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, and Alceu S. Britto Jr.

Pontifícia Universidade Católica do Paraná, Curitiba, Brazil
LIA - Artificial Intelligence Academic League
ligaIA@ppgia.pucpr.br

Abstract. We present a frame-independent audio–text system for the 3rd Ambivalence/Hesitancy Video Recognition Challenge at the 11th Affective & Behavior Analysis in-the-Wild (ABAW) Workshop. Videos are divided into overlapping 5-s windows aligned with transcript timestamps. Each window combines prosodic audio descriptors, emotion-oriented RoBERTa embeddings, and 74 psycholinguistic features representing uncertainty, hedging, and attitudinal conflict. Temporal cross-attention fuses audio and text, while the support features condition gated Multiple Instance Learning (MIL) pooling. A five-seed ensemble achieves an average precision of 0.875 and a macro-F1 of 0.722 on the 525-video labeled public-test split. Notably, our submission ranked third overall on the official challenge leaderboard, with a macro-F1 of 0.7455. Source code is available at https://github.com/Liga-de-IA-PUCPR/abaw-11-ah-challenge/.

Keywords: Ambivalence · Hesitancy · Multimodal learning · Cross-attention · Multiple instance learning

1 Introduction

Ambivalence and Hesitancy (A/H) are relevant behavioral signals in digital behavior-change interventions because they may indicate that a person is uncertain about, or simultaneously attracted to and resistant toward, a proposed action [41]. Ambivalence is not equivalent to neutrality or simple indecision: it arises when positive and negative evaluations of the same object are simultaneously accessible, producing evaluative conflict even when their net difference is close to zero [27, 35, 40]. Such conflict can increase information processing and weaken attitude–intention consistency, making ambivalence especially relevant when a person is considering behavioral change [11, 25]. Hesitancy is the temporally unfolding expression of reduced commitment and may be conveyed by delayed responses, silent or filled pauses, lengthening, hedges, repairs, slower speech, and rising or more variable intonation [5, 24, 37, 39]. As these markers

Accepted to ABAW @ ECCV 2026. The final published version will be available on *Springer*.

reflect partially distinct linguistic and acoustic processes, hesitancy should not be reduced to the occurrence of a single filler or pause [4].

Recognizing A/H in videos is challenging for two main reasons. First, the relevant evidence may occur only during a short portion of a video rather than persist throughout the complete response. A positive video may contain long neutral or fluent stretches interleaved with a brief pause, hedge, self-correction, or explicit evaluative conflict. Consequently, the video-level label does not imply that every temporal segment contains diagnostic evidence. This setting is related to weakly supervised temporal localization, in which a global label may be explained by a sparse subset of discriminative snippets whose relevance must be inferred without using dense annotations during training [33]. Uniform temporal averaging can therefore dilute localized A/H cues with substantially longer non-diagnostic intervals. Second, useful evidence is distributed across modalities. Response latency, pauses, speech timing, pitch variation, epistemic hedges, contrastive constructions, and conflicting positive and negative evaluations may all contribute to the final decision [32, 36]. Explicitly modeling these cues is especially relevant when the amount of labeled data is limited and generic high-dimensional embeddings may not reliably expose the constructs of interest.

Such challenges motivate our central research question: can attention-based temporal aggregation combined with explicit prosodic and psycholinguistic support features improve audio–text A/H recognition over generic embeddings and uniform temporal pooling, without relying on visual-frame input? Our objective is to develop a frame-independent audio–text method for video-level A/H recognition that identifies informative temporal segments while explicitly representing behavioral markers of uncertainty and attitudinal conflict. Although visual behavior may provide complementary evidence, audio and language already contain well-established indicators of A/H; studying these modalities without visual frames also makes it possible to isolate their contribution and removes the need for frame decoding and visual feature extraction.

We hypothesize that **(H1)** attention-based Multiple Instance Learning (MIL) pooling provides a more effective video-level aggregation mechanism than uniform mean pooling, as A/H evidence is temporally sparse in the BAH corpus. We further hypothesize that **(H2)** explicit support features encoding prosodic uncertainty, textual hedging, and attitudinal conflict provide a complementary inductive bias beyond what generic audio and text embeddings capture alone.

To investigate these hypotheses, we develop an audio–text architecture that operates exclusively on audio and timestamped transcripts while preserving the video-level prediction setting. Videos are divided into overlapping 5-s windows that match the typical duration of an A/H episode in the Behavioral Ambivalence/Hesitancy (BAH) dataset [20]. For every window, we extract prosodic audio descriptors, an emotion-oriented text embedding, and an interpretable support vector grounded in psycholinguistic descriptions of uncertainty and ambivalence. Cross-attention combines the aligned audio and text representations, after which the support features are incorporated into the window representation. Gated

MIL pooling then assigns greater weight to the most informative windows, and a five-seed ensemble is used to reduce training variance.

The main contributions of this work are threefold:

- A frame-independent audio–text architecture that combines temporal cross-attention with gated MIL pooling for localized video-level A/H recognition;
- A 74-dimensional psycholinguistic support representation encoding question context, prosodic uncertainty, textual hedging, emotion dynamics, and attitudinal ambivalence;
- An experimental analysis of temporal aggregation and explicit support features, showing that the support representation provides the largest observed improvement.

The experimental results on a 525-video labeled public test set show that replacing uniform mean pooling with attention-based MIL increases average precision from 0.828 to 0.835 and macro-F1 from 0.701 to 0.705. Adding the complete support representation and averaging the predictions of five independently initialized models further increases average precision to 0.875 and macro-F1 to 0.722. These results indicate that explicit prosodic and psycholinguistic descriptors provide useful complementary information for audio–text A/H recognition, while the smaller improvement obtained from MIL provides more limited evidence for the advantage of learned temporal aggregation.

The remainder of this paper is organized as follows. Section 2 reviews the literature on prosodic and linguistic uncertainty, attitudinal ambivalence, multimodal fusion, and temporal aggregation. Section 3 presents the architecture and feature representations. Section 4 describes the dataset, training procedure, and evaluation protocol. Section 5 reports and discusses the results. Finally, Section 6 summarizes the findings and outlines directions for future work.

2 Related Work

Research on A/H recognition draws on speech prosody, linguistic uncertainty, attitudinal ambivalence, and multimodal temporal aggregation. These complementary areas motivate the domain-informed support features, cross-modal fusion, and temporal pooling used in our approach.

2.1 Prosodic and linguistic markers of uncertainty

Speech conveys a speaker's degree of certainty through timing, intonation, loudness, and voice quality. Doubtful utterances typically exhibit longer response latencies, more pauses, slower speech, rising terminal intonation, lower or less stable intensity, and greater fundamental-frequency variability [5, 24, 37, 39]. Such features can be operationalized through automatic estimates of speech and articulation rate based on syllable-nucleus detection [12, 26], together with compact voice-quality descriptors such as jitter and shimmer [16].

The literature distinguishes uncertainty, as a speaker state, from disfluencies, as discrete, observable events. Filled pauses can be detected from acoustic and textual representations, with acoustic models generally providing stronger frame-level evidence and multimodal combinations yielding additional gains [9]. However, filler counts alone do not represent continuous variations in prosodic uncertainty and may be unreliable when derived from Automatic Speech Recognition (ASR) transcripts. We therefore model timing, pitch, loudness, speech rate, and voice quality directly rather than treating filled-pause detection as the primary task. Text provides complementary evidence through hedges and other epistemic markers.

The CoNLL-2010 shared task [17] established uncertainty detection as a context-dependent problem: expressions such as "may," "suggest," and "I think" function as hedges only in particular linguistic contexts. Hedge detection has also been approached through distant supervision and shallow linguistic features, which are especially useful in label-scarce settings [18]. These findings motivate explicit features for graded epistemic commitment, contrast, and polarity shifts.

2.2 Attitudinal ambivalence and mixed affect

Psychological models describe ambivalence as the simultaneous activation of positive and negative, or approach and avoidance, tendencies. Behavioral studies have shown that conflicting appetitive and aversive evidence affects gaze and response trajectories, producing measurable approach–avoidance conflict [10, 19].

Psychometric models therefore represent positive and negative components separately. Kaplan's ambivalence-indifference problem shows that a net polarity score cannot distinguish weak positive and negative evidence from strong but conflicting evidence [27]. Conflict can instead be represented using quantities such as $\min(P, N)$ and the similarity-intensity index using Eq. (1).

$$\frac{P+N}{2} - |P-N| \,. \tag{1}$$

The index increases when positive and negative components are simultaneously strong [40]. Related work also distinguishes cognitive–affective inconsistency from subjectively experienced ambivalence [11]. A similar principle appears in multi-label emotion recognition, where opposite-valence emotions may co-occur rather than being treated as mutually exclusive. Models such as MEDA explicitly capture correlations among emotions and can therefore represent mixed affective states [14]. Together, these findings motivate separate positive and negative intensities, ambivalence indices, emotion-distribution uncertainty, and temporal variation in valence.

2.3 Cross-modal fusion and temporal aggregation

Attention mechanisms provide a flexible framework for modeling interactions between modalities [42]. In affective computing, audio-text cross-attention has been

used to capture complementary semantic and paralinguistic information, thereby improving emotion recognition over unimodal or simple late-fusion systems [38].

Related work has extended cross-modal attention to visual streams [13]. These studies motivate our use of audio-to-text cross-attention over aligned temporal windows. As A/H labels are provided at the video level while evidence may occur only in a few temporal segments, Multiple Instance Learning (MIL) provides a natural formulation. A video can be represented as a bag of windows, with attention-based pooling learning the contribution of each instance [23].

Similar formulations have been applied to temporally localized psychological and clinical evidence, including stress, autism spectrum disorder, and pain recognition [6, 15]. Hard-top-k aggregation may produce incomplete localization when the evidence is discontinuous, motivating softer attention-based alternatives [43]. We therefore use gated soft MIL pooling [23], allowing multiple windows to contribute with learned continuous weights rather than selecting a fixed number of instances.

2.4 A/H recognition systems

The BAH dataset [20] and its associated challenge are recent, and most directly related systems are challenge submissions or preprints evaluated on different validation, public-test, or private-test splits. Their reported scores should therefore be compared cautiously. Most existing systems combine visual, audio, and textual information.

The broader evolution of the ABAW benchmark and its associated affective and behavioral analysis tasks has been documented in successive workshop and competition reports [28, 29].

Cabacas-Maso et al. [8] ensemble fusion models over face, audio, text, and pose representations. Bui et al. [7] use visual, audio, and textual embeddings with speaker-normalized features and a certainty-weighted loss. Bakin and Savchenko [1] combine frame-level face, audio, and text classifiers through late fusion and transcript-based gating, while Kumar et al. [30] enrich multimodal embeddings with hand-crafted hesitation markers. Ryumina et al. [34] instead use text as an anchor modality and introduce gated residual contributions from audio, face, and scene streams.

Related baselines for earlier or differently evaluated versions of the task include video architectures and multimodal-LLM prompting [21]. In contrast to these predominantly audiovisual systems, our method uses only audio and timestamped transcripts. It compensates for the absence of a visual stream through explicit prosodic, epistemic, and attitudinal support features. These descriptors are fused with the cross-attended audio–text representation before gated MIL pooling, allowing them to influence both window representations and temporal aggregation weights.

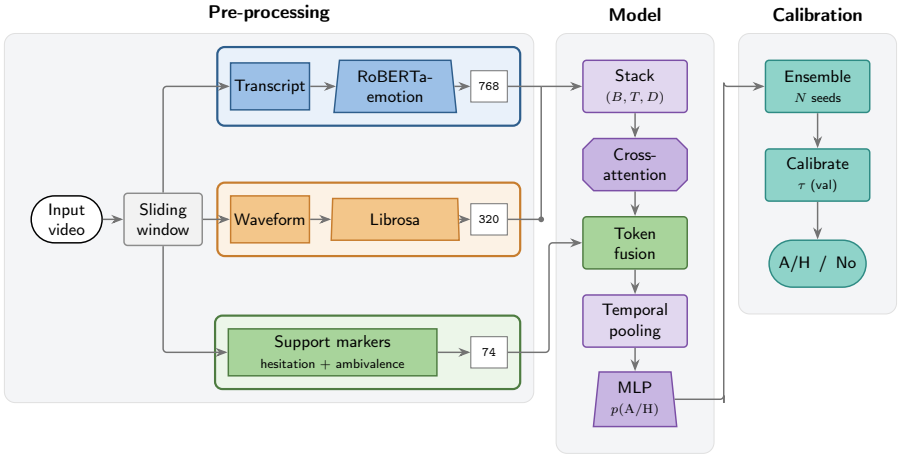

Fig. 1: Overview of the audio–text cross-attention architecture with psycholinguistic support features.

3 Proposed Method

Figure 1 provides an overview of the proposed architecture, which integrates audio-text cross-attention, psycholinguistic support features, gated MIL pooling, and a video-level classifier.

3.1 Pre-processing and temporal alignment

The audio track is extracted from each video, converted to mono 16-kHz FLAC, and segmented into 5-s windows with a 2.5-s hop. The transcripts in the BAH dataset include Whisper-generated text and timestamps. We use these timestamps to concatenate all transcript segments that overlap each audio window. Timestamp resets at internal Whisper segment boundaries are corrected by accumulating an offset, which produces a monotonic timeline.

The target is the global video label $y_v \in \{0, 1\}$. The released participant-wise splits prevent the same participant from appearing in more than one partition. The complete video is represented as a variable-length sequence of T aligned windows.

3.2 Window representations

Each temporal window t is represented by three complementary vectors: an audio descriptor $\mathbf{a}_t \in \mathbb{R}^{320}$, a text embedding $\mathbf{x}_t \in \mathbb{R}^{768}$, and a psycholinguistic support vector $\mathbf{s}_t \in \mathbb{R}^{74}$. The audio and text vectors provide generic acoustic, semantic, and affective representations, whereas the support vector explicitly encodes cues associated with attitudinal ambivalence and hesitancy.

Audio representation. The 320-dimensional audio descriptor is extracted with `librosa` and contains MFCCs and their first- and second-order derivatives, spectral centroid, bandwidth, roll-off, flatness, spectral contrast, chroma, zero-crossing rate, RMS energy, pitch statistics, voiced fraction, and tempo. When applicable, frame-level descriptors are summarized by their mean, standard deviation, minimum, and maximum. This representation was selected after preliminary experiments in which generic self-supervised speech embeddings did not improve performance.

Text representation. Text is encoded with the TweetEval RoBERTa emotion encoder[1] [2, 31]. We remove the classification head, apply attention-masked mean pooling over the token representations, and perform ℓ_2 normalization. The resulting 768-dimensional vector represents the semantic and affective content of each window.

Psycholinguistic support representation. The support vector comprises three blocks: a 7-dimensional question-context descriptor, an 18-dimensional acoustic–prosodic block, and a 49-dimensional textual block. Table 1 summarizes their composition, dimensionality, and temporal granularity.

Table 1: Composition and temporal granularity of the psycholinguistic support vector. Response-level features are broadcast to every window.

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

The acoustic support features are computed from the raw waveform using `librosa` for timing, pause, and loudness descriptors and Praat through Parselmouth for F0, jitter, and shimmer. The analysis uses a frame length of 2048

[1] `https://huggingface.co/cardiffnlp/twitter-roberta-base-emotion`

samples and a hop of 512 samples at 16 kHz. Textual features are computed from the ASR transcript using the VADER sentiment lexicon [22] and the four-class output of the TweetEval emotion model [2], with labels {anger, joy, optimism, sadness}.

Feature granularity. Each video contains a single question–answer pair, with approximately 24 words per window and 130 words per complete response, distributed over about nine windows. The question type is constant throughout the response and is therefore represented by a 7-dimensional one-hot vector broadcast to every window.

Ambivalence is also modeled primarily at the response level because opposing evaluations may occur in different windows. A single short window may therefore contain insufficient sentiment-bearing content to estimate positive and negative evidence reliably. We compute the 20 ambivalence features over the reconstructed transcript and broadcast them to every window. Four emotion dynamics are likewise computed at the response level: the standard deviation and range of window-level valence, pole conflict, and overall response valence.

All acoustic–prosodic features are computed locally for each window. The epistemic-hedge and contrast/polarity-shift groups are also window-specific. The remaining two emotion-dynamics features—emotion entropy and the gap between the two highest emotion probabilities—are computed per window. Thus, the emotion-dynamics block has mixed granularity, as indicated in Table 1.

Acoustic–prosodic features. The 18 acoustic–prosodic features quantify the speaker's state of (un)certainty rather than detecting isolated disfluency events such as filled pauses [5, 24]. A voice-activity detector based on `librosa.effects.split` with `top_db` = 30 segments each window into speech intervals and internal silences.

- *Timing and pauses* (6) comprise onset latency, the proportion of low-energy frames, pauses per second, mean pause duration, the proportion of time spent in long pauses, and voiced fraction. Low-energy frames have RMS below 0.1 times the window peak; pauses and long pauses correspond to silences of at least 0.15 s and more than 0.5 s, respectively [5, 24, 37, 39];
- *Speech rate and articulation* (3) comprise speech rate, articulation rate, and phonation ratio. Syllable nuclei are detected as intensity peaks within 25 dB of the loudest frame and separated by a dip of at least 2 dB, following de Jong and Wempe [26]. The measures are derived from the number of nuclei, total duration, and phonation time [12, 24];
- *Pitch and intonation* (4) are computed from the Praat F0 contour over voiced frames using autocorrelation in the 65–500 Hz range. They comprise the coefficient of variation, semitone range, global slope, and terminal slope over the final 30% of voiced frames. Greater variability and rising terminal intonation are associated with uncertainty [5, 24];
- *Loudness* (2) is represented by the mean and coefficient of variation of frame-level RMS energy, since lower and less stable intensity may indicate doubt [24];

– *Voice quality and uncertainty score* (3) comprise local jitter and shimmer, computed with Praat and motivated by the eGeMAPS parameter set [16], and a fixed-weight uncertainty score used only for inspection.[2]

Textual features. The 49 textual features represent ambivalence, emotion dynamics, discourse contrast, polarity shifts, and epistemic hedging.

– *Ambivalence indices* (20) are computed from two polarity sources: VADER positive and negative proportions and the emotion-model probabilities, where positive evidence is defined as joy plus optimism and negative evidence as anger plus sadness. For each source, we retain positive intensity P, negative intensity N, and four psychometric measures:

$$\frac{P+N}{2}, \qquad |P-N|, \qquad \min(P,N), \qquad \frac{P+N}{2} - |P-N|. \tag{2}$$

These correspond to mean intensity, polar discrepancy, Kaplan's minimum [27], and the Griffin similarity–intensity index [11,40]. Retaining P and N separately avoids mapping neutral and ambivalent responses to the same net score $P-N$ [27].
The remaining eight ambivalence features describe an approach–avoidance axis. Desire and resistance markers are identified using hand-built lexicons released with the code. Their frequencies are represented as raw counts, per-word rates, and per-100-word rates, together with the product and minimum of the desire and resistance components.
– *Emotion dynamics* (6) include two window-level and four response-level measures. Window-level features are the Shannon entropy of the four-class emotion distribution and the gap between its two highest probabilities. Response-level features are the standard deviation and range of window valence $P-N$, pole conflict $\min(P,N)$, and overall response valence [14].
– *Contrast and polarity shift* (5) comprise the frequency of contrastive discourse markers in raw, per-word, and per-100-word forms, the absolute difference between VADER compound scores in the first and second halves of the window, and a binary indicator of whether polarity changes sign. Contrastive markers include *but*, *however*, *although*, *though*, *yet*, *on the other hand*, and *at the same time*.
– *Epistemic hedges* (18) represent six categories: first-person epistemic expressions, adverbial hedges, approximators, explicit uncertainty, agentless attribution, and the aggregate hedge count. Each category is represented as a raw count, a per-word rate, and a per-100-word rate [17,18].

3.3 Cross-attention and token fusion

Audio and text are independently projected to a common width $d = 512$. Multi-head cross-attention uses the audio sequence as query and the aligned text

[2] $s = \mathrm{clip}(0.20\,s_{\mathrm{lat}} + 0.20\,s_{\mathrm{pause}} + 0.15\,s_{\mathrm{rise}} + 0.15\,s_{\mathrm{F0cv}} + 0.15\,s_{\mathrm{loud}} + 0.15\,s_{\mathrm{slow}}, 0, 1)$. Each component is clipped to $[0,1]$, and the F0-related terms are gated by the voiced fraction.

sequence as key and value:

$$\mathbf{H} = \mathrm{LN}(\mathbf{A} + \mathrm{MHA}(\mathbf{A}, \mathbf{X}, \mathbf{X})), \tag{3}$$

where padding windows are masked and LN denotes layer normalization.

The heterogeneous support features are normalized over valid windows only, projected to 512 dimensions, and concatenated with the cross-modal representation before another linear projection:

$$\widetilde{\mathbf{h}}_t = \mathrm{LN}\left(W_f[\mathbf{h}_t \| g(\mathbf{s}_t)]\right). \tag{4}$$

This token fusion places the interpretable cues before temporal aggregation. Consequently, a window containing strong uncertainty or conflict cues can receive greater MIL attention. Experiments showed that this placement was more effective than concatenating the support vector only after video pooling.

3.4 MIL pooling, classifier, and ensemble

Gated attention pooling assigns a score to each valid window [23]:

$$q_t = \mathbf{w}^\top \left[\tanh(V\widetilde{\mathbf{h}}_t) \odot \sigma(U\widetilde{\mathbf{h}}_t)\right], \tag{5}$$

$$\alpha_t = \frac{\exp(q_t)}{\sum_{j=1}^{T} \exp(q_j)}, \qquad \mathbf{z} = \sum_{t=1}^{T} \alpha_t \widetilde{\mathbf{h}}_t. \tag{6}$$

The video vector $\mathbf{z}$ passes through a 512-unit ReLU layer, dropout (0.3), and a binary output layer. The model is optimized with binary cross-entropy and AdamW using a learning rate of 3×10^{-4}, weight decay of 0.05, a batch size of 32, gradient clipping at 1.0, and early stopping based on validation average precision.

Five models with seeds $\{42, 1, 2, 3, 4\}$ are trained with identical settings, and their video probabilities are averaged into a single score $s_v \in [0, 1]$ per video.

3.5 Threshold calibration

Binary predictions are obtained by applying a threshold τ to the ensemble score s_v. The threshold is selected exclusively on the validation split and frozen for test evaluation. As the validation set contains only 124 videos, directly maximizing macro-F1 over $\{0, 0.01, \ldots, 1\}$ may be unstable.

For the ensemble, we smooth the validation macro-F1 curve with a moving average of width 0.1 and define the near-optimal plateau, as denoted in Eq. (7).

$$\mathcal{P} = \left\{\tau : \widetilde{F}_1(\tau) \geq \max_{\tau'} \widetilde{F}_1(\tau') - 0.01\right\}. \tag{7}$$

We select the center of its longest contiguous interval. For individual models, we additionally evaluate base-rate matching, setting τ to the $(1 - p_+)$ quantile of validation scores using Eq. (8). Both strategies use validation information only.

$$p_+ = \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} y_v, \qquad \tau = Q_{1-p_+}\left(\{s_v\}_{v \in \mathcal{V}}\right). \tag{8}$$

4 Experimental Protocol

The BAH dataset [20] contains 1,427 labeled videos from 300 Canadian participants. The released splits contain 778 training, 124 validation, and 525 public-test videos. Our final configuration fits the model weights on the training set, while the threshold used for the reported macro-F1 is calibrated on the validation set.

We evaluate video-level recognition using macro-F1 across the positive and negative classes and additionally report positive-class Average Precision (AP). All reported results were produced by the supplied implementation on the labeled public test split. All experiments were conducted on a single Apple M3 Pro chip (14-core GPU, 18 GB unified memory), using the Metal Performance Shaders (MPS) backend for GPU acceleration.

5 Results and Discussion

Table 2 reports the main results on the 525-video labeled public-test split of BAH. The experiments isolate the effects of temporal pooling, the acoustic and textual support-feature groups, the complete support representation, and probability averaging across independently initialized models.

Table 2: Performance on the BAH public test split containing 525 labeled videos [20].

Method	Macro-F1	AP
Comparison methods		
ConflictAwareAH [3] (Fennec)	0.694	N/A
Mean-pooling baseline	0.701	0.828
Proposed MIL variants		
Audio + text, without support features	0.705	0.835
Audio support only	0.636	0.744
Text support only	0.698	0.842
All support features, single seed	0.710	0.857
All support features, five-seed ensemble	**0.722**	**0.875**

Temporal aggregation. Replacing uniform temporal mean pooling with attention-based MIL, while retaining only the audio and text embeddings, increases AP from 0.828 to 0.835 and macro-F1 from 0.701 to 0.705. The improvement is consistent across both metrics but small in magnitude. This result is directionally consistent with H1, as learned temporal weighting yields a modest improvement over uniform averaging. However, the gain was obtained from a single training run and is comparable in magnitude to the seed-to-seed variation observed elsewhere in our experiments. We therefore regard this result as suggestive rather than conclusive evidence in support of H1. Furthermore, the learned MIL weights are

neither visualized nor compared with the corpus' time-resolved A/H annotations in the present study, although such annotations make this analysis possible. We leave this investigation to future work and, accordingly, interpret the learned weights only as internal aggregation coefficients rather than validated indicators of the temporal location of A/H evidence. Our empirical conclusions are therefore restricted to video-level recognition.

Acoustic support features. Adding the acoustic–prosodic support block decreases macro-F1 from 0.705 to 0.636 and AP from 0.835 to 0.744. These features may be noisy in short or weakly voiced windows and partly redundant with the 320-dimensional base audio representation, which already includes timing, pitch, energy, and spectral descriptors. The acoustic support block therefore provides no independent benefit in the present configuration.

Textual support features. Textual support increases AP from 0.835 to 0.842 but decreases macro-F1 from 0.705 to 0.698. The explicit hedging, polarity, emotion, and ambivalence markers therefore improve ranking quality, although this gain does not transfer to threshold-dependent classification, possibly because of calibration sensitivity or overlapping class-score distributions.

Complete support representation. Combining contextual, acoustic, and textual support produces the strongest single-model result, increasing AP from 0.835 to 0.857 and macro-F1 from 0.705 to 0.710. This controlled comparison supports H2 for the complete representation rather than for each feature group independently. The full representation outperforms both isolated blocks, suggesting complementarity among the support cues, although the current ablation does not directly test their interactions. As the support features are fused before temporal pooling, they may influence both the window representations and the internal MIL weighting. Establishing whether these learned weights correspond to meaningful A/H episodes, however, requires a dedicated temporal analysis and is therefore beyond the video-level evaluation performed here.

Ensemble effect. Averaging five full-support models increases AP from 0.857 to 0.875 and macro-F1 from 0.710 to 0.722. As shown in Table 3, most of the gain is already obtained with three models, and performance saturates thereafter. Five seeds yield the highest AP and stable macro-F1 under smooth calibration, although they do not dominate every thresholding strategy. The ensemble gain is therefore distinct from, and additional to, the contribution of the support representation.

Official challenge comparison. Table 4 reports the final leaderboard of the 3rd Ambivalence/Hesitancy (AH) Video Recognition Challenge. Our submission (PUCPR) achieved a macro-F1 of 0.7455 and ranked third overall, behind AffectVision (0.7959) and RAS (0.7824), while narrowly outperforming AIM for Emo (0.7431). The winning AffectVision approach combines affect-specialized textual, acoustic, and visual representations with interpretable hesitation cues

Table 3: Effect of ensemble size on the test set under smooth and base-rate threshold calibration, with ⋆ indicating the selected configuration.

Seeds	$F1_{test}$ (smooth)	$F1_{test}$ (base rate)	AP_{test}
1 (single)	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

through Affective Marker Fusion, followed by an AP-weighted ensemble with a fixed decision threshold [30]. The second-ranked RAS approach adopts a text-centered multimodal strategy in which linguistic information serves as the anchor and acoustic, facial, and scene representations provide complementary information through gated residual adjustments [34]. In contrast, our system relies exclusively on audio and timestamped transcripts, without processing visual frames. Its third-place result therefore demonstrates competitive challenge performance while retaining a frame-independent design. The leaderboard score is reported separately from the controlled experiments in Table 2, which use the labeled public-test split to analyze the individual components of the proposed method.

Table 4: Final leaderboard of the 3rd Ambivalence/Hesitancy (AH) Video Recognition Challenge. Our submission (PUCPR) ranked third overall.

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

Overall interpretation. The results provide limited but directionally positive evidence for H1 and stronger evidence for H2. MIL yields a small improvement over mean pooling, whereas the complete support representation provides the clearest single-model gain. Acoustic and textual support behave differently in isolation, and probability averaging provides a further improvement, particularly in AP. All results are obtained without visual-frame features, establishing a frame-independent audio–text baseline for video-level A/H recognition.

6 Conclusions

We presented a frame-independent audio–text approach for video-level Ambivalence and Hesitancy (A/H) recognition. The method combines prosodic audio descriptors and emotion-oriented text embeddings through cross-attention, incorporates psycholinguistic support features before gated Multiple Instance Learning (MIL) pooling, and averages five independently initialized models.

The experiments separate the effects of temporal aggregation, support features, and ensembling. Gated MIL provides a small improvement over mean pooling, offering limited support for H1. The complete support representation yields the strongest single-model result and supports H2, although its acoustic and textual components are not consistently beneficial when used independently. The five-seed ensemble further improves performance, reaching an AP of 0.875 and a macro-F1 of 0.722. These results show that explicit uncertainty and ambivalence cues complement generic audio and text representations without requiring visual-frame processing. Notably, on the official challenge leaderboard, our submission achieved a macro-F1 of 0.7455 and ranked third overall.

The study is limited by the small, single-dataset evaluation, possible instability in threshold calibration, dependence on Whisper-generated transcripts, and the poor performance of the acoustic support block when used alone. Excluding visual information also leaves potentially complementary facial and gestural evidence unexplored.

Future work should evaluate the method on additional datasets, characterize the effect of ASR errors, and further refine the acoustic support features. We also plan to systematically vary the window length and overlap to quantify the effect of temporal granularity, and to conduct a broader threshold-sensitivity analysis. Regarding temporal modeling, explicit absolute or relative positional encodings could be incorporated to provide the model with direct information about window order. In addition, visualizing the learned MIL weights and comparing them with temporal A/H annotations would help determine whether the internal weighting aligns with meaningful A/H episodes. Finally, facial expressions, head motion, body gestures, and other visual cues could be incorporated as a complementary stream to quantify their incremental contribution over the proposed frame-independent audio–text baseline.

Acknowledgments

The authors gratefully acknowledge the *Pontifícia Universidade Católica do Paraná* (PUCPR) for the financial support that enabled participation in the conference. The authors also acknowledge support from the *Conselho Nacional de Desenvolvimento Científico e Tecnológico* (CNPq) under Grants 406030/2023-5 and 306878/2022-4, and from the *Fundação Araucária*, in partnership with the *Secretaria da Ciência, Tecnologia e Ensino Superior do Estado do Paraná* (SETI-PR), under Grant No. 653/2025 (FA/UNIVERSAL).

References

1. Bakin, A., Savchenko, A.V.: HSEmotion team at the 11th ABAW challenge: Multi-task learning and ambivalence/hesitancy video recognition. arXiv preprint arXiv:2607.12774 (2026). https://doi.org/10.48550/arXiv.2607.12774
2. Barbieri, F., Camacho-Collados, J., Espinosa Anke, L., Neves, L.: TweetEval: Unified benchmark and comparative evaluation for tweet classification. In: Findings of the Association for Computational Linguistics: EMNLP (2020). https://doi.org/10.18653/v1/2020.findings-emnlp.148
3. Bekhouche, S.E., Telli, H., Benlamoudi, A., Herrouz, S.E., Taleb-Ahmed, A., Hadid, A.: Conflict-aware multimodal fusion for ambivalence and hesitancy recognition. arXiv preprint arXiv:2603.15818 (2026). https://doi.org/10.48550/arXiv.2603.15818
4. Betz, S., Bryhadyr, N., Türk, O., Wagner, P.: Cognitive load increases spoken and gestural hesitation frequency. Languages **8**(1), 71 (2023). https://doi.org/10.3390/languages8010071
5. Brennan, S.E., Williams, M.: The feeling of another's knowing: Prosody and filled pauses as cues to listeners about the metacognitive states of speakers. Journal of Memory and Language **34**, 383–398 (1995). https://doi.org/10.1006/jmla.1995.1017
6. Brügge, N.S., Korda, A., Borgwardt, S., Andreou, C., Giannakakis, G., Handels, H.: Bag-level multiple instance learning for acute stress detection from video data. In: International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC) (2025). https://doi.org/10.5220/0013364900003911
7. Bui, T.H., Nguyen, H.H., Huynh, V.T.: CF-Net: Conflict fusion with speaker normalisation and certainty weighting for ambivalence/hesitancy recognition. arXiv preprint arXiv:2607.13976 (2026). https://doi.org/10.48550/arXiv.2607.13976
8. Cabacas-Maso, J., Benito-Altamirano, I., Ventura, C.: A calibrated multimodal ensemble for ambivalence/hesitancy recognition: System description and private-test submission strategy. arXiv preprint arXiv:2607.12176 (2026). https://doi.org/10.48550/arXiv.2607.12176
9. Chatziagapi, A., Sgouropoulos, D., Karouzos, C., Melistas, T., Giannakopoulos, T., Katsamanis, A., Narayanan, S.: Audio and asr-based filled pause detection. In: International Conference on Affective Computing and Intelligent Interaction (ACII) (2022). https://doi.org/10.1109/ACII55700.2022.9953889
10. Chen, M., Reutter, M., Pauli, P., Gamer, M., Pittig, A.: Overcoming automatic behavioral tendencies in approach-avoidance conflict decisions. Psychophysiology **62**, e70101 (2025). https://doi.org/10.1111/psyp.70101
11. Conner, M., Wilding, S., van Harreveld, F., Dalege, J.: Cognitive-affective inconsistency and ambivalence: Impact on the overall attitude-behavior relationship. Personality and Social Psychology Bulletin **47**(4), 673–687 (2021). https://doi.org/10.1177/0146167220945900
12. Cucchiarini, C., Strik, H., Boves, L.: Quantitative assessment of second language learners' fluency by means of automatic speech recognition technology. The Journal of the Acoustical Society of America **107**(2), 989–999 (2000). https://doi.org/10.1121/1.428279
13. Das, A., Sarma, M.S., Hoque, M.M., Siddique, N., Dewan, M.A.A.: AVaTER: Fusing audio, visual, and textual modalities using cross-modal attention for emotion recognition. Sensors **24**, 5862 (2024). https://doi.org/10.3390/s24185862

14. Deng, J., Ren, F.: Multi-label emotion detection via emotion-specified feature extraction and emotion correlation learning. IEEE Transactions on Affective Computing **14**(1), 475–486 (2023). https://doi.org/10.1109/TAFFC.2020.3034215
15. Duan, W., Li, J., Ouyang, G.: Facial instance learning for video-based asd diagnosis. In: International Conference on Mechatronics and Machine Vision in Practice (M2VIP) (2024). https://doi.org/10.1109/M2VIP62491.2024.10746003
16. Eyben, F., Scherer, K.R., Schuller, B.W., Sundberg, J., André, E., Busso, C., Devillers, L.Y., Epps, J., Laukka, P., Narayanan, S.S., Truong, K.P.: The geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. IEEE Transactions on Affective Computing **7**(2), 190–202 (2016). https://doi.org/10.1109/TAFFC.2015.2457417
17. Farkas, R., Vincze, V., Móra, G., Csirik, J., Szarvas, G.: The CoNLL-2010 shared task: Learning to detect hedges and their scope in natural language text. In: Conference on Computational Natural Language Learning (2010)
18. Ganter, V., Strube, M.: Finding hedges by chasing weasels: Hedge detection using wikipedia tags and shallow linguistic features. In: Proceedings of the ACL-IJCNLP 2009 Conference Short Papers. pp. 173–176 (2009)
19. Garcia-Guerrero, S., O'Hora, D., Zgonnikov, A., Scherbaum, S.: The action dynamics of approach-avoidance conflict during decision-making. Quarterly Journal of Experimental Psychology **76**(1), 160–179 (2023). https://doi.org/10.1177/17470218221087625
20. González-González, M., Belharbi, S., Zeeshan, M.O., Sharafi, M., Aslam, M.H., Pedersoli, M., Koerich, A.L., Bacon, S.L., Granger, E.: BAH dataset for ambivalence/hesitancy recognition in videos for digital behavioural change. International Conference on Learning Representations (2026)
21. Gonz'alez-Gonz'alez, M., Belharbi, S., Zeeshan, M.O., Sharafi, M., Aslam, M.H., Sia, L., Richet, N., Pedersoli, M., Koerich, A.L., Bacon, S.L., Granger, E.: Multimodal ambivalence/hesitancy recognition in videos for personalized digital health interventions. arXiv preprint arXiv:2604.11730 (2026). https://doi.org/10.48550/arXiv.2604.11730
22. Hutto, C.J., Gilbert, E.: VADER: A parsimonious rule-based model for sentiment analysis of social media text. In: International AAAI Conference on Web and Social Media (ICWSM). vol. 8, pp. 216–225 (2014). https://doi.org/10.1609/icwsm.v8i1.14550
23. Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: International Conference on Machine Learning (2018)
24. Jiang, X., Pell, M.D.: The sound of confidence and doubt. Speech Communication **88**, 106–126 (2017). https://doi.org/10.1016/j.specom.2017.01.011
25. Jonas, K., Diehl, M., Brömer, P.: Effects of attitudinal ambivalence on information processing and attitude-intention consistency. Journal of Experimental Social Psychology **33**(2), 190–210 (1997). https://doi.org/10.1006/jesp.1996.1317
26. de Jong, N.H., Wempe, T.: Praat script to detect syllable nuclei and measure speech rate automatically. Behavior Research Methods **41**(2), 385–390 (2009). https://doi.org/10.3758/BRM.41.2.385
27. Kaplan, K.J.: On the ambivalence-indifference problem in attitude theory and measurement: A suggested modification of the semantic differential technique. Psychological Bulletin **77**(5), 361–372 (1972). https://doi.org/10.1037/h0032590
28. Kollias, D., Tzirakis, P., Cowen, A., Zafeiriou, S., Kotsia, I., Granger, E., Pedersoli, M., Bacon, S., Baird, A., Gagne, C., Shao, C., Hu, G., Belharbi, S., Aslam, M.H.: Advancements in affective and behavior analysis: The 8th ABAW

workshop and competition. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 5572–5583 (2025). https://doi.org/10.1109/CVPRW67362.2025.00554
29. Kollias, D., Tzirakis, P., Cowen, A., Zafeiriou, S., Kotsia, I., Granger, E., Pedersoli, M., Bacon, S., Madsen, J., Belharbi, S., Aslam, M.H., Shao, C., Hu, G.: From affect to complex behavior: Advancing multimodal human-centered ai at the 10th ABAW workshop & competition. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. pp. 5302–5311 (June 2026)
30. Kumar, V., Mishra, A., Lone, H.R.: Simple features and honest calibration for ambivalence and hesitancy recognition in video. arXiv preprint arXiv:2607.11120 (2026). https://doi.org/10.48550/arXiv.2607.11120
31. Liu, Y., et al.: RoBERTa: A robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692 (2019). https://doi.org/10.48550/arXiv.1907.11692
32. Qi, X., Wen, Y., Zhang, P., Huang, H.: MFGCN: Multimodal fusion graph convolutional network for speech emotion recognition. Neurocomputing **611**, 128646 (2025). https://doi.org/10.1016/j.neucom.2024.128646
33. Ren, H., Yang, W., Zhang, T., Zhang, Y.: Proposal-based multiple instance learning for weakly-supervised temporal action localization. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2394–2404 (2023). https://doi.org/10.1109/CVPR52729.2023.00237
34. Ryumina, E., et al.: Team RAS in 11th ABAW competition: Multimodal ambivalence recognition approach. arXiv preprint arXiv:2607.14702 (2026). https://doi.org/10.48550/arXiv.2607.14702
35. Schneider, I.K., Schwarz, N.: Mixed feelings: The case of ambivalence. Current Opinion in Behavioral Sciences **15**, 39–45 (2017). https://doi.org/10.1016/j.cobeha.2017.05.012
36. Shang, Y., Fu, T.: Multimodal fusion: A study on speech-text emotion recognition with the integration of deep learning. Intelligent Systems with Applications **24**, 200436 (2024). https://doi.org/10.1016/j.iswa.2024.200436
37. Smith, V.L., Clark, H.H.: On the course of answering questions. Journal of Memory and Language **32**(1), 25–38 (1993). https://doi.org/10.1006/jmla.1993.1002
38. Sun, L., Liu, B., Tao, J., Lian, Z.: Multimodal cross- and self-attention network for speech emotion recognition. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 4275–4279 (2021). https://doi.org/10.1109/ICASSP39728.2021.9414654
39. Swerts, M., Krahmer, E.: Audiovisual prosody and feeling of knowing. Journal of Memory and Language **53**, 81–94 (2005). https://doi.org/10.1016/j.jml.2005.02.003
40. Thompson, M.M., Zanna, M.P., Griffin, D.W.: Let's not be indifferent about (attitudinal) ambivalence. In: Attitude Strength: Antecedents and Consequences. Lawrence Erlbaum (1995). https://doi.org/http://doi.org/10.4324/9781315807041
41. Torre, D.L., et al.: Modeling engagement with a digital behavior change intervention (HeartSteps II): An exploratory system identification approach. Journal of Biomedical Informatics **158**, 104721 (2024). https://doi.org/10.1016/j.jbi.2024.104721
42. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems (2017). https://doi.org/10.48550/arXiv.1706.03762
43. Wang, J., Kong, D., Li, J., Wang, X., Yin, B.: Semantic prototype guided sparse temporal interaction for weakly supervised temporal action localization. ACM Transactions on Multimedia Computing, Communications, and Applications **22**(6) (2026). https://doi.org/10.1145/3807956