 ± 0.0112 | 0.5491 ± 0.0430 |
| All three | **0.6096 ± 0.0171** | **0.5570 ± 0.0130** |

# Team 10: HSEmotion

[Paper][Code]

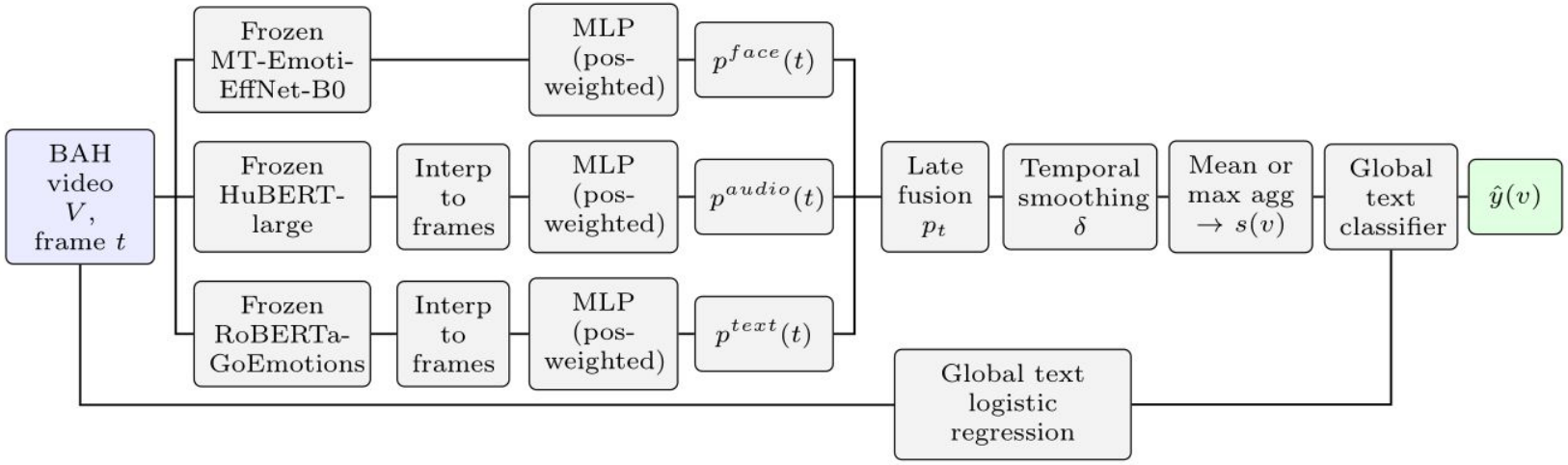



**Fig. 2:** Proposed A/H pipeline: frozen face, audio, and text descriptors; frame-level MLPs; late fusion; temporal aggregation; optional global-text gate for video-level Macro F1.

- Multimodal framework (face, audio, transcript) with late fusion

# Team 10: HSEmotion

[Paper][Code]

**Table 4:** Reported video-level Macro F1 on BAH from prior ABAW challenge publications.

| Method | Macro F1 |
|---|---|
| ABAW-10 baseline [22] | 0.343 |
| Lenovo PCIE [22] | 0.675 |
| LEYA [43] | 0.714 |
| Fennec [2] | 0.715 |
| VisPBF [39] | 0.727 |
| ABAW-11 baseline (Video-LLaVA zero-shot) [27] | 0.283 |
| Ours | **0.731** |

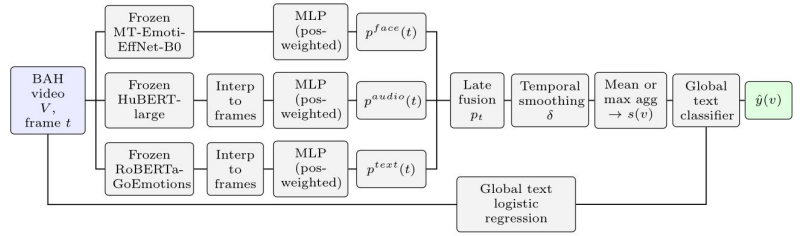



**Fig. 2:** Proposed A/H pipeline: frozen face, audio, and text descriptors; frame-level MLPs; late fusion; temporal aggregation; optional global-text gate for video-level Macro F1.

**Table 5:** Frame-level ablation of the proposed A/H pipeline on ABAW-11 validation ($\tau = 0.5$ unless noted).

| Configuration | Weighted F1 | Macro F1 |
|---|---|---|
| Face MLP only | 0.71 | 0.52 |
| Audio MLP only | 0.67 | 0.53 |
| Text MLP only | 0.77 | 0.59 |
| Early fusion MLP ($\tau = 0.4$) | 0.76 | 0.60 |
| Late fusion ($\tau = 0.55$) | 0.79 | 0.59 |

**Table 6:** Video-level ablation of the proposed A/H pipeline on ABAW-11 validation (124 videos) and the public test split (525 videos). The validation majority baseline always predicts no A/H.

| Configuration | Split | Macro F1 | AP |
|---|---|---|---|
| Majority class | val | 0.28 | 0.60 |
| RandomForest on concatenated features | val | 0.67 | 0.82 |
| Early fusion MLP | val | 0.68 | 0.85 |
| Late fusion, mean aggregation | val | 0.72 | 0.85 |
| Late fusion, max aggregation | val | 0.71 | 0.83 |
| Late fusion, mean | public | 0.69 | 0.79 |
| Late fusion, mean + hard global-text gate | public | 0.71 | 0.80 |
| Late fusion, max | public | 0.69 | 0.82 |
| Late fusion, max + hard global-text gate | public | 0.73 | 0.82 |
| Global text classifier only | public | 0.73 | 0.87 |

# Team 11: GGL

[Paper][Code]

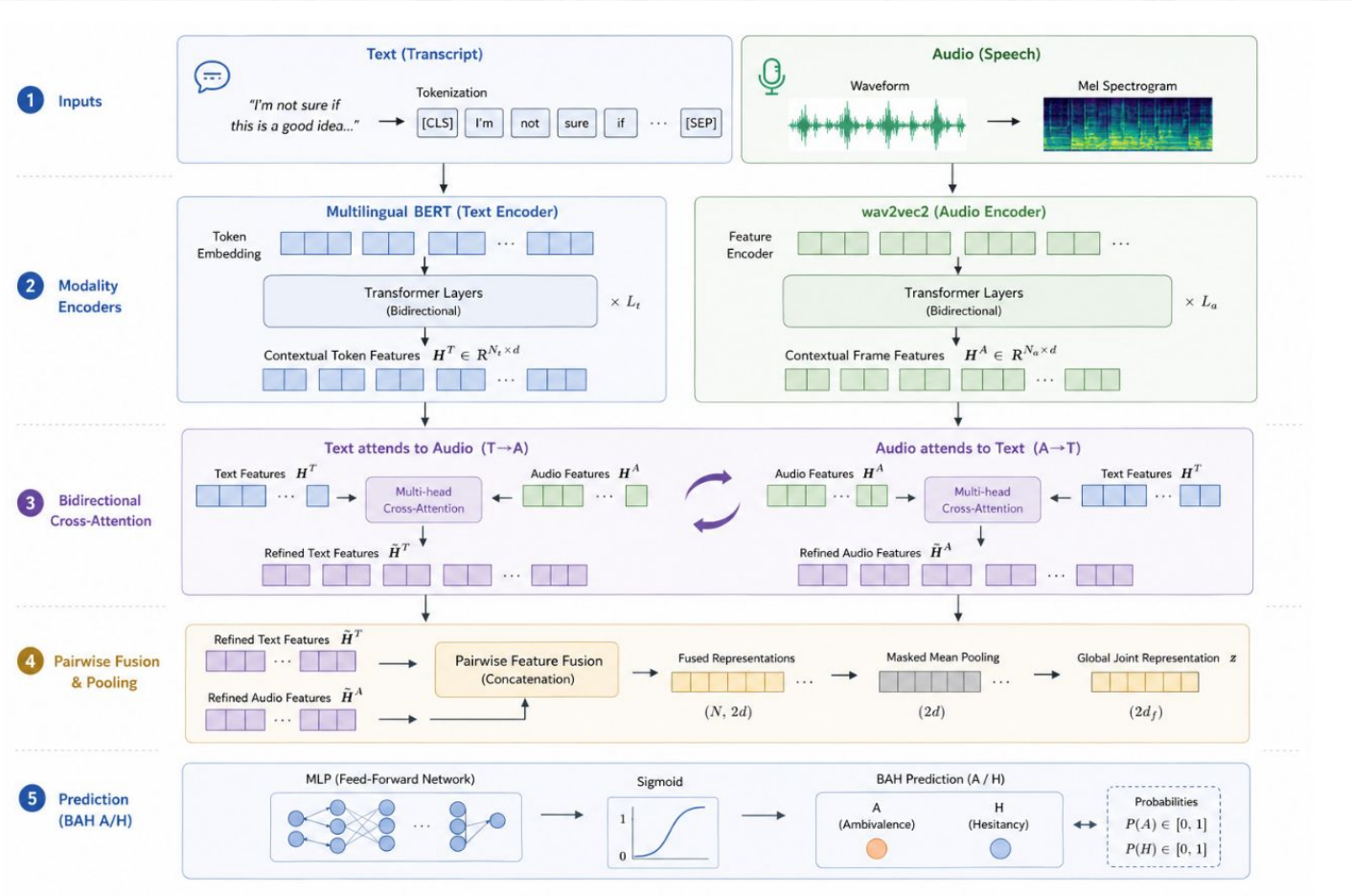



Figure 1: Overview of the proposed audio–text framework for video-level A/H recognition. The transcript and waveform are encoded by multilingual BERT and wav2vec 2.0, projected into a shared space, and exchanged through stacked bidirectional cross-attention blocks. Masked mean pooling and interaction-based fusion produce the final binary prediction.

- Audio-text framework with bidirectional cross-attention

# Team 11: GGL

[Paper][Code]

Table 1: Video-level performance on the BAH test set. Precision, recall, and F1-score are computed with A/H as the positive class. All values are percentages.

| Accuracy | Precision | Recall | F1-score | Macro-F1 |
|---|---|---|---|---|
| 72.19 | 75.00 | 81.13 | 77.95 | 70.16 |

Table 2 presents the corresponding confusion matrix. The model correctly identifies 258 of the 318 A/H videos and 121 of the 207 No-A/H videos. Its recall is higher for A/H than for No A/H, indicating that the model is more sensitive to the presence of A/H but produces more false positives for the No-A/H class.

Table 2: Confusion matrix on the BAH test set.

| Ground truth | Predicted No A/H | Predicted A/H |
|---|---|---|
| No A/H | 121 | 86 |
| A/H | 60 | 258 |

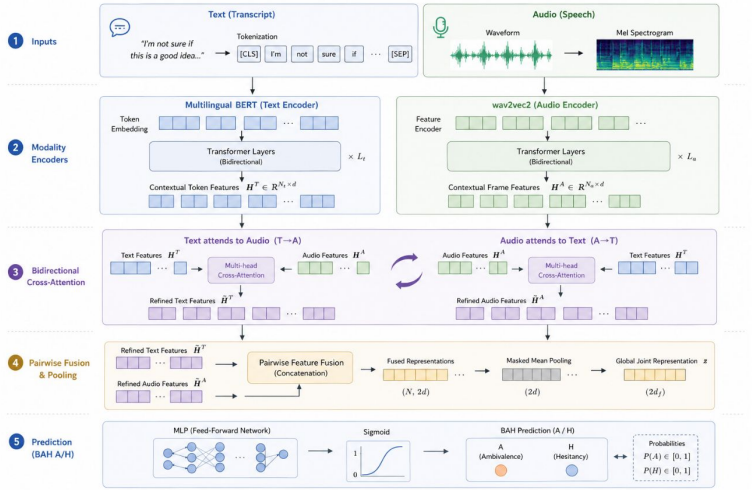



Figure 1: Overview of the proposed audio–text framework for video-level A/H recognition. The transcript and waveform are encoded by multilingual BERT and wav2vec 2.0, projected into a shared space, and exchanged through stacked bidirectional cross-attention blocks. Masked mean pooling and interaction-based fusion produce the final binary prediction.

# Team 12: AVULAB

[Paper][Code]

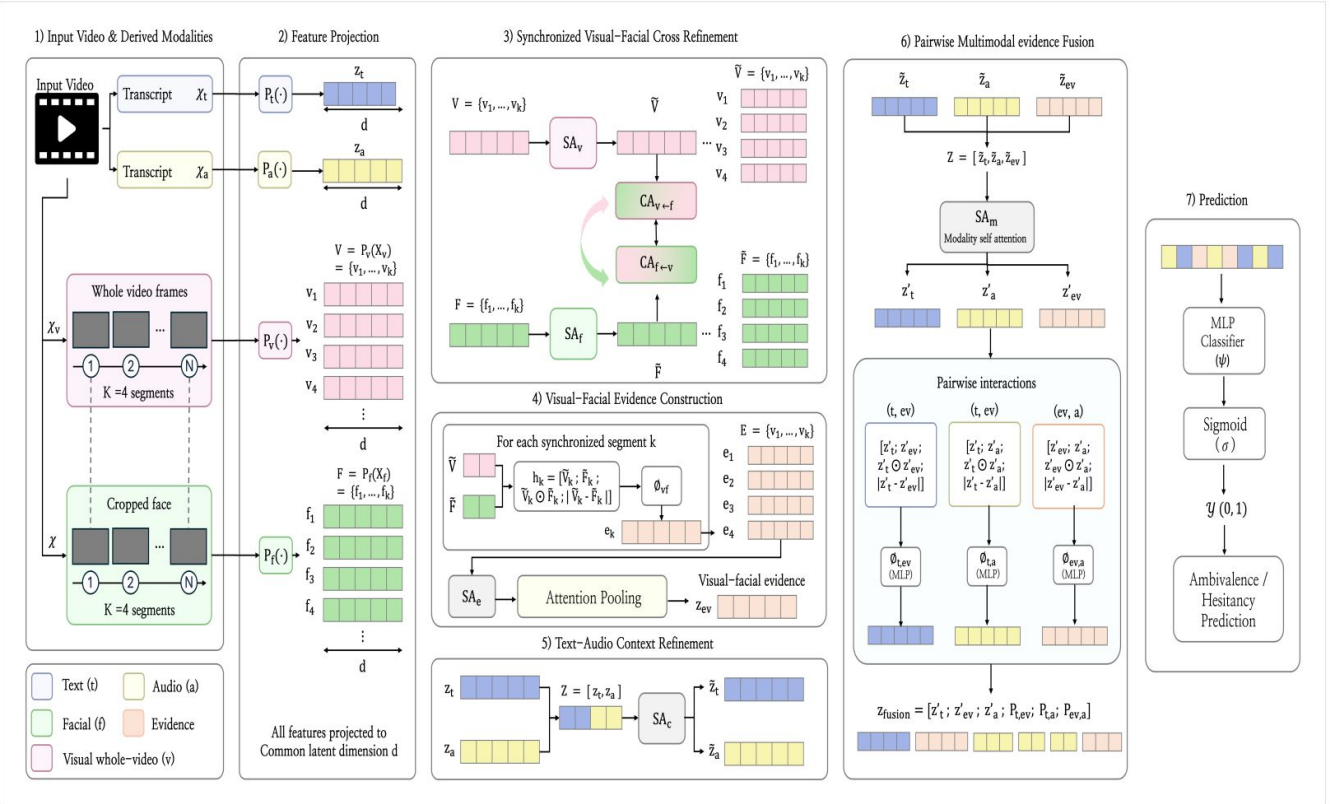



Figure 1: Overview of the proposed Synchronized Visual-Facial Cross-Refinement (SVF-CR) framework. The model extracts textual, whole-video visual, cropped-face facial, and acoustic features from an input video. Whole-video and facial segment tokens are temporally synchronized and refined through intra-modal self-attention and bidirectional visual-facial cross-attention. The refined tokens are then used to construct visual-facial evidence, which is finally fused with textual and acoustic evidence through pairwise multimodal evidence fusion.

- Multimodal framework (face, frame, audio, transcript)
- Face and frame tokens are synchronized

# Team 12: AVULAB

[Paper][Code]

Table 3: Ablation of synchronized visual-facial evidence modeling modules. All variants use text, enhanced visual evidence, and audio.

| Variant | BACC | MF1 | AUC | AUPRC |
|---|---|---|---|---|
| Visual pooling | 0.6848 | 0.6796 | 0.7663 | 0.8313 |
| Face pooling | 0.6955 | 0.6992 | 0.7757 | **0.8476** |
| Anchor-based fusion | 0.6961 | 0.6799 | 0.7595 | 0.8256 |
| Text-conditioned anchor fusion | 0.6883 | 0.6800 | 0.7635 | 0.8291 |
| Text-conditioned consistency-discrepancy evidence | 0.6878 | 0.6902 | 0.7630 | 0.8329 |
| Sync. consistency-discrepancy evidence | 0.7106 | 0.7099 | 0.7802 | 0.8463 |
| SVF-CR w/o VF cross-attention | 0.6926 | 0.6980 | **0.7825** | 0.8436 |
| SVF-CR w/o reverse refinement | 0.6789 | 0.6814 | 0.7617 | 0.8336 |
| SVF-CR with text pooling | 0.6932 | 0.6964 | 0.7770 | 0.8417 |
| Contextual SVF-CR | 0.6955 | 0.6944 | 0.7598 | 0.8287 |
| **SVF-CR** | **0.7179** | **0.7156** | 0.7794 | 0.8387 |

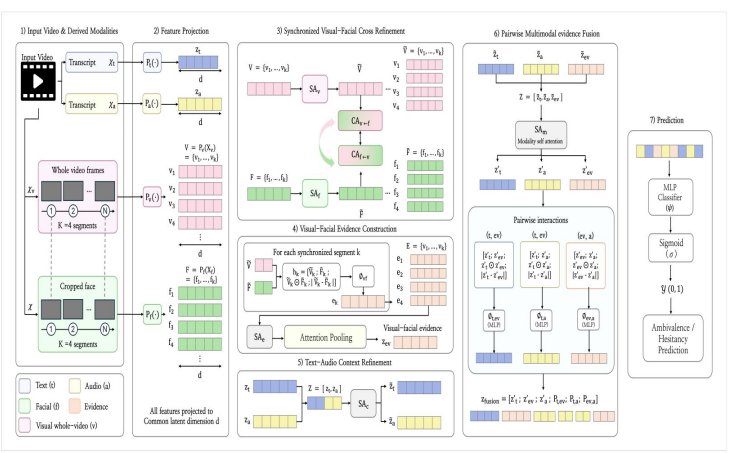



Figure 1: Overview of the proposed Synchronized Visual-Facial Cross-Refinement (SVF-CR) framework. The model extracts textual, whole-video visual, cropped-face facial, and acoustic features from an input video. Whole-video and facial segment tokens are temporally synchronized and refined through intra-modal self-attention and bidirectional visual-facial cross-attention. The refined tokens are then used to construct visual-facial evidence, which is finally fused with textual and acoustic evidence through pairwise multimodal evidence fusion.

Table 1: Main public evaluation results.

| Method | ACC | BACC | MF1 | AUC | AUPRC |
|---|---|---|---|---|---|
| Global token-cross | 0.7219 | 0.7097 | 0.7094 | 0.7660 | 0.8336 |
| **SVF-CR** | **0.7257** | **0.7179** | **0.7156** | **0.7794** | **0.8387** |

Table 2: Effect of modality combinations using the proposed SVF-CR framework.

| Modalities | BACC | MF1 | AUC | AUPRC |
|---|---|---|---|---|
| Text | 0.6867 | 0.6923 | 0.7779 | **0.8462** |
| eVisual | 0.5861 | 0.5817 | 0.5978 | 0.6772 |
| Audio | 0.6972 | 0.7005 | 0.7690 | 0.8387 |
| Text + eVisual | 0.6913 | 0.6919 | 0.7706 | 0.8437 |
| Text + Audio | 0.7012 | 0.7076 | **0.7813** | 0.8461 |
| eVisual + Audio | 0.7032 | 0.7039 | 0.7676 | 0.8395 |
| **Text + eVisual + Audio** | **0.7179** | **0.7156** | 0.7794 | 0.8387 |

# Leaderboard: Performance on the challenge private test set

[Link]

| Rank | Teams | AVGF1 (Macro F1) | Github | arXiv |
|---|---|---|---|---|
| 1 | AffectVision | 0.7959 | link | link |
| 2 | RAS | 0.7824 | link | link |
| 3 | PUCPR | 0.7455 | link | link |
| 4 | AIM for Emo | 0.7431 | link | link |
| 5 | AIWELL | 0.7367 | link | link |
| 5 | CASIA-26 | 0.7367 | link | link |
| 6 | Tamimi | 0.7364 | link | link |
| 7 | CPR | 0.7167 | link | link |
| 8 | IUSD | 0.6841 | link | link |
| 9 | MINDH Lab | 0.6324 | link | link |
| 10 | HSEmotion | 0.5623 | link | link |
| 11 | GGL | 0.5136 | link | link |
| 12 | AVULAB | 0.4858 | link | link |
| | Baseline | 0.3448 | – | – |

[ABAW11 Leaderboards]

# Leaderboard: Previous challenges

- ABAW 10th, AH 2nd, CVPR 2026: [link]