Ambivalence/Hesitancy (AH) Video Recognition Challenge, 2nd edition, ABAW10th - CVPR 2026

Task: Video classification

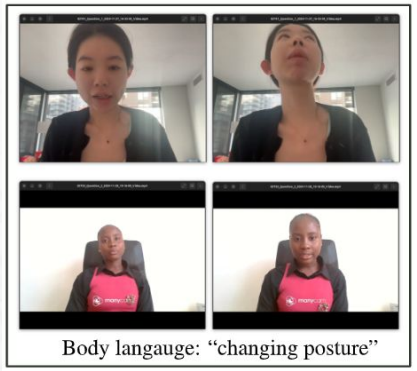
Jan-Mar 2026

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

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:	January 31, 2026
Test set release:	March 9, 2026
Final submission deadline (Predictions, Code and ArXiv paper):	March 15, 2026
Winners Announcement:	March 17, 2026
Final Paper Submission Deadline:	March 18, 2026
Review decisions sent to authors; Notification of acceptance:	April 7, 2026
Camera ready version:	April 10, 2026

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

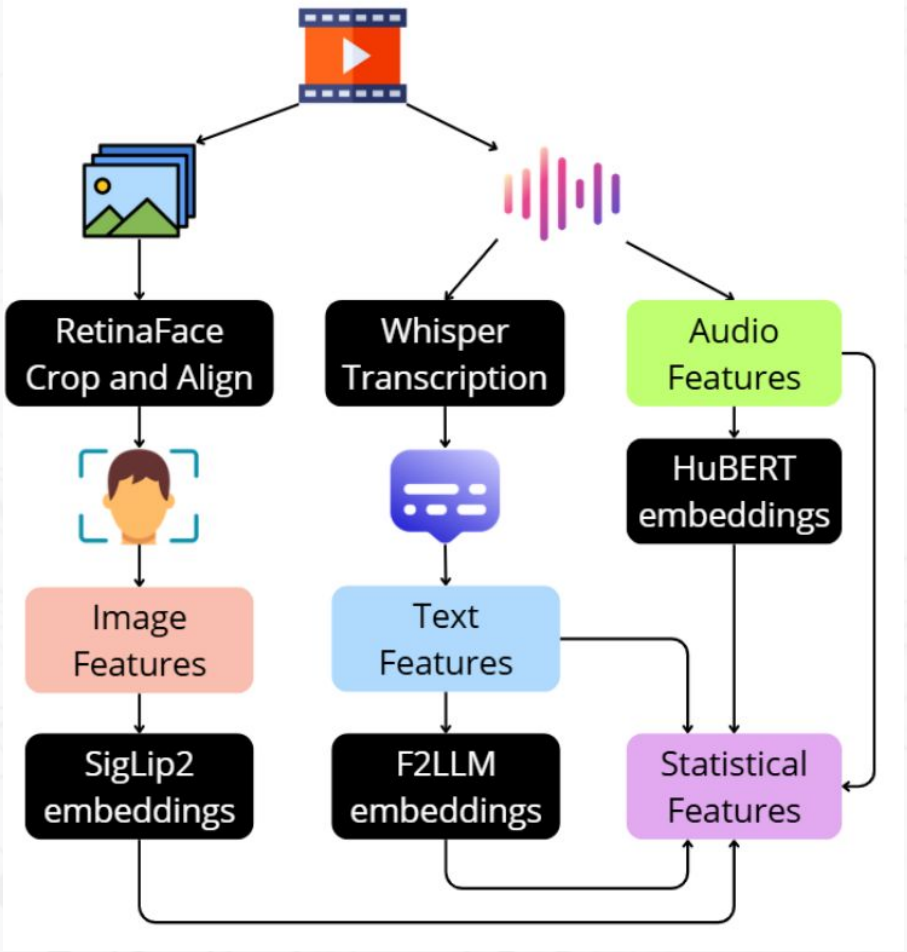
Task Ambivalence/Hesitancy (AH) Video Recognition Challenge

Teams	Method
VisPBF	Standard multimodal model w/ensembling
Fennec	Cross-modality features difference
LEYA	Standard multimodal model
Lenovo PCIE	Finetune MLLM adapted to short segments
Time Visão	Cross-modality features difference

- 9 teams registered
- 5 teams submitted predictions
- All 5 papers are specialized for ambivalence/hesitancy recognition
- Link to papers
- Link to leaderboard

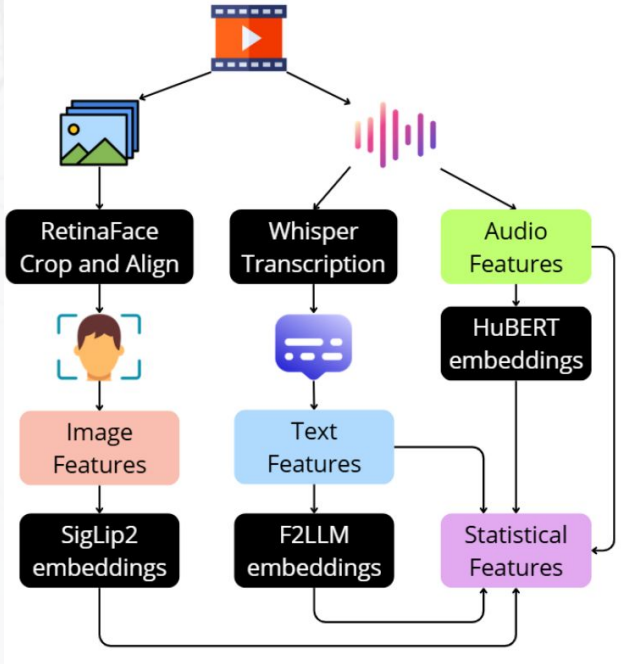
Methods

Methods: VisPDF team



- Standard multimodal model: visual, audio, transcript, handcraft-statistics (structured aggregation of temporal data)
- Ensembling-based
- Test 3 classifiers: MLP, random forest, gradient boosted decision tree
- Best model across 15 combinations of modalities was selected using lowest binary-cross entropy over validset
- Hard voting over 15 models using Particle Swarm Optimization (PSO)

Methods: VisPDF team



- Standard multimodal model: visual, audio, transcript, handcraft-statistics (structured aggregation of temporal data)
- Ensembling-based
- Test 3 classifiers: MLP, random forest, gradient boosted decision tree
- Best model across 15 combinations of modalities was selected using lowest binary-cross entropy over validset
- Hard voting over 15 models using Particle Swarm Optimization (PSO)

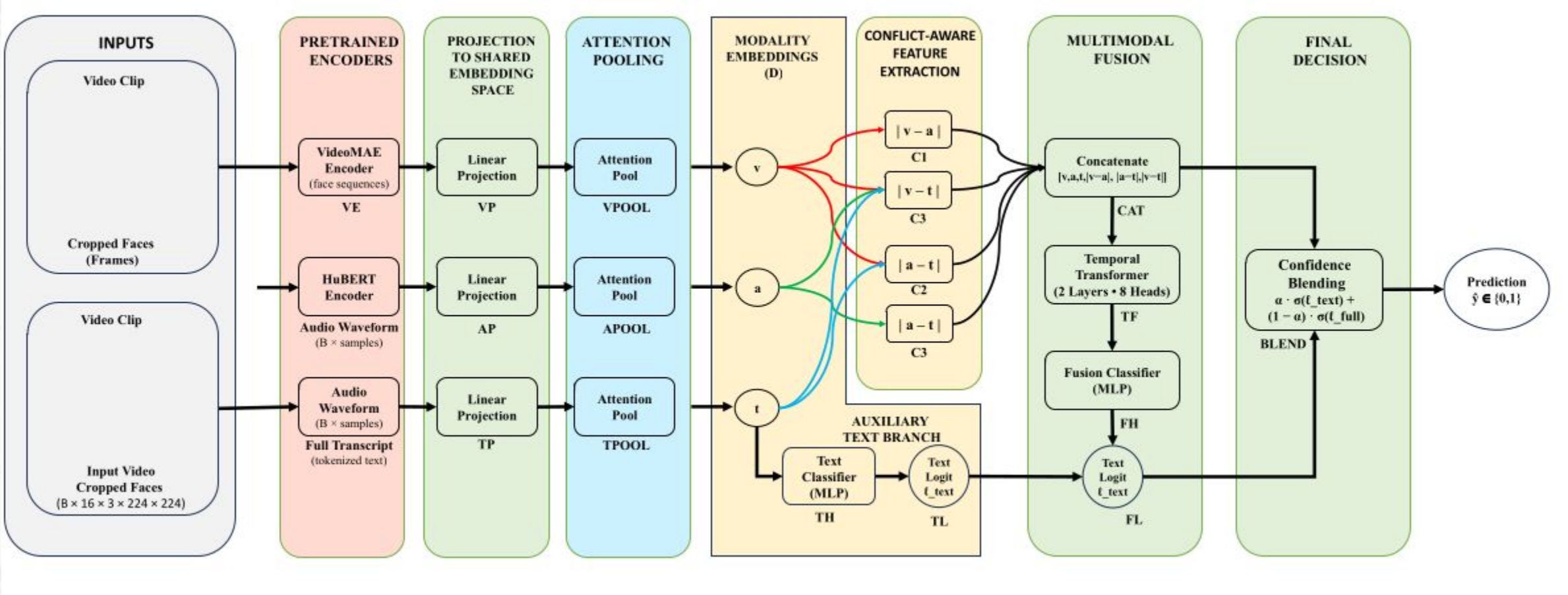
Modality Combination	MLP		RF		GBDT		Winner
	BCE	F1	BCE	F1	BCE	F1	(Best BCE)
Text	**0.5730**	0.7275	0.6234	0.6613	0.6309	0.6779	MLP
Audio	**0.6751**	0.6317	0.6950	0.5992	0.6922	0.5973	MLP
Video	0.7465	0.5226	0.6963	0.4638	**0.6956**	0.4704	GBDT
Stats	0.6496	0.6928	**0.6341**	0.6406	0.6403	0.6317	RF
Text+Audio	**0.5925**	0.7014	0.6322	0.6409	0.6393	0.6544	MLP
Text+Video	0.6884	0.5363	**0.6242**	0.6686	0.6316	0.6693	RF
Text+Stats	**0.5937**	0.6512	0.6203	0.6935	0.6294	0.6773	MLP
Audio+Video	0.7170	0.4959	0.6936	0.5989	**0.6921**	0.5921	GBDT
Audio+Stats	**0.6612**	0.6282	0.6698	0.6261	0.6679	0.6232	MLP
Video+Stats	0.7283	0.6228	**0.6593**	0.6521	0.6621	0.6494	RF
Text+Aud+Vid	0.6872	0.5008	**0.6349**	0.6544	0.6414	0.6529	RF
Text+Aud+Sts	**0.6001**	0.6314	0.6282	0.6601	0.6370	0.6574	MLP
Text+Vid+Sts	0.7334	0.5426	**0.6209**	0.6838	0.6302	0.6744	RF
Aud+Vid+Sts	0.6897	0.5377	**0.6732**	0.6288	0.6736	0.6220	RF
All Modalities	0.6962	0.5951	**0.6271**	0.6601	0.6362	0.6540	RF

Public BAH validation set

Penalty (λ)	Train F1-M	Val F1-M	Test F1-M
0.0 (0%)	0.9743	0.7355	0.7399
0.2 (20%)	**0.9820**	0.7355	**0.7465**
0.4 (40%)	0.9653	**0.7578**	0.7409
0.6 (60%)	0.9653	**0.7578**	0.7409
0.8 (80%)	0.9781	0.7486	0.7424

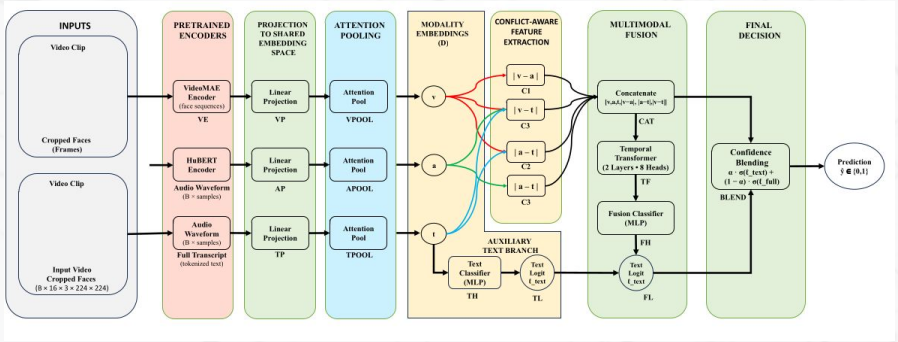
Public BAH test set (ensembling)

Methods: Fennec team



Model A/H via difference at feature level across modalities.

Methods: Fennec team

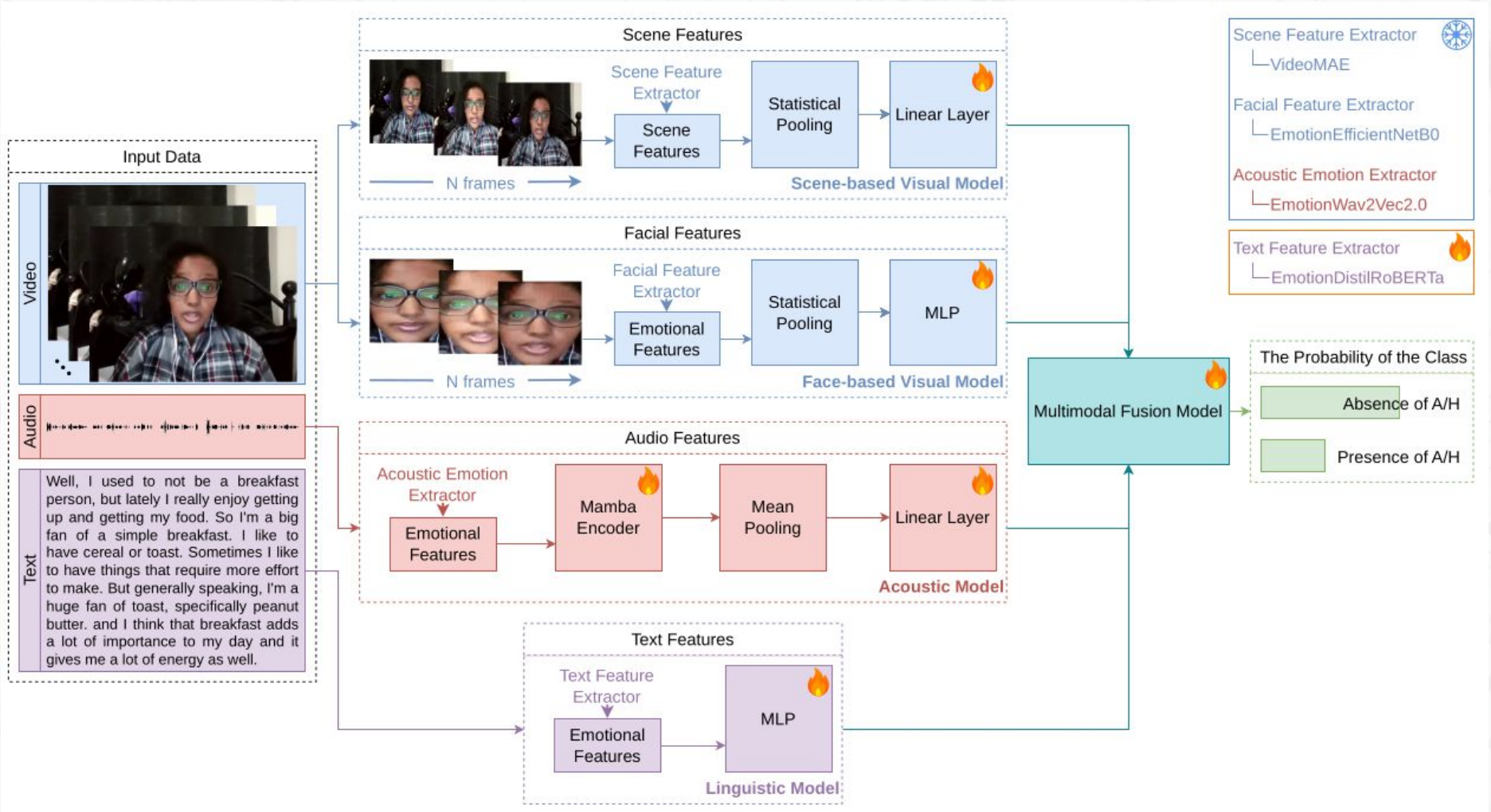


Method	Macro F1
BAH: Zero-shot M-LLM (vision only) [7]	0.283
BAH: Video-FocalNet base [7]	0.566
BAH: LFAN (V+A+T, co-attention) [7]	0.593
BAH: Zero-shot M-LLM + transcript [7]	0.634
Ours — single model (1 or 5 windows)	0.690–0.692
Ours — ensemble (2 ckpts × 5 win.)	**0.694**

Public BAH test set

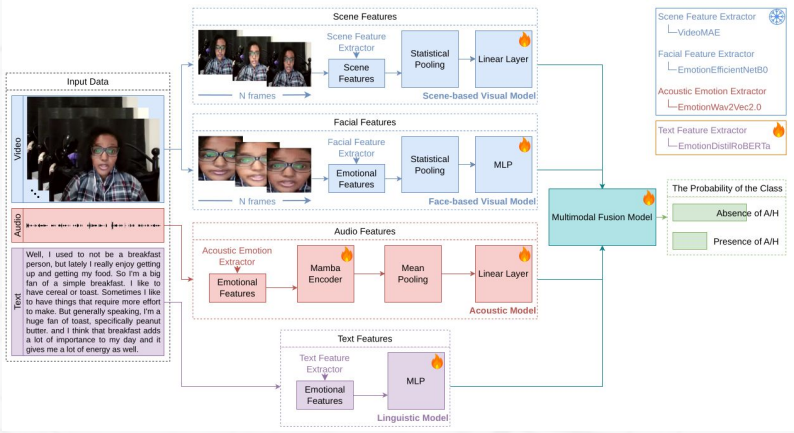
Model A/H via difference at feature level across modalities.

Methods: LEYA team



Standard multimodal framework: visual (face, and scene), audio, and transcript.

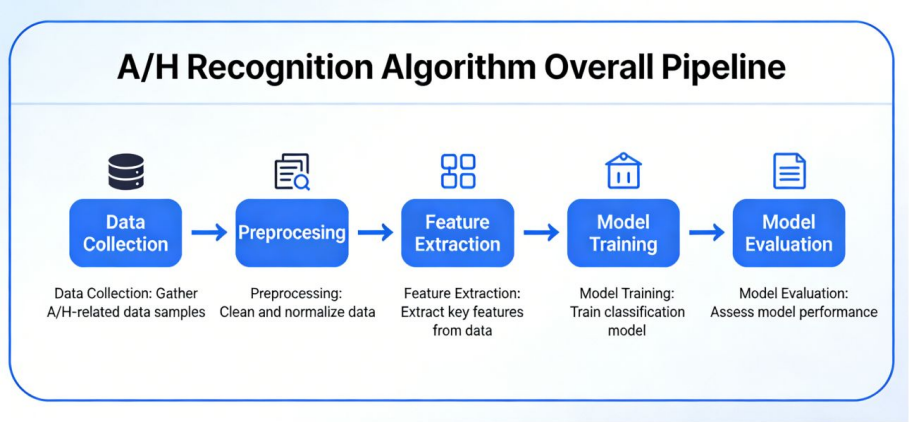
Methods: LEYA team



ID	Model Configuration			BAH sub-corpus			
	Modality	Features	Classifier	Devel. / Valid. (MF1, %)	Test (MF1, %)	Average (MF1, %)	Final test (MF1, %)
1	Face	EmotionEfficientNetB0 + Statistical Features	MLP	65.29	60.05	62.67	–
2	Scene	VideoMAE	Linear Layer	61.71	62.21	61.96	–
3	Audio	EmotionWav2Vec2.0 + Mamba	Linear Layer	67.20	70.87	69.03	–
4	Text	TF-IDF	Logistic Regression	68.30	67.75	68.03	–
5	Text	TF-IDF	CatBoost	65.56	72.02	68.79	–
6	Text	Fine-tuned EmotionTextClassifier	MLP	69.28	70.72	70.00	–
7	Text	Fine-tuned EmotionDistilRoBERTa	MLP	68.54	71.49	70.02	–
8	Models IDs 1, 2, 3 and 4	Multimodal Fusion Model	Linear Layer	80.79	77.03	78.91	–
9	Models IDs 1, 2, 3 and 5	Multimodal Fusion Model	Linear Layer	77.91	78.54	78.22	–
10	Models IDs 1, 2, 3 and 6	Multimodal Fusion Model	Linear Layer	78.35	77.03	77.69	–
11	Models IDs 1, 2, 3 and 7	Multimodal Fusion Model	Linear Layer	85.38	79.94	82.66	68.32
12	Models IDs 1, 2, 3 and 7	Multimodal Fusion Model with Prototype Head	Linear Layer	83.79	82.72	83.25	65.21
13	Models IDs 1, 2, 3 and 7	Ensemble of Five Multimodal Fusion Models	Linear Layer	81.94	80.64	81.29	70.17
14	Models IDs 1, 2, 3 and 7	Ensemble of Five Multimodal Fusion Models with Prototype Head	Linear Layer	83.00	80.77	81.89	71.43

Standard multimodal framework: visual (face, and scene), audio, and transcript.

Methods: Lenovo PCIE team



Initialize processed dataset $\mathcal{D} \leftarrow \emptyset$
for each video $V_i \in \mathcal{V}$ **do**
  Read annotation info a_i from $\mathcal{A}$
  if a_i indicates A/H-negative **then**
    Keep the whole video as one sample
    $\hat{V}_i \leftarrow V_i$
    Split $\hat{V}_i$ into clips of length at most Δ
    Assign label $y_i = 0$ to all clips
    Add all clips into $\mathcal{D}$
  else
    Obtain annotated A/H time intervals $\{(t_k^{start}, t_k^{end})\}_{k=1}^{K}$
    for each annotated interval (t_k^{start}, t_k^{end}) **do**
      Crop sub-video $\hat{V}_{ik}$ from V_i using (t_k^{start}, t_k^{end})
      Split $\hat{V}_{ik}$ into clips of length at most Δ
      Assign label $y_i = 1$ to all clips
      Add all clips into $\mathcal{D}$
    end for
  end if
end for
Construct multimodal instruction-tuning samples using video, audio, and task prompt
Fine-tune model M on $\mathcal{D}$

Train

for each test video V_j **do**
  Split V_j into clips $\{c_1, c_2, \ldots, c_N\}$
  for each clip c_n **do**
    Build multimodal input with video, audio, and prompt
    Use M to predict clip label $\hat{y}_n \in \{0, 1\}$
  end for
  Aggregate clip predictions:

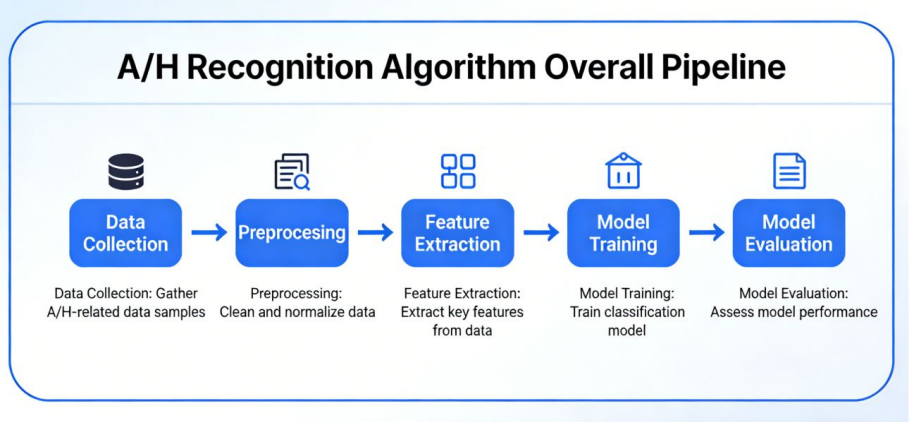
Test

$$\hat{y}_{video} = \max(\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_N)$$

end for
return Video-level predictions for all test videos

- MLLM (Qwen3-Omni-30B-A3B) finetuned (LoRA) on BAH using visual and audio modalities.
- Adapted to work on short segments (5s)
- Use text prompts (question). Binary answer: yes/no

Methods: Lenovo PCIE team



22: **for** each test video V_j **do**
23: Split V_j into clips $\{c_1, c_2, \ldots, c_N\}$
24: **for** each clip c_n **do**
25: Build multimodal input with video, audio, and prompt
26: Use M to predict clip label $\hat{y}_n \in \{0, 1\}$
27: **end for**
28: Aggregate clip predictions: Test

$$\hat{y}_{video} = \max(\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_N)$$

29: **end for**
30: **return** Video-level predictions for all test videos

Public BAH test set

Model Configuration	Accuracy
Qwen3-Omni-v1	81.9%
Qwen3-Omni-v2	79.8%
Qwen3-Omni-v3	65.3%
Qwen3-Omni (Majority Vote)	85.1%

- MLLM (Qwen3-Omni-30B-A3B) finetuned (LoRA) on BAH using visual and audio modalities.
- Adapted to work on short segments (5s)
- Use text prompts (question). Binary answer: yes/no

Methods: Time Visão team

3.2. Temporal Modeling and Fusion

Each temporal modality is processed by a 2-layer BiLSTM(hidden dim 64) with attention pooling, then projected to D=128 dimensions. We compare three fusion strategies:

Fusion A (Implicit): $\mathbf{f}_A = [\mathbf{h}'_v; \mathbf{h}'_a; \mathbf{h}'_t]$

Fusion B (Divergence): $\mathbf{f}_B = [|\mathbf{h}'_v - \mathbf{h}'_a|; |\mathbf{h}'_v - \mathbf{h}'_t|; |\mathbf{h}'_a - \mathbf{h}'_t|]$

Fusion C (Combined): $\mathbf{f}_C = [\mathbf{f}_A; \mathbf{f}_B]$

The fused vector is classified by a 3-layer MLP with dropout (p=0.3). Training uses BCEWithLogitsLoss with class weighting, AdamW with differentiated learning rates (5×10^{-5} for BERT, 5×10^{-4} for other parameters), cosine annealing over 30 epochs, gradient clipping at 1.0, and early stopping with patience 8.

- Pairwise difference between modalities features
- Visual (action units), audio, and transcript.

Methods: Time Visão team

3.2. Temporal Modeling and Fusion

Each temporal modality is processed by a 2-layer BiLSTM(hidden dim 64) with attention pooling, then projected to D=128 dimensions. We compare three fusion strategies:

Fusion A (Implicit): $\mathbf{f}_A = [\mathbf{h}'_v; \mathbf{h}'_a; \mathbf{h}'_t]$

Fusion B (Divergence): $\mathbf{f}_B = [|\mathbf{h}'_v - \mathbf{h}'_a|; |\mathbf{h}'_v - \mathbf{h}'_t|; |\mathbf{h}'_a - \mathbf{h}'_t|]$

Fusion C (Combined): $\mathbf{f}_C = [\mathbf{f}_A; \mathbf{f}_B]$

The fused vector is classified by a 3-layer MLP with dropout (p=0.3). Training uses BCEWithLogitsLoss with class weighting, AdamW with differentiated learning rates (5×10^{-5} for BERT, 5×10^{-4} for other parameters), cosine annealing over 30 epochs, gradient clipping at 1.0, and early stopping with patience 8.

Model	Val F1	Test F1
Unimodal:		
Visual AUs (XGBoost)	0.6194	0.5642
Audio Wav2Vec (LSTM)	0.5218	0.6141
Text BERT	0.5758	0.5904
Multimodal (raw AUs):		
Fusion A (implicit)	0.6788	0.6604
Fusion B (divergence)	0.6524	**0.6808**
Fusion C (combined)	0.6700	0.6766
Multimodal (windowed AUs):		
Fusion B (divergence)	0.6912	0.6602
Challenge baseline [7]	—	0.2827

Public BAH test set

- Pairwise difference between modalities features
- Visual (action units), audio, and transcript.

Leaderboard: Performance on the challenge private test set

Teams	AVGF1 (Macro F1)	Github	arXiv
VisPBF	0.7266	link	link
Fennec	0.7151	link	link
LEYA	0.7142	link	link
Lenovo PCIE	0.6748	link	link
Time Visão	0.5362	link	link
Baseline	0.3428	–	–