Audio–Text Cross-Attention with Psycholinguistic Support Features for Ambivalence/Hesitancy Recognition

Luiz F. B. F. Martins · Rodrigo W. Pisaia · Matheus M. Girardi · Isabella V. Berkembrock · João A. Almeida
Andre G. Hochuli · Rayson Laroca · Alceu S. Britto Jr.
Pontifícia Universidade Católica do Paraná (PUCPR/BRAZIL) · LIA – Artificial Intelligence Student League · ligaIA@ppgia.pucpr.br

11th ABAW Workshop @ ECCV 2026 | 3rd Ambivalence/Hesitancy Video Recognition Challenge

ECCV 2026 | MALMÖ | SEPT 8–12

PAPER & CODE

3rd PLACE
Official Challenge Leaderboard

1. MOTIVATION

Why A/H matters

Ambivalence/Hesitancy (A/H) reflects uncertainty, reduced commitment, or conflicting positive and negative evaluations during decision-making.

TEMPORALLY SPARSE | MULTIMODAL | BEHAVIORAL

- Pauses and delayed responses
- Slower speech, pitch variation, and intonation changes
- Hedging, self-corrections, and epistemic markers
- Simultaneous positive and negative statements

Approaches exploit multimodal representations:
- ▸ visual / facial behavior;
- ▸ speech / audio;
- ▸ textual information.

2. RESEARCH GAP & CONTRIBUTIONS

Research Question

Can attention-based temporal aggregation + explicit prosodic and psycholinguistic support improve audio–text A/H recognition over generic embeddings and uniform mean pooling?

Hypotheses

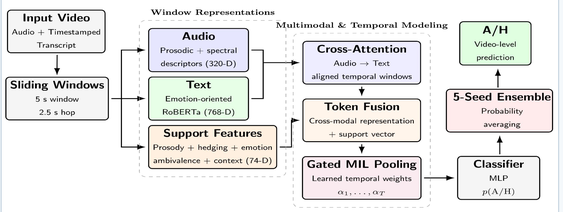
H1 Attention-based MIL improves temporal aggregation over mean pooling.

H2 Explicit psycholinguistic support complements generic audio–text embeddings.

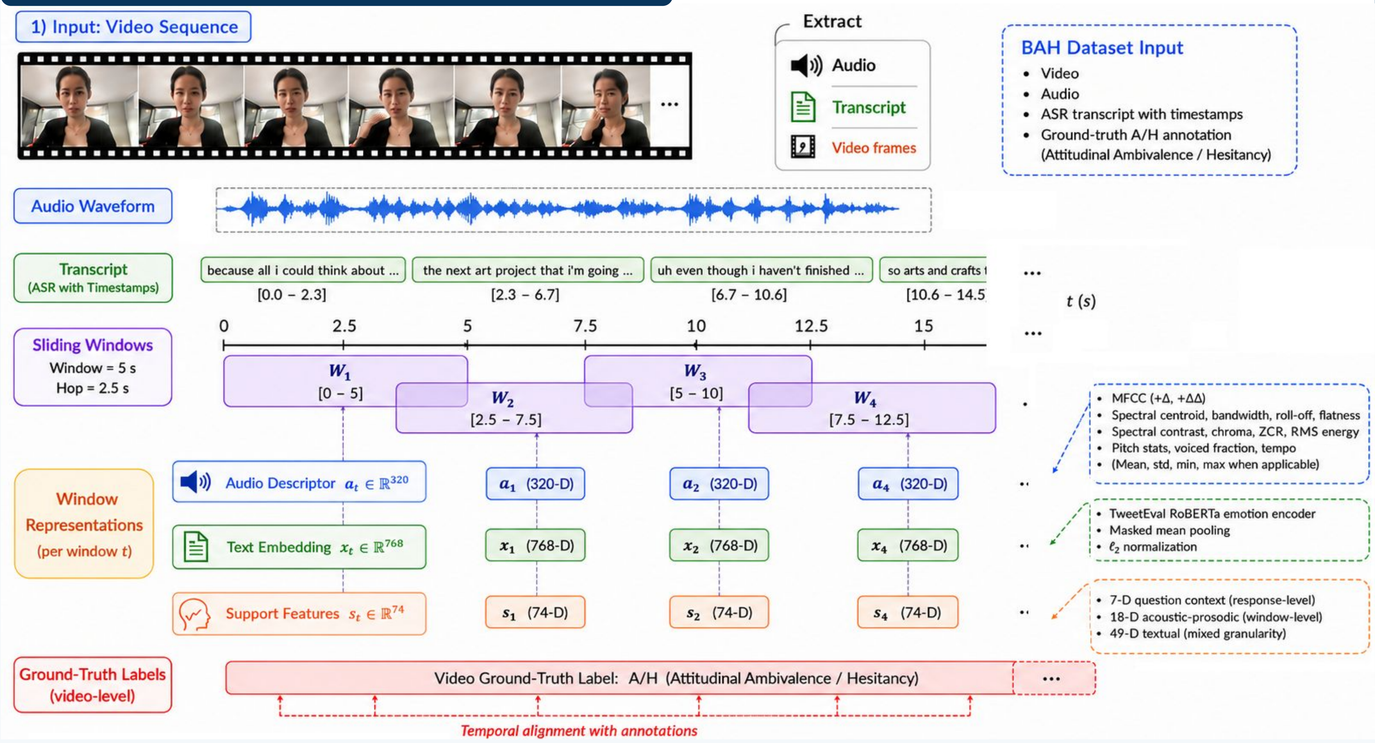
Contributions

- ✓ Audio→text cross-attention over aligned temporal windows.
- ✓ 74-D interpretable support vector for uncertainty, hedging, context, and ambivalence.
- ✓ Gated MIL pooling to learn which windows contribute most to video-level A/H.
- ✓ Ensemble Based Approach

3. METHOD OVERVIEW



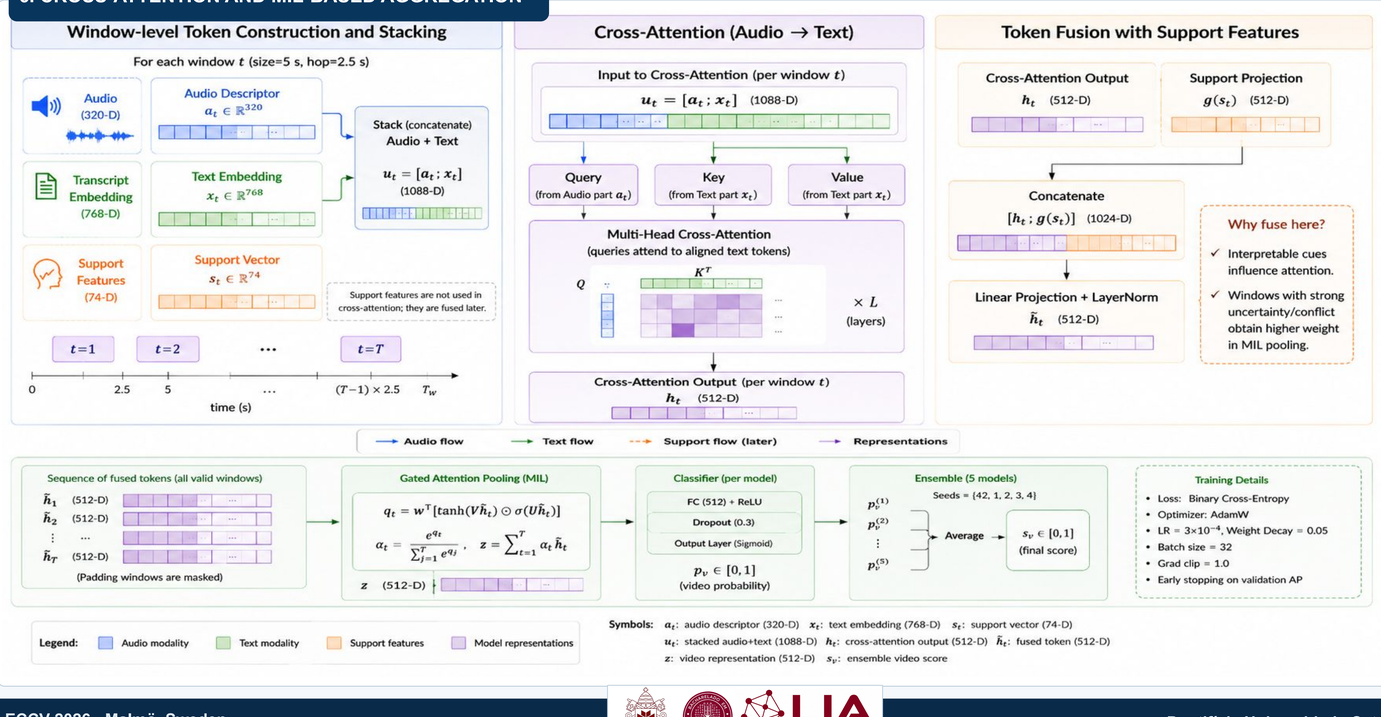
4. WINDOW-BASED REPRESENTATIONS AND TEMPORAL ALIGNMET



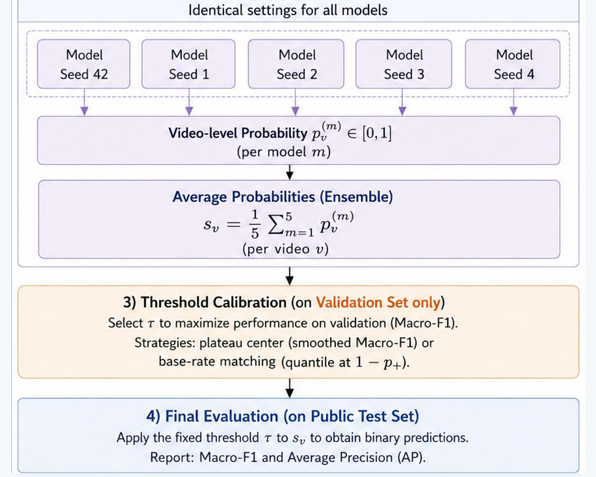
5. PSYCHOLINGUISTIC SUPPORT VECTOR (74-D)

Block	Feature group	Granularity	Dim.
Context	Question type	Response	7
Acoustic–prosodic	Timing and pauses	Window	6
	Speech rate and articulation	Window	3
	Pitch and intonation	Window	4
	Loudness	Window	2
	Voice quality and uncertainty score	Window	3
	Acoustic–prosodic subtotal		18
Textual	Ambivalence indices	Response	20
	Emotion dynamics	Mixed	6
	Contrast and polarity shift	Window	5
	Epistemic hedges	Window	18
	Textual subtotal		49
Total			**74**

6. CROSS-ATTENTION AND MIL-BASED AGGREGATION



7. EXPERIMENTAL PROTOCOL



8. RESULTS: FEATURE ABLATION

Method	Macro-F1	AP
Comparison methods		
ConflictAwareAH (Fennec)	0.694	N/A
Mean-pooling baseline	0.701	0.828
Proposed MIL variants		
Audio + text, no support	0.705	0.835
Audio support only	0.636	0.744
Text support only	0.698	0.842
All support, single seed	0.710	0.857
All support, 5 seeds	**0.722**	**0.875**

BAH public-test split: 525 labeled videos

- ▸ **MIL vs. mean pooling:** small but consistent gain (ΔF1 = +0.004; ΔAP = +0.007).
- ▸ **Acoustic support:** degrades AP (0.835 → 0.744), likely due to redundancy/noise.
- ▸ **Text support:** improves AP (0.835 → 0.842), but slightly reduces Macro-F1.
- ▸ **All support: best single-model result (F1 = 0.710; AP = 0.857), supporting H2.**
- ▸ **5-seed ensemble:** further improves performance (F1 = 0.722; AP = 0.875).

9. RESULTS: ENSEMBLE SIZE ABLATION

Seeds	$F1_{smooth}$	$F1_{base\text{-}rate}$	AP
1	0.7107	0.7081	0.8572
3	0.7213	**0.7252**	0.8719
5 ⋆	0.7226	0.7191	**0.8750**
7	**0.7245**	0.7124	0.8726

- ▸ Performance gain saturates beyond three models
- ▸ Five-seed ensemble provides the best trade-off:
  - ▸ ranking quality;
  - ▸ classification stability;
  - ▸ computational cost

10: LEADERBOARD

Rank	Team	Macro-F1
1	AffectVision	0.7959
2	RAS	0.7824
3	**PUCPR (Ours)**	**0.7455**
4	AIM for Emo	0.7431
5	AIWELL	0.7367
5	CASIA-26	0.7367