

Announcement of 2nd Edition: Ambivalence/Hesitancy (AH) Video Recognition Challenge ABAW10th - CVPR 2026

ABAW10th website (It will be updated soon): <https://affective-behavior-analysis-in-the-wild.github.io/10th>

Part 1: How to register?

AH Video Recognition Challenge

To participate in this challenge, please follow the registration procedure below:

Please fill out our [form](#) according to [these](#) steps, and submit it. It involves signing an EULA and uploading it through the same form. The form and the EULA must be completed and signed by a person holding a full-time faculty position at a university, higher education institution, or an equivalent organization. The signee cannot be a student (undergraduate, postgraduate, Ph.D., or postdoctoral).

Once the form is submitted, with the signed EULA, we will contact you to provide details for access to the BAH video dataset. The BAH dataset includes raw videos, cropped-aligned faces at each frame, video- and frame-level labels, audio transcripts with timestamps, annotators' cues, participants meta-data, pre-defined data splits (training, validation and test sets), and documentation.

Part 2: Details of the challenge

Ambivalence/Hesitancy (AH) Video Recognition Challenge

About Ambivalence and Hesitancy

Ambivalence and hesitancy (A/H), a closely related construct, is the primary reason why individuals delay, avoid, or abandon health behaviour changes. It is a subtle and conflicting emotion that sets a person in a state between positive and negative orientations, or between acceptance and refusal to do something. This often constitutes a barrier to initiating behaviour change and a trigger for discontinuing interventions or change efforts.

To detect A/H in videos, experts typically rely on recognizing conflicting emotions, positive and negative, using modality cues such as facial, language, audio, and body language cues. Commonly, emotional conflicts occur between modalities (cross-modality), making it relatively easy to detect. However, a conflict within the same modality could also lead to A/H, and its detection may be more challenging. In our recent work [1], we introduced video-based A/H recognition, along with a first dataset, named BAH, for A/H recognition in videos. Results indicate that standard deep multimodal models and fusion techniques provide a low level of accuracy, suggesting that they are not well equipped for this new task. We described several challenges in A/H recognition that should be considered in designing better models. We note that other combinations of modalities could be helpful as well, such as full body (full frame), and body- and head-pose. Teams can explore improving standard multimodal models, temporal modeling, multimodal alignment and specialized fusion to detect conflicting emotions (positive and negative affect) within and across modalities, in addition to multimodal LLMs with specialized parameter-efficient fine-tuning (PEFT). Domain adaptation could also be considered to personalize deep learning models, as the BAH dataset is based on participants [2, 3].

The first edition of our challenge concerns frame-level prediction of A/H which took place last year within [ABAW8 @ CVPR 2025](#), and it attracted several teams. We are excited to announce this second edition of our challenge along with [ABAW10 @ CVPR 2026](#). In this new edition, the challenge consists of building models for **video-level** prediction to answer the question: is there A/H in a given video? Teams can have access to our larger video BAH dataset with rich annotations. We are looking forward to your participation and innovative methods for the task of A/H recognition in videos.

Dataset

Upon registration for the AH video recognition challenge, teams will be granted access to a new, fully annotated at video- and frame-level version of the BAH dataset [1] that was collected for multimodal recognition of A/H in videos. It contains 1,427 videos with a total duration of 10.60 hours, captured from 300 participants across Canada, answering a predefined set of questions to elicit A/H. It is intended to mirror real-world online personalized behaviour change interventions. BAH is fully annotated by experts to provide timestamps that indicate where A/H occurs, and frame- and video-level annotations with A/H cues. Speech-to-text transcripts, their timestamps, cropped and aligned faces, and participants' metadata are

also provided. Since A and H manifest similarly in practice, we provide a binary annotation indicating the presence or absence of both A and H, without distinction.

The challenge aims at the design of innovative models to predict A/H at the **video-level** to indicate whether or not a video contains A/H (1: presence of A/H, 0: absence of A/H). Each participant in the dataset may have up to seven videos. The dataset is divided participant-wise into training, validation, and test sets. Teams will have access to the BAH dataset with full supervision. For performance evaluation, they can train their models on the BAH training set using any type of supervision, and report the performance on its public test set using the code in [`bah_metrics.py`](#). A second unlabeled private test set will be released to the teams before the end of the challenge. Teams must submit by email to the AH recognition challenge organizers a file of their predictions per-video using this private test set. They are allowed to provide multiple trials (up to 5 trials) within the week of the test period. We will compute the performance and the best trial will be used to rank teams and announce the winners. Teams can submit all 5 trials at once, or one trial at a time. This last option allows us to send teams the trial performance as feedback to adjust their approach if needed for the next trial. More details of the submission format will be communicated on the date of the test release.

Goal of the Challenge and Rules

Teams participating in this challenge will have access to the BAH video dataset [1] with full supervision. The dataset is composed of 1,427 videos from 300 participants, where each one contributed up to seven videos. Teams are required to develop their methods to recognize A/H at the video-level (binary task). Given a video, can we predict whether there is or not A/H? Different learning setups could be considered: supervised/self-supervised, domain adaptation and personalization, zero-/few-shot learning, etc. Standard multimodal models could be used, in addition to multimodal LLMs and other recent architectures. Teams are advised to develop solutions tailored for A/H recognition.

Teams are allowed to use any publicly available or private pre-trained model and any public or private dataset (that contains any type of annotations, e.g. valence/arousal, basic or compound emotions, action units). Other datasets for ambivalence/hesitancy, if available, could be used, in addition to the BAH dataset, but they must be disclosed in the paper.

Performance Assessment

Teams must submit by email to the AH recognition challenge organizers a file of their predictions per-video using this private test set. They are allowed to provide multiple trials (up to 5 trials) within the week of the test period. We will compute the performance and the best trial will be used to rank teams and announce the winners. Teams can submit all 5 trials at once, or one trial at a time. This last option allows us to send teams the trial performance as feedback to adjust their approach if needed for the next trial. More details of the submission format will be communicated on the date of the test release.

The performance measure (P) is the average F1 score (Macro F1) at the video level across both classes (presence (1) and absence (0) of A/H) over the private test set, and will be used to rank teams. We will also report the average precision score (AP) of the positive class (1).

Baseline Results

A performance of $P = 0.2827$ was obtained on the BAH public test set, using a baseline model (zero-shot setup with Multimodal-LLM (M-LLM), Video-LLaVA, with a simple prompt and vision modality only (code: <https://github.com/sbelharbi/zero-shot-m-llm-bah-prediction>)). See more details in [1].

Additionally, teams could build on top of standard multimodal models that leverage vision, audio, and text modality, such as the one used in [1], and adapt it from frame-level prediction to video-level prediction: <https://github.com/sbelharbi/bah-dataset>

Teams can explore improving standard multimodal models, temporal modeling, multimodal alignment, and multimodal LLMs with specialized parameter-efficient fine-tuning (PEFT). Domain adaptation and personalization could also be considered.

[1]: González-González M, Belharbi S, Zeeshan MO, Sharafi M, Aslam MH, Pedersoli M, Koerich AL, Bacon SL, Granger E. “BAH Dataset for Ambivalence/Hesitancy Recognition in Videos for Behavioural Change”. <https://arxiv.org/pdf/2505.19328>, ICLR, 2026.

[2]: Sharafi M, Belharbi S, Salem HB, Etemad A, Koerich AL, Pedersoli M, Bacon S, Granger E. “Personalized Feature Translation for Expression Recognition: An Efficient Source-Free Domain Adaptation Method”. <https://arxiv.org/pdf/2508.09202>. ICLR, 2026.

[3]: Zeeshan MO, Aslam MH, Belharbi S, Koerich AL, Pedersoli M, Bacon S, Granger E. “Subject-based domain adaptation for facial expression recognition”. <https://arxiv.org/pdf/2312.05632>, FG conference, 2024.