BAH DATASET FOR AMBIVALENCE/HESITANCY RECOGNITION IN VIDEOS FOR DIGITAL BEHAVIOURAL CHANGE

Anonymous authors
Paper under double-blind review

ABSTRACT

Ambivalence and hesitancy (A/H), a closely related construct, is the primary reasons 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. It manifests by a discord in affect between multiple modalities or within a modality, such as facial and vocal expressions, and body language. Although experts can be trained to recognize A/H as done for in-person interactions, integrating them into digital health interventions is costly and less effective. Automatic A/H recognition is therefore critical for the personalization and cost-effectiveness of digital behaviour change interventions. However, no datasets currently exists for the design of machine learning models to recognize A/H. This paper introduces the Behavioural Ambivalence/Hesitancy (BAH) dataset 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 predefined questions to elicit A/H. It is intended to mirror real-world online personalized behaviour change interventions. BAH is annotated by three experts to provide timestamps that indicate where A/H occurs, and frame- and video-level annotations with A/H cues. Video transcripts, cropped and aligned faces, and participants' meta-data are also provided. Since A and H manifest similarly in practice, we provide a binary annotation indicating the presence or absence of A/H. Additionally, this paper includes benchmarking results using baseline models on BAH for frame- and video-level recognition, zero-shot prediction, and personalization using source-free domain adaptation. The limited performance highlights the need for adapted multimodal and spatio-temporal models for A/H recognition. Results for specialized methods for fusion are shown to assess the presence of conflict between modalities, and for temporal modelling for within-modality conflict are essential for better A/H recognition. The data, code, and pretrained weights are publicly available.

1 INTRODUCTION

Emotion recognition plays a growing role in a range of health-related domains (Siddiqi et al., 2024), including disease prevention (Jin, 2024), diagnosis (Jiang et al., 2024; Maki et al., 2013), treatment monitoring (Dhuheir et al., 2021; Pepa et al., 2021; Suraj et al., 2022), and digital health promotion (Arabian et al., 2023; Subramanian et al., 2022), by supporting adaptive and responsive interventions (Liu et al., 2024b; Sinha et al., 2020). Emotion recognition technologies can support behaviour change interventions (Guo et al., 2024) by identifying affective states relevant to motivation, adherence, and engagement. Health-related behaviour change focuses on strategies to support individuals in adopting and maintaining healthy behaviours to prevent or manage chronic diseases, reduce early mortality, and improve mental health and well-being (Davidson & Scholz, 2020). Achieving and maintaining long-term behaviour change is a complex process. It often includes overcoming ambivalence and hesitancy (A/H) (McDonald et al., 2002; Michie et al., 2013a;b;

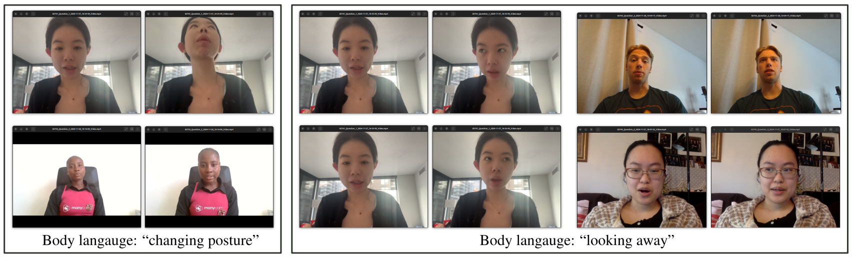


Figure 1: Examples of body language cues used by annotators to identify the occurrence of A/H: “looking away,” and “changing posture.”

Voisard et al., 2024), a closely related constructs, as they are the primary reason for individuals to delay, avoid, or abandon health behaviour changes (Conner & Armitage, 2008; Conner & Sparks, 2002; Manuel & Moyers, 2016; Miller & Rose, 2015; Van Gent et al., 2024; Williams, 2024). A/H is a subtle and conflicting emotion manifested by a discord in affect between multiple modalities or within a modality, such as facial and vocal expressions, and body language. This conflict sets a person in a state between positive and negative orientations, or between acceptance and refusal to do something, which often constitutes a barrier to initiating behaviour change and a trigger for discontinuing interventions or change efforts. Healthcare providers (e.g., clinicians, therapists) often identify A/H through a combination of speech and non-verbal cues (e.g., facial expressions and tone) (Heisel & Mongrain, 2004; Labbé et al., 2022; Miller & Rose, 2015) during in-person interactions. However, integrating them into digital interventions and ehealth is costly and less effective. Therefore, designing robust automated methods for A/H recognition can provide a cost-effective alternative that can adapt to individual users, and operate seamlessly within real-time, and resource-limited environments.

Recent research on machine learning (ML) in emotion recognition focuses mainly on seven basic discrete emotions, e.g, ‘Happy’, ‘Sad’, and ‘Surprised’ (Belharbi et al., 2024a; Liu et al., 2024a; Xue et al., 2022). Other models in the literature predict ordinal levels, including pain and stress estimation (Aslam et al., 2024; Chaptoukaev et al., 2023; Zeeshan et al., 2024; Nasimzada et al., 2024), or continuous predictions such as valence-arousal (Dong et al., 2024; Praveen & Alam, 2024a; Praveen et al., 2023; 2021). However, real-world scenarios present more complex cases of emotions. Recently, there has been an increased interest in designing robust affect models for compound emotions, a case where a mixture of basic emotions is manifested (Kollias, 2023; Richet et al., 2024). In particular, compound emotions commonly occur in daily interactions. However, they are more difficult to discern as they are subtle, ambiguous, and resemble basic emotions. A/H recognition is related to such a task where intention and attitudes are conflicted or in a in-between state, between willingness and resistance (MacDonald, 2015), or positive and negative affect (Armitage & Conner, 2000). This can manifest in how individuals express themselves and can be recognized (Hayashi et al., 2023) in their facial expression, tone, verbal, and body language (Figure 1). As a result, A/H exhibits a multimodal nature that comes as the result of subtle interconnection between different cues; in addition to manifesting within a modality. Unfortunately, such discord is extremely difficult to spot; a task that requires human training. This is a tedious and expensive procedure, leading to ineffective and less scalable eHealth interventions under limited resources. Integrating automated and reliable tools for A/H recognition can have a major impact in improving eHealth interventions. Although A/H is a common topic in behavioural science (Conner & Armitage, 2008; Hohman et al., 2016; Manuel & Moyers, 2016), it remains unexplored in the ML community, and as such, in the design of eHealth components. A possible reason is the lack of the necessary and specialized data for training and evaluation of ML models.

To address this limitation, we introduce in this work a first Behavioural Ambivalence/Hesitancy (`BAH`) dataset collected for subject-based multimodal recognition of A/H in videos. Through a collaboration between behavioural science and machine learning teams, we have collected a large video dataset, `BAH`, from 9 provinces in Canada. A data capture protocol is set in place to recruit diverse participants, including the development of a web-platform for video capturing, a dedicated storage server, and a specific annotation protocol. Our behavioural team designed seven questions to elicit responses regarding behaviours and to identify possible instances where participants are displaying A/H. Via our web-platform, participants are presented these questions and asked to record themselves while answering via their device camera with a microphone. Participants are guided in the platform by an avatar throughout the entire data capture session. Our dataset was developed with real-world digital health applications in mind, particularly for use in personalized behaviour change interventions where users interact with an avatar that prompts them with predefined questions. To mirror this setting, participants in our study responded to structured but genuine questions about behaviours they engage in or avoid, based on their actual experiences. This setup encourages authentic, spontaneous expression of ambivalence and hesitancy (A/H), while still maintaining consistency across participants. The BAH dataset was intentionally designed to maximize ecological validity. Participants had control over their environment and delivery, which contributed to naturalistic responses that reflect how people express A/H in real life.

The BAH dataset is composed of 300 participants. This amounts to a total of 1,427 videos (~ 10.60 hours) where 778 videos contain A/H (~ 1.79 hours). It has 916,618 total frames where 156,255 contain A/H. Three of our behavioural experts annotated the data at video- and frame-level to assess when A/H occurs. Because ambivalence and hesitancy manifest similarly in practice, the annotation is framed under a binary form (A/H vs. non-A/H). In addition, the video cues used by the annotators are reported such as facial expressions, body language, audio and language in addition to highlighting where there is inconsistency between the modalities. The BAH dataset is made public and it is provided with the raw videos with audio, cropped and aligned faces, detailed annotation/cues for video- and frame-level, audio transcript/timestamps/language, and participants meta-data such as age, ethnicity and more.

Our main contributions are summarized as follows. **(1)** A novel video dataset named BAH is proposed for automated subject-based multimodal recognition of A/H in videos that can be used in tools for digital health intervention systems. To mirror real-world digital interventions, participants in our study responded to structured but genuine questions about behaviours that elicit A/H using our online platform while recording themselves. Behavioural experts labelled the data at the video- and frame-levels under a binary form to indicate the presence/absence of A/H as they both manifest similarly. BAH can be used to develop and evaluate standard and personalized ML models for classification task, and build insights about A/H for digital behaviour change interventions. **(2)** Preliminary benchmarking results for baseline models on BAH for frame- and video-based emotion recognition. Results allowed exploring the impact of key factors, including the impact of using temporal context, multimodal information, and feature fusion. They also showed the need for specialized modalities fusion, and temporal modelling for better performance in the new task of A/H recognition. Baseline results are also shown for other tasks – zero-shot prediction and personalization through subject-based domain adaptation. Our code and dataset is public.

2 The BAH Dataset

2.1 Dataset Collection and Annotation

Capture. The BAH dataset contains Q&A videos. Its is constructed by collecting samples from participants over the age of 18 across Canada. Data collection and annotation process is presented in Figure 2. To proceed with the data collection, we developed "Automatic Expression Recognition" (AER) web-based platform (www.aerstudy.ca) where participants could record their responses to specific questions using their own computers or devices with camera and microphone. Users receive secure credentials to access the data collection platform, or they can create their own account. Participants first complete a brief survey to provide demographic information and indicate consent preferences (e.g., inclusion in secure datasets, challenges, or publications). They are then redirected to the AER platform, where they test their camera and microphone and choose an avatar to interact during data capture. The avatar guides them through seven questions. The session takes approximately 30 minutes. Participants are recruited and compensated via Prolific company (www.prolific.com), which also ensures population diversity and allows submission processing.

Participants answer seven questions designed by our behavioural team (Table 1), each one intended to elicit neutral, positive, negative, ambivalent, willing, resistant, and hesitant answers. Once the question is presented, the recording of the participant response starts. Skipping questions is allowed. At the end of each question, the participant has the option to rate their emotional response using a Likert-like 5-point scale. This self-rating is only employed for our analysis and does not serve as

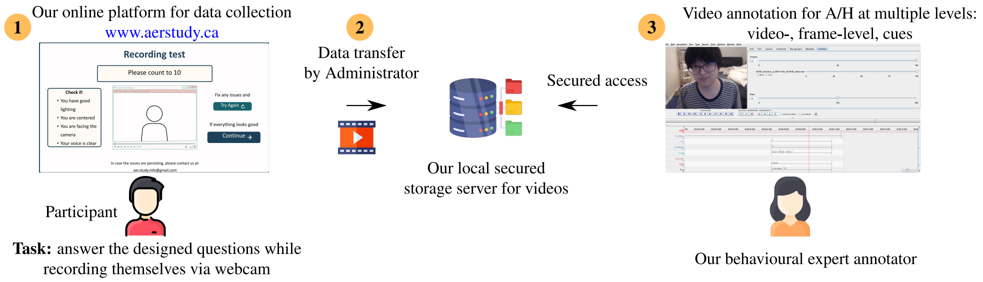


Figure 2: BAH dataset collection and annotation procedure. First, a participant access our web platform. They go through initial test/calibration to ensure the quality of the data. An avatar guides them throughout the entire process. Seven questions are presented to the participant. They are recorded while answering them. Once the data is captured, it is transferred by the Administrator to our local server. It is then annotated at several levels by an expert to determine when A/H occurs.

annotation. The order of the questions is randomized. In addition, participants are not aware what each questions is expected to illicit as emotion. During this capture procedure, several information is gathered including contact information of the participant, their demographics, consent, video recordings, survey responses, and software usage data (such as the time spent on each question). The participants' data is systematically downloaded and transferred to local secured server storage by the team for annotation and further analysis.

The study obtained human ethics approval from the two collaborating universities following all standard ethical practices. The dataset was collected between September 2024 and April 2025 in batches. This allowed us to adjust the targeted population (regarding participants' sex and Canadian province of residency) to ensure the dataset diversity.

We note that while BAH dataset is sourced exclusively from Canadian participants, its sample collection was designed to be representative in terms of province and sex, two key demographic variables relevant to health behaviour research in Canada. Additionally, the dataset includes individuals from a wide range of backgrounds, including different ethnicities and national origins, which reflects the cultural diversity of the Canadian population. This diversity, often lacking in many existing datasets, enhances the relevance and applicability of our findings within the Canadian context. Despite the geographic limitation, we consider this dataset a meaningful starting point for understanding ambivalence and hesitancy in health-related behaviour. our future plan is to expand to other countries and cultural settings to support broader generalizability.

Annotation. Three annotators were trained in expression recognition, specialized in identifying A/H, and in the annotation process of audio-visual data. A two-stage process was used: first, a

Question no.	Response	Prompt
1	Neutral	Tell us about an activity you commonly do after waking up.
2	Positive	Talk about an activity that brings you joy, for example, a hobby. Tell us why.
3	Negative	Talk about an activity you dislike doing, for example, a chore or something you find boring or annoying. Tell us why.
4	Ambivalent	Tell us about something you enjoy doing but wish you stopped doing (like a guilty pleasure) or something you don't do but wish you did.
5	Willing	Tell us about an activity you are almost always willing to do, for example with friends, at work, at home.
6	Resistant	Tell us about something people around you do, but that you would not be willing to do, for example, with friends, at work, at home.
7	Hesitant	Tell us about something you could have done already but haven't done yet, for example, something you are procrastinating or haven't made up your mind about.

Table 1: The 7 questions (prompts) designed by our experts to create our videos for BAH dataset. To avoid influencing the participants answers, they are only shown prompts without indicating the expected emotion/response.

global-level annotation determined the presence of A/H in each video; then, a frame-level annotation identified the precise segments where A/H occurred, specifying the start and end times (i.e., onset and offset) of each instance. Annotators also provided certainty ratings, and for some segments, indicated the cues that supported their judgment. To identify A/H, annotators tracked expressions across well-established modalities, (facial expressions, body language, audio, and language) and flagged cases where inconsistencies between modalities were observed. Each A/H segment has a timestamp start/end in addition to its own cues and modalities determined by an annotator allowing for better precision. Multiple A/H segments could be present within a video. We do not include an "apex" nor a continuous annotation, as ambivalence and hesitancy do not reliably exhibit a peak moment of maximum intensity. Instead, they tend to manifest as sustained or fluctuating states, making the concept of an apex incompatible with their typical temporal structure. The videos were annotated following a codebook created specifically for the study. Videos were annotated using the ELAN 8 (archive.mpi.nl/tla/elan) software (Figure 2).

The annotation process followed a structured training protocol supported by a detailed training manual. Annotators first received a conceptual introduction to A/H, followed by hands-on training in using the ELAN annotation software. Practical application was conducted using a standardized set of videos from the dataset. This phase also introduced annotators to the codebook, emphasizing the cue list, with examples spanning facial, vocal, verbal, and bodily expressions. Annotators received feedback and additional sessions were provided when further alignment was needed. Only after this training phase, and a final assessment, did annotators proceed to independent annotation. To ensure large labeled data, `BAH` dataset is divided among the three annotators where each video is labeled by one annotator. However, we conducted an analysis over 10 random videos with multi-annotator setup to assess inter-annotator agreement. We considered Fleiss's Kappa measure (Fleiss, 1971) for more than 2-annotator case. At global level, the score measure is 0.65 (substantial agreement) while at frame level, we have 0.41 (moderate agreement). When considering only videos where all three annotators agreed they have A/H, the score is 0.50.

To promote consistency, annotators were instructed to flag cases of uncertainty or complexity. These cases were discussed collaboratively, often through co-annotation. A consistent lead annotator facilitated resolution efforts, ensuring that decisions reflected a shared interpretation. In parallel, a comprehensive annotation protocol guided how videos were managed, accessed, and annotated. Annotators followed standardized procedures: (1) watch the video without taking notes to understand the participant and context; (2) re-watch the video to identify A/H segments and record start and end times; (3) reassess and refine selected segments; (4) identify and assign cues using the codebook; and (5) if needed, watch the video without audio or visual elements to isolate specific signals. Annotators were also encouraged to consult other videos from the same participant to establish expressive baselines in ambiguous cases.

The presence of A/H is assigned a single label (1), while its absence is assigned the label 0. Each video has a global- and frame-level label which can be used to train and evaluate ML models. The provided cues can also be used for interpretability aspect as well as to build insights on how people express A/H. The dataset is structured subject-wise which can be also useful for personalization training scenarios. The `BAH` dataset that is being made public contains only videos of participants consented for their data to be made public.

2.2 Dataset Variability

The dataset is designed to approximate the demographic distribution of sex and provincial representation in Canada. The `BAH` dataset is composed of 300 participant across Canada from nine provinces where 25.7% of participants is from British Columbia followed by Alberta with 19.7% and Ontario with 17.3% . All participants agreed to be part of this dataset. However, 61 participants (20.3%)[1] did not consent to be in publications while only seven participants (2.3%) did not consent to be part of challenges. The recorded videos are majority in English language and very few are in French language. Each participant can record up to seven videos where 113 participants have recorded the full seven videos. We obtained an average of ~ 4.75 videos/participant where each participant has an average of ~ 2.59 videos with A/H which is equivalent to ~ 520.85 frames of A/H (or ~ 21.49 seconds of A/H). The dataset amounts a total of 1,427 videos (~ 10.60 hours)

[1]The list of these participants is provided within the shared files of the `BAH` dataset.

where 778 videos contain A/H (~ 1.79 hours). This amounts to 916,618 total frames where 156,255 contain A/H. Since captured videos represent answers to questions, they are relatively short. BAH dataset has an average video duration of 26.76 ± 16.47 (seconds) with a minimum and maximum duration of 3 and 96 seconds.

An important characteristic of this dataset is the duration of the A/H segments in videos. BAH counts a total of 443 videos with multiple A/H segments and 332 videos with only one A/H segment. In total, there are 1,504 A/H segments. In particular, the duration of segments varies but it is brief with an average of 4.29 ± 2.45 seconds which is equivalent to 103.89 ± 58.70 frames. The minimum and maximum A/H segment is 0.004 seconds (1 frame), and 23.8 seconds (572 frames), respectively.

In terms of participants age, the dataset covers a large range from 18 to 74 years old. In particular, 37.7% of the participants covers the range 25-34 years, followed by the range of 35-44 years with 24.3%, then the range of 18-24 with 20.7%. In terms of sex, 52.0% are female, while 47.3 are male. As for ethnicity variation, White comes with 54.0% of the participants, followed by Asian with 21.0%, and Mixed with 10.7%, then Black with 9.7%. Large part of the participants are not students (67.0%) which limits common issues in recruit bias.

The public BAH dataset contains the row videos, detailed A/H annotation at video- and frame-level, cues, and per participant demographic information including age, birth country, Canada province where the participant lives, ethnicity, ethnicity simplified, sex, student status, consent to use recordings in publications. More details about the dataset diversity are provided in the appendix.

2.3 Ethical Consideration, Dataset Accessibility and Intended Uses

The collected data of human participants follows tightly ethical considerations. The project to collect BAH data was approved by ethical committees from both collaborating universities. Once recruited, participants have access to the full consent form prior to accessing the data capture platform and starting their data capture procedure. They are provided with details of the study, as well as a list of the potential risks and benefits of participating in the study. They are be instructed to read the consent form thoroughly and they are provided with a clear and simple video that summarizes the consent form. Participants are then able to decide the type of access they want the researcher to have to their audiovisual data, including if they want their images to be used for publications and presentations. In addition, these options are presented again at the end of the data capture procedure, just in case they change their mind around their participation in the study or the use of their data after they have finished recording their responses. At the end of the study, participants receive, via email, a copy of the consent form that includes their choices about data usage and the contact information for the team should they have any further questions. Note that participants are given numerical codes for anonymity.

Following the guidelines of the funding agency , the BAH dataset is made public with open credentialed access for research purposes. To access the dataset, users are required to fill in a request form and sign an End-User License Agreement (EULA) as commonly done to ensure dataset security. Upon access approval, the user will receive a link to download the full dataset, including row videos, detailed annotation, cues, participants' meta-data, cropped-and-aligned-faces, frames, audio transcripts. BAH uses a proprietary license for research purposes. The dataset is hosted in a secured server as it is intended for long-term availability. Our public code is under an open-source license (BSD-3-Clause license). The code website will be used as a permanent page for the dataset that will reflect any future updates. Despite all our precautions, our dataset may still could be misused. We consider a thorough review of requests before granting data access. Reviewers can directly download the BAH dataset via the link provided in the appendix. Please read Sec.A in appendix before proceeding to download the dataset.

Our primary goal of building BAH dataset is to make public a first and unique dataset for A/H recognition in videos. Given the content of the dataset, its multimodal aspect, and the provided annotation, it can be used to train and evaluate ML models for A/H recognition in videos at frame- and/or video-level with different learning scenarios. Since data is subject-based, it can also be used for personalization using domain adaptation, for instance. The provided cues used by annotators can also be used for interpretability learning, and further analysis to get more insight on our understanding of A/H in human behaviours. Such understanding and recognition of A/H can be leveraged in downstream tasks such as behavioural change, interventions and recommendations in clinics or via automated systems such as virtual trainers/assistants.

2.4 Experimental Protocol

Dataset split. The dataset is divided randomly based on participants into 3 sets: train (195 participants), validation (30 participants) and test (75 participants) set. We ensured that the 3 splits represent the total data distribution. The train and validation sets amounts to 3/4 of the total participants, while 1/4 goes to the test set. Videos of one participants belong to one and one set only. The details of each set is presented in Table 2. The split files are provided along with the dataset files. They contain the split in terms of videos and frames ready to use. Note that the dataset is highly imbalanced as depicted in Table 3, especially at frame level where only 17.04% contains A/H. This factor should be accounted for during training and evaluation. The dataset can be used for training at video- and/or frame-level. The participant identifiers are provided in the splits allowing subject-based learning scenarios.

Data subsets	Train	Validation	Test	Total
Number participants	195	30	75	300
Number participants with A/H	144	27	75	246
Number videos	778	124	525	1427
Number videos with A/H	385	75	318	778
Number frames	501,970	79,538	335,110	916,618
Number frames with A/H	76,515	13,984	65,756	156,255
Total duration (hour)	5.80	0.92	3.87	10.60
Total duration with A/H (hour)	0.87	0.16	0.75	1.79

Table 2: `BAH` dataset split into train, validation, and test sets.

Data subsets	Train (%)	Validation (%)	Test (%)	Total (%)
Participants with A/H	73.85	90.00	100.00	82.00
Videos with A/H	49.49	60.48	60.57	54.51
Frames with A/H	15.24	17.58	19.62	17.04
Duration with A/H	15.09	17.41	19.44	16.88

Table 3: Imbalance rate of `BAH` dataset split across train, validation, and test sets: (Total # items with A/H)/(Total # items).

Evaluation metrics. We refer here to the positive class as the class with label 1 indicating the presence of A/H, while negative class is the class 0 indicating the absence of A/H. To account for the imbalance in `BAH` dataset, we use adapted standard evaluation metrics: - `F1` score of the positive class. - Average `F1` (`AVGF1`) score which is the unweighted mean of `F1` of the positive and negative class. - Average precision score (`AP`) of the positive class which accounts for the performance sensitivity to the model's confidence. For `AP` score, a threshold list between 0 and 1 is used with a step of 0.001. Evaluation code of all measures is provided along with the public code of this dataset.

3 Baseline Results

This section provides preliminary results of different baseline models on our `BAH` dataset. In particular, we provide performance of models for the supervised frame-level classification task. We consider a 2-class classification problem where each frame is annotated, and models predict two outputs: one for the positive class (presence of A/H), and a second for the absence of A/H. Supervised video-level classification performance is included in the appendix. In addition, the results of other tasks – zero-shot prediction, and personalization through unsupervised domain adaptation – are also included.

We initially focus on the impact of using single vs multimodal learning for frame-level classification. Then, the performance of different individual modalities are explored, along with their multimodal fusion. In addition, we investigate the impact of temporal modelling and context vs single frame learning. In the following, we present the pre-processing of the three different used modalities:

	Without context		With context (TCN)	
Backbone	AVGF1	AP	AVGF1	AP
APViT (Xue et al., 2022)	0.5051	0.1906	0.5019	0.2069
ResNet18 (He et al., 2016)	0.5074	0.1940	0.5079	0.1993
ResNet34 (He et al., 2016)	**0.5138**	0.1952	0.4998	0.1984
ResNet50 (He et al., 2016)	0.4737	0.1942	0.4985	0.1915
ResNet101 (He et al., 2016)	0.4929	**0.1967**	**0.5165**	**0.2070**
ResNet152 (He et al., 2016)	0.4889	0.1843	0.5084	0.2058

Table 4: Visual modality performance on test set of BAH at frame-level classification: impact of architecture and context.

visual (facial), audio (vocal), and text transcripts (textual), and describe the baseline models used in each case.

3.1 Pre-processing of Modalities

1) Visual. All frames from each video are extracted, and for each frame, faces are located using RetinaFace model (Deng et al., 2019), cropped, then aligned. The face with the highest score is stored in case of multiple faces are detected in a frame. Faces are resized to 256×256 and stored as RGB images with a file name that maintains the order of frames. The video frame rate is 24 FPS. **2) Audio**. We follow standard procedure to process audio data (Praveen & Alam, 2024b; Richet et al., 2024; Zhang et al., 2023b). For audio modality, we first convert videos to single audio channels (mono) with a 16k sampling rate into wav format. The log melspectrograms features are extracted using Vggish model (Hershey et al., 2017)(github.com/harritaylor/torchvggish). A hope of 1/FPS of the raw video is used to extract the spectrograms to synchronize audio with other modalities. **3) Text**. The collected data captures the audio of participants. We consider audio transcripts as an extra modality that can help recognizing A/H since text is a significant cue used by annotators. To this end, we transcribe the audio of each video, and detect the language using Whisper model (Radford et al., 2023) (Whisper large-v3 multilingual: huggingface.co/openai/whisper-large-v3). We provide the timestamp of each transcript. Word-level features are then extracted using BERT Base Uncased model (Devlin et al., 2019)(pypi.org/project/pytorch-pretrained-bert/). A word may span more than one frame. To synchronize with other modalities, a word-level features is repeated per its timestamp for all the frames that correspond to the word.

Note that both cropped and aligned faces, and video transcripts are shipped with the shared public BAH dataset. Researchers can choose to use them or build their own.

3.2 Pre-training of Visual Backbone

For audio and text modality, features are extracted offline and stored as described above. For visual modality, we explore different architectures including CNN- and ViT-based(Dosovitskiy et al., 2021). In particular, we explore ResNet family including ResNet18, 34, 50, 101, and 152 (He et al., 2016). For ViT family, we consider a recent model designed for basic emotion recognition, APViT (Xue et al., 2022). First, we pre-train each model on basic emotion recognition task including these emotions: “Anger”, “Disgust”, “Fear”, “Happiness”, “Sadness”, “Surprise”. To this end, we collected a large mixed dataset composed of 3 public common datasets for emotion recognition using images: RAF-DB (Li et al., 2017), and AffectNet (Mollahosseini et al., 2019), and Aff-wild2 (Kollias & Zafeiriou, 2019). This amounts to more than 0.54 millions training images. Models are trained for basic emotion classification for 60 epochs with a batch size of 1,424 samples using 4 parallel NVIDIA A100 GPUs with 40 GB of memory. Standard cross-entropy loss and Stochastic Gradient descent (SGD) are used for training. Once pretrained, each model is further fine tuned on our BAH train set for A/H recognition. To account for class imbalance, we perform under-sampling of negative class over train set. This is achieved by randomly sampling the negative class samples to be the same as the positive class. The weights of both models will be made public. The backbones of each model are used later for feature extraction of visual modality. Note that we compared variant visual modalities including cropped faces, full-frame, and head-pose. Results show that cropped faces yield better results. Details is included in supplementary materials.

3.3 Importance of Contextual Learning

In this section, we aim to answer the question whether context modelling can help better A/H recognition. This is particularly interesting since A/H does not occur instantly, but within a context. Text and audio modality already capture context in their features making using them to answer this question less efficient. However, we can obtain frame features with and without any context. Therefore, we consider visual modality to answer our question. To this end, we study different visual backbones.

In the case of modelling without context, models simply train on independent frames without considering any context or dependency between them. Inference is done in the same way. In the case of modelling with context, both training and inference leverages temporal dependency between frames. To this end, a window of adjacent frames is fed to the model. We then use temporal convolutional network (TCN) (Bai et al., 2018) after the visual backbone to capture relations between frames embeddings. The window length defines the extent of the context. Table 4 shows the obtained results. Regardless of the context, we observe a very low performance over `AP` highlighting the difficulty of recognizing A/H based on images alone. In particular `AP` is below 0.2070. On the other hand, `AVGF1` reaches 0.5138. As a reference, predicting every frame as negative class yields an `AVGF1` of 0.4459. We note that overall, using context boost performance of all metrics across all architectures. This is expected as A/H does not usually occur at a single frame but withing a context. This makes its recognition from a single frame challenging. We recall that the average A/H segments spans 103 frames (or 4.29 seconds). Future works should account for temporal context for better performance. However, as we will show in the ablations in the appendix, very large context could lead to poor performance. Note that large ResNet models seems to yield better performance overall. Unless mentioned otherwise, all our next experiments will use ResNet101.

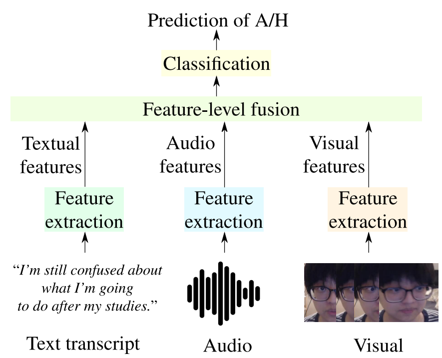


Figure 3: Multimodal model used to produce baseline performance (Richet et al., 2024).

3.4 Multimodal Baselines

Since A/H is multimodal by nature, we explore the impact of using different modalities, including visual, audio, and transcript using the model presented in Figure 3 (Richet et al., 2024). Results are reported in Table 5. Using text modality alone yields better performance, `AP` of 0.2510 and `AVGF1` of 0.5497, compared to visual or text modalities over both metrics. Combining a pair of modalities improves furthermore the performance to 0.5586 for `AVGF1`, and 0.2609 for `AP` with the case of audio and text. Combing the three modalities slightly reduces performance in terms of `AP`. This may suggest that better and more adapted fusion techniques are needed to recognized conflicts between modalities.

Modalities	`AVGF1`	`AP`
Visual	0.5165	0.2070
Audio	0.4658	0.2238
Text	0.5497	0.2519
Visual + Audio	0.5205	0.2225
Visual + Text	0.5547	0.2479
Audio + Text	**0.5586**	**0.2609**
Visual + Audio + Text	0.5502	0.2548

Table 5: Multimodal models performance on test set of `BAH` at frame-level classification. For visual modality, ResNet101 backbone is used.

Table 6 shows the impact of using different feature fusing techniques including simple concatenation (CAN) (Zhang et al., 2023b), co-attention (LFAN) (Zhang et al., 2023b), transformer-based fusion (MT) (Waligora et al., 2024), and cross-attention fusion (JMT) (Waligora et al., 2024). We observe that the way of leveraging the interaction between the three modalities is a key factor. LFAN and CAN fusion lead over both metrics. Future works should pursue more adapted methods to A/H. Ambivalence and hesitancy are usually expressed as a conflict between willingness and resistance. This can be perceived through a parallel affect conflict between modalities and or within modalities.

For instance, a participant could say a sentence to convey a meaning but their facial expression, body behaviour, or tone may carry a contradictory emotion. Understanding such subtly and interconnection between different cues in different modalities could play an important role in designing robust methods for A/H recognition in videos.

We believe our new and unique dataset has brought a new challenging research direction to better understand complex and subtle human emotions that is A/H. Given the multimodal nature of A/H, our BAH dataset provides an essential and valuable toolkit for the research community to design and evaluate their methods. Important key and critical downstream tasks could potentially benefits from these methods including but not limited to clinical interviews, interventions, behavioural changes, and automated assistants such as online trainers. Our preliminary results suggest that leveraging context, multimodality, and their fusion could lead to better A/H recognition performance.

Fusion type	AVGF1	AP
LFAN (Zhang et al., 2023b) *(cvprw,2023)*	0.5502	0.2548
CAN (Zhang et al., 2023b) *(cvprw,2023)*	**0.5526**	**0.2631**
MT (Waligora et al., 2024) *(cvprw,2024)*	0.5137	0.2134
JMT (Waligora et al., 2024) *(cvprw,2024)*	0.5241	0.2139

Table 6: Feature fusion performance on test set of BAH at frame-level classification.

4 Recommendations for Future Works in A/H Recognition

The newly introduced A/H recognition task requires the model to be able to detect affect conflict cross-modalities in addition to within-modality. Standard multimodal models are trained to yield predictions aligned with the output supervision. This automatic training may focus on learning label patterns and miss acquiring a mechanism to understand affect conflict. To build more interpretable A/H recognition systems, we recommend a 2-level framework. The first level should focus on modelling affect per modality in an independent way. Off-the-shelf pretrained sentiment analysis (Sharma et al., 2025) models could be used. This first level should be separated from A/H since we can not detect it at modality level yet, at least for the cross-modality case. At the second level, a dedicated fusion mechanism should be used to assess whether there is affect conflict cross-modalities to make a decision. This module does a deeper work than simply fusing features as commonly done. It should acquire an understanding of affect conflict to be able to detect it. Such modular and interpretable framework allows introducing priors about affect conflicts.

Statistics extracted from annotators cues could be leveraged to constrain the model find conflicts cross-modalities. A specialized temporal modelling should be used to detect within-modality conflict based on the output of the first level. Statistics from A/H segments durations should be considered as A/H happens briefly. Context is also important to detect within-modality cases. Segmented body could help as well since it contains important cues and less noise compared to full frame. Our results showed that standard multimodal models and fusion technique yield modest performance. Future works should focus on designing specialized frameworks for A/H recognition.

5 Conclusion

This work introduces a new and unique multimodal and subject-based video dataset, BAH, for A/H recognition in videos. BAH contains 300 participants across 9 provinces in Canada. Recruited participants answer 7 designed questions to elicit A/H while recording themselves via webcam and microphone via our web-platform. The dataset amounts to 1,118 videos for a total duration of 10.60 hours with 1.79 hours of A/H. It was annotated by our behavioural team at video- and frame-level.

Our initial benchmarking yielded limited performance highlighting the difficulty of A/H recognition. Our results showed also that leveraging context, multimodality, and adapted feature fusion is a first good direction to design robust models. Our dataset and code are made public.

The following appendix contains related work, more detailed and relevant statistics about the datasets and its diversity, dataset limitations, implementation details, and additional results.

REFERENCES

H. Arabian, T. Abdulbaki Alshirbaji, R. Schmid, V. Wagner-Hartl, J. Chase, and K. Moeller. Harnessing wearable devices for emotional intelligence: Therapeutic applications in digital health. *Sensors*, 23(19):8092, 2023.

C. J Armitage and M. Conner. Attitudinal ambivalence: A test of three key hypotheses. *Personality and Social Psychology Bulletin*, 26(11):1421–1432, 2000.

M.H. Aslam, M.O. Zeeshan, S. Belharbi, M. Pedersoli, Al.L. Koerich, S. Bacon, and E. Granger. Distilling privileged multimodal information for expression recognition using optimal transport. In *International Conference on Automatic Face and Gesture Recognition (FG)*, 2024.

S. Bai, J.Z. Kolter, and V. Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *CoRR*, abs/1803.01271, 2018.

S. Belharbi, M. Pedersoli, A. L. Koerich, S. Bacon, and E. Granger. Guided interpretable facial expression recognition via spatial action unit cues. In *International Conference on Automatic Face and Gesture Recognition (FG)*, 2024a.

S. Belharbi, M. Pedersoli, A.L. Koerich, S. Bacon, and E. Granger. Spatial action unit cues for interpretable deep facial expression recognition. In *AI and Digital Health Symposium*, 2024b.

M. Bishay, P. Palasek, S. Priebe, and I. Patras. Schinet: Automatic estimation of symptoms of schizophrenia from facial behaviour analysis. *IEEE Transactions on Affective Computing*, 12(4): 949–961, 2019.

J. Bonnard, A. Dapogny, F. Dhombres, and K. Bailly. Privileged attribution constrained deep networks for facial expression recognition. In *ICPR*, 2022.

E.H. Bradley, L.A. Curry, and K.J. Devers. Qualitative data analysis for health services research: developing taxonomy, themes, and theory. *Health services research*, 42(4):1758–1772, 2007.

C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J.N. Chang, S. Lee, and S.S. Narayanan. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42:335–359, 2008.

H. Chaptoukaev, V. Strizhkova, M. Panariello, B. Dalpaos, A. Reka, V. Manera, S. Thummler, E. ISMAILOVA, N. Evans, F. Bremond, M. Todisco, M. A Zuluaga, and L. M. Ferrari. StressID: a multimodal dataset for stress identification. In *NeurIPS*, 2023.

J. Choe, S.J. Oh, S. Chun, S. Lee, Z. Akata, and H. Shim. Evaluation for weakly supervised object localization: Protocol, metrics, and datasets. *TPAMI*, pp. 1–1, 2022.

Y. Chu, L. Liao, Z. Zhou, C.-W. Ngo, and R. Hong. Towards multimodal emotional support conversation systems. *CoRR*, abs/2408.03650, 2024.

M. Conner and C.J. Armitage. Attitudinal ambivalence. 2008.

M. Conner and P. Sparks. Ambivalence and attitudes. *European review of social psychology*, 12(1): 37–70, 2002.

K.W. Davidson and U. Scholz. Understanding and predicting health behaviour change: a contemporary view through the lenses of meta-reviews. *Health psychology review*, 14(1):1–5, 2020.

M. De-la Torre, E. Granger, P. VW Radtke, R. Sabourin, and D.O. Gorodnichy. Partially-supervised learning from facial trajectories for face recognition in video surveillance. *Information fusion*, 24: 31–53, 2015.

J. Deng, J. Guo, Y. Zhou, J. Yu, I. Kotsia, and S. Zafeiriou. Retinaface: Single-stage dense face localisation in the wild. *CoRR*, abs/1905.00641, 2019.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, pp. 4171–4186, 2019.

M. Dhuheir, A. Albaseer, E. Baccour, A. Erbad, M. Abdallah, and M. Hamdi. Emotion recognition for healthcare surveillance systems using neural networks: A survey. In *2021 International Wireless Communications and Mobile Computing (IWCMC)*, pp. 681–687, 2021.

L. Dong, X. Wang, S. Setlur, V. Govindaraju, and I. Nwogu. Ig3d: Integrating 3d face representations in facial expression inference. *CoRR*, abs/2408.16907, 2024.

A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

Y. Fan, J. Lam, and V. Li. Facial action unit intensity estimation via semantic correspondence learning with dynamic graph convolution. In *AAAI*, 2020.

J.L. Fleiss. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5): 378, 1971.

R. Guo, H. Guo, L. Wang, M. Chen, D. Yang, and B. Li. Development and application of emotion recognition technology—a systematic literature review. *BMC psychology*, 12(1):95, 2024.

X. Guo, B. Zhu, L.F. Polanía, C. Boncelet, and K.E. Barner. Group-level emotion recognition using hybrid deep models based on faces, scenes, skeletons and visual attentions. In *Proceedings of the 20th ACM international conference on multimodal interaction*, pp. 635–639, 2018.

J.A. Hall, J.A. Harrigan, and R. Rosenthal. Nonverbal behavior in clinician—patient interaction. *Applied and preventive psychology*, 4(1):21–37, 1995.

J. Han, L. Xie, J. Liu, and X. Li. Personalized broad learning system for facial expression. *Multimedia Tools and Applications*, 2020.

D. Hayashi, S.D.L. Carvalho, P.A.B. Ribeiro, R.C.M. Rodrigues, T.M. São-João, K. Lavoie, S. Bacon, and M. E. Cornelio. Methods to assess ambivalence towards food and diet: a scoping review. *Journal of Human Nutrition and Dietetics*, 36(5):2010–2025, 2023.

K. He, X. Zhang, S.g Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

Z. He, Z. Li, F. Yang, L. Wang, J. Li, C. Zhou, and J. Pan. Advances in multimodal emotion recognition based on brain–computer interfaces. *Brain sciences*, 10(10):687, 2020.

M.J. Heisel and M. Mongrain. Facial expressions and ambivalence: Looking for conflict in all the right faces. *Journal of Nonverbal Behavior*, 28:35–52, 2004.

S. Hershey, S. Chaudhuri, D.P.W. Ellis, J.F. Gemmeke, A. Jansen, R.C. Moore, M. Plakal, D. Platt, R.A. Saurous, B. Seybold, M. Slaney, R.J. Weiss, and K.W. Wilson. Cnn architectures for large-scale audio classification. In *ICASSP*, 2017.

Z.P. Hohman, W.D. Crano, and E.M. Niedbala. Attitude ambivalence, social norms, and behavioral intentions: Developing effective antitobacco persuasive communications. *Psychology of Addictive Behaviors*, 30(2):209, 2016.

S. Hornstein, K. Zantvoort, U. Lueken, B. Funk, and K. Hilbert. Personalization strategies in digital mental health interventions: a systematic review and conceptual framework for depressive symptoms. *Frontiers in digital health*, 5:1170002, 2023.

G. M. Jacob and B. Stenger. Facial action unit detection with transformers. In *CVPR*, 2021.

D. Jiang, L. Yan, and F. Mayrand. Emotion expressions and cognitive impairments in the elderly: review of the contactless detection approach. *Frontiers in Digital Health*, 6:1335289, 2024.

H. Jin. A comparative analysis of single and multi-modality-based emotion recognition for disease prevention. In *International Conference on Artificial Intelligence and Communication (ICAIC)*, volume 185, pp. 323, 2024.

R. Khanna, N. Robinson, M. O'Donnell, H. Eyre, and E. Smith. Affective computing in psychotherapy. *Advances in Psychiatry and Behavioral Health*, 2(1):95–105, Sep 2022.

D. Kollias. Multi-label compound expression recognition: C-expr database & network. In *CVPR*, 2023.

D. Kollias and S. Zafeiriou. Expression, affect, action unit recognition: Aff-wild2, multi-task learning and arcface. *CoRR*, 2019.

D. Kollias, P. Tzirakis, A. S. Cowen, S. Zafeiriou, I. Kotsia, E. Granger, M. Pedersoli, S. L. Bacon, A. Baird, C. Gagne, C. Shao, G. Hu, S. Belharbi, and M. H. Aslam. Advancements in Affective and Behavior Analysis: The 8th ABAW Workshop and Competition. 3 2025. doi: 10.6084/m9.figshare.28524563.v4.

J. Kossaifi, R. Walecki, Y. Panagakis, J. Shen, M. Schmitt, F. Ringeval, J. Han, V. Pandit, A. Toisoul, B. Schuller, et al. Sewa db: A rich database for audio-visual emotion and sentiment research in the wild. *TPAMI*, 43(3):1022–1040, 2019.

K. Kraack. A multimodal emotion recognition system: Integrating facial expressions, body movement, speech, and spoken language. *CoRR*, abs/2412.17907, 2024.

S. Labbé, I. Colmegna, V. Valerio, V.G. Boucher, S. Peláez, A.I. Dragomir, C. Laurin, E.M. Hazel, S.L. Bacon, and K.L. Lavoie. Training physicians in motivational communication to address influenza vaccine hesitation: a proof-of-concept study. *Vaccines*, 10(2):143, 2022.

I. Lee, E. Lee, and S.B. Yoo. Latent-ofer: Detect, mask, and reconstruct with latent vectors for occluded facial expression recognition. In *ICCV*, 2023.

S. Li and W. Deng. Deep emotion transfer network for cross-database facial expression recognition. In *ICPR*, 2018.

S. Li, W. Deng, and J. Du. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *CVPR*, 2017.

J. Liang, D. Hu, and J. Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *ICML*, 2020.

B. Lin, Y. Ye, B. Zhu, J. Cui, M. Ning, P. Jin, and L. Yuan. Video-llava: Learning united visual representation by alignment before projection, 2024.

C. Liu, X. Zhang, X. Liu, T. Zhang, L. Meng, Y. Liu, Y. Deng, and W. Jiang. Facial expression recognition based on multi-modal features for videos in the wild. In *CVPR*, 2023a.

D. Liu, H. Zhang, and P. Zhou. Video-based facial expression recognition using graph convolutional networks. In *ICPR*, 2021a.

H. Liu, R. An, Z. Zhang, B. Ma, W. Zhang, Y. Song, Y. Hu, W. Chen, and Y. Ding. Norface: Improving facial expression analysis by identity normalization. *ECCV*, 2024a.

X. Liu, L. Jin, X. Han, J. Lu, J. You, and L. Kong. Identity-aware facial expression recognition in compressed video. In *ICPR*, 2021b.

Y. Liu, W. Wang, Ch. Feng, H. Zhang, Z. Chen, and Y. Zhan. Expression snippet transformer for robust video-based facial expression recognition. *Pattern Recognition*, 138:109368, 2023b.

Y. Liu, Y. Zhang, and Y. Wang. Application of deep learning-based image processing in emotion recognition and psychological therapy. *Traitement du Signal*, 41(6):2923, 2024b.

H. Lokhande, C. Garware, T. Kudale, and R. Kumar. Personalized well-being interventions (pwis): A new frontier in mental health. In *Affective Computing for Social Good: Enhancing Well-being, Empathy, and Equity*, pp. 183–200. 2024.

I. Loshchilov and F. Hutter. SGDR: stochastic gradient descent with warm restarts. In *ICLR*, 2017.

C. Luo, S. Song, W. Xie, L. Shen, and H. Gunes. Learning multi-dimensional edge feature-based AU relation graph for facial action unit recognition. In *IJCAI*, 2022.

N.E. MacDonald. Vaccine hesitancy: Definition, scope and determinants. *Vaccine*, 33(34):4161–4164, 2015.

Y. Maki, H. Yoshida, T. Yamaguchi, and H. Yamaguchi. Relative preservation of the recognition of positive facial expression "happiness" in alzheimer disease. *International Psychogeriatrics*, 25 (1):105–110, 2013.

J.K. Manuel and T.B. Moyers. The role of ambivalence in behavior change. *Addiction*, 111(11): 1910–1912, November 2016.

J. Mao, R. Xu, X. Yin, Y. Chang, B. Nie, A. Huang, and Y. Wang. Poster++: A simpler and stronger facial expression recognition network. *Pattern Recognition*, pp. 110951, 2024.

H.P. McDonald, A.X. Garg, and R.B. Haynes. Interventions to enhance patient adherence to medication prescriptions: scientific review. *Jama*, 288(22):2868–2879, 2002.

S. Michie, M. Richardson, M. Johnston, C. Abraham, J. Francis, W. Hardeman, M.P. Eccles, J. Cane, and C.E. Wood. The behavior change technique taxonomy (v1) of 93 hierarchically clustered techniques: building an international consensus for the reporting of behavior change interventions. *Annals of behavioral medicine*, 46(1):81–95, 2013a.

S. Michie, R. West, and B. Spring. Moving from theory to practice and back in social and health psychology. 2013b.

W.R. Miller and G.S. Rose. Motivational interviewing and decisional balance: contrasting responses to client ambivalence. *Behavioural and cognitive psychotherapy*, 43(2):129–141, 2015.

J.A. Miranda Calero, L. Gutiérrez-Martín, E. Rituerto-González, E. Romero-Perales, J.M. Lanza-Gutiérrez, C. Peláez-Moreno, and C. López-Ongil. Wemac: Women and emotion multi-modal affective computing dataset. *Scientific data*, 11(1):1182, 2024.

A. Mollahosseini, B. Hassani, and M. H. Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Trans. Affect. Comput.*, 10(1):18–31, 2019.

S. Murtaza, S. Belharbi, M. Pedersoli, and E. Granger. A realistic protocol for evaluation of weakly supervised object localization. In *WACV*, 2025.

J. Nasimzada, J. Kleesiek, K. Herrmann, A. Roitberg, and C. Seibold. Towards synthetic data generation for improved pain recognition in videos under patient constraints. *CoRR*, abs/2409.16382, 2024.

A. O'Donnell, M. Addison, L. Spencer, H. Zurhold, M. Rosenkranz, R. McGovern, E. Gilvarry, M.-S. Martens, U. Verthein, and E. Kaner. Which individual, social and environmental influences shape key phases in the amphetamine type stimulant use trajectory? a systematic narrative review and thematic synthesis of the qualitative literature. *Addiction*, 114(1):24–47, 2019.

C. Ortiz, T. López-Cuadrado, A. Ayuso-Álvarez, C. Rodríguez-Blázquez, and I. Galán. Co-occurrence of behavioural risk factors for non-communicable diseases and mortality risk in spain: a population-based cohort study. *BMJ open*, 15(1):e093037, 2025.

M. Pantic and L.J.M. Rothkrantz. Toward an affect-sensitive multimodal human-computer interaction. *Proceedings of the IEEE*, 91(9):1370–1390, 2003.

L. Pepa, L. Spalazzi, M. Capecci, and M. G. Ceravolo. Automatic emotion recognition in clinical scenario: a systematic review of methods. *IEEE Transactions on Affective Computing*, 14(2): 1675–1695, 2021.

S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Conference of the Association for Computational Linguistics*, pp. 527–536, 2019.

R. G. Praveen and Ja. Alam. Inconsistency-aware cross-attention for audio-visual fusion in dimensional emotion recognition. *CoRR*, abs/405.12853, 2024a.

R. G. Praveen, E. Granger, and P. Cardinal. Cross attentional audio-visual fusion for dimensional emotion recognition. In *International Conference on Automatic Face and Gesture Recognition (FG 2021)*, 2021.

R. G. Praveen, P. Cardinal, and E. Granger. Audio–visual fusion for emotion recognition in the valence–arousal space using joint cross-attention. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 5(3):360–373, 2023.

R.G. Praveen and J. Alam. Recursive joint cross-modal attention for multimodal fusion in dimensional emotion recognition. In *ABAW workshop at CVPR*, 2024b.

A. Radford, J. Wook Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever. Robust speech recognition via large-scale weak supervision. In *ICLM*, 2023.

N. Richet, S. Belharbi, H. Aslam, O. Zeeshan, A. L. Koerich, M. Pedersoli, S. Bacon, and E. Granger. Textualized and feature-based models for compound multimodal emotion recognition in the wild. In *ABAW workshop at ECCV*, 2024.

F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne. Introducing the recola multimodal corpus of remote collaborative and affective interactions. In *International conference and workshops on automatic face and gesture recognition (FG)*, pp. 1–8, 2013.

E. Sánchez-Lozano, G. Tzimiropoulos, and M. F. Valstar. Joint action unit localisation and intensity estimation through heatmap regression. In *BMVC*, 2018.

D. Sejdinovic, B. Sriperumbudur, A. Gretton, and K. Fukumizu. Equivalence of distance-based and rkhs-based statistics in hypothesis testing. *The annals of statistics*, pp. 2263–2291, 2013.

M. Sharafi, S. Belharbi, H.B. Salem, A. Etemad, A.L. Koerich, M. Pedersoli, S. Bacon, and E. Granger. Personalized feature translation for expression recognition: An efficient source-free domain adaptation method. *CoRR*, abs/2508.09202, 2025a.

M. Sharafi, E. Ollivier, M.O. Zeeshan, S. Belharbi, M. Pedersoli, A.L. Koerich, S. Bacon, and E. Granger. Disentangled source-free personalization for facial expression recognition with neutral target data. In *International Conference on Automatic Face and Gesture Recognition (FG)*, 2025b.

N.A. Sharma, A.S. Ali, and M.A. Kabir. A review of sentiment analysis: tasks, applications, and deep learning techniques. *International journal of data science and analytics*, 19(3):351–388, 2025.

M. H. Siddiqi, I. Ahmad, Y. Alhwaiti, and F. Khan. Facial expression recognition for healthcare monitoring systems using neural random forest. *IEEE Journal of Biomedical and Health Informatics*, 2024.

S. Sinha, S.Kr. Mishra, and S. Bilgaiyan. Emotion analysis to provide counseling to students fighting from depression and anxiety by using cctv surveillance. In *Machine Learning and Information Processing: Proceedings of ICMLIP 2019*, pp. 81–94, 2020.

B. Subramanian, J. Kim, M. Maray, and A. Paul. Digital twin model: A real-time emotion recognition system for personalized healthcare. *IEEE Access*, 10:81155–81165, 2022.

A. Suraj, A. S Kaushik, and K. Bai. Patient monitoring using emotion recognition. *International Journal for Research in Applied Science and Engineering Technology*, 10(8):1197–1201, 2022.

I. Sutskever, J. Martens, G. Dahl, and G. Hinton. On the importance of initialization and momentum in deep learning. In *ICML*, 2013.

A.D. Vairamani. *Advancements in Multimodal Emotion Recognition: Integrating Facial Expressions and Physiological Signals*, pp. 217–240. Cham, 2024.

M.J. Van Gent, M.C. Onwezen, R.J. Renes, and M. Handgraaf. Betwixt and between: A systematic review on the role of ambivalence in environmental behaviours. *Journal of Environmental Psychology*, 97:102311, 2024.

B. Voisard, A.I. Dragomir, V.G. Boucher, G. Szczepanik, Simon L. Bacon, and K.L. Lavoie. Training physicians in motivational communication: An integrated knowledge transfer study protocol. *Health Psychology*, 2024.

P. Waligora, H. Aslam, O. Zeeshan, S. Belharbi, A. L. Koerich, M. Pedersoli, S. Bacon, and E. Granger. Joint multimodal transformer for emotion recognition in the wild. In *ABAW workshop at CVPR*, 2024.

S. Walter, S. Gruss, H. Ehleiter, J. Tan, H.C. Traue, P. Werner, A. Al-Hamadi, S. Crawcour, A.O. Andrade, and G. Moreira da Silva. The biovid heat pain database data for the advancement and systematic validation of an automated pain recognition system. In *International Conference on Cybernetics (CYBCO)*, pp. 128–131, 2013.

J. Wang and L. Kawka. Gimefive: Towards interpretable facial emotion classification. *CoRR*, abs/2402.15662, 2024.

Y. Wang, S. Yan, Y. Liu, W. Song, J. Liu, Y. Chang, X. Mai, X. Hu, W. Zhang, and Z. Gan. A survey on facial expression recognition of static and dynamic emotions. *CoRR*, abs/2408.15777, 2024.

B. Williams. *Practical human behaviour change for the health and welfare of animals*. 2024.

Z. Wu and J. Cui. La-net: Landmark-aware learning for reliable facial expression recognition under label noise. In *ICCV*, 2023.

F. Xue, Q. Wang, and G. Guo. Transfer: Learning relation-aware facial expression representations with transformers. In *ICCV*, 2021.

F. Xue, Q. Wang, Z. Tan, Z. Ma, and G. Guo. Vision transformer with attentive pooling for robust facial expression recognition. *IEEE Transactions on Affective Computing*, 2022.

J. Yan, P. Li, C. Du, K. Zhu, X. Zhou, Y. Liu, and J. Wei. Multimodal emotion recognition based on facial expressions, speech, and body gestures. *Electronics (2079-9292)*, 13(18), 2024.

S. Yang, J. Van de Weijer, L. Herranz, and S. Jui. Exploiting the intrinsic neighborhood structure for source-free domain adaptation. *NeurIPS*, 2021.

M.O. Zeeshan, M.H. Aslam, S. Belharbi, A.L. Koerich, M. Pedersoli, S. Bacon, and E. Granger. Subject-based domain adaptation for facial expression recognition. In *International Conference on Automatic Face and Gesture Recognition (FG)*, 2024.

Muhammad Osama Zeeshan, Marco Pedersoli, Alessandro Lameiras Koerich, and Eric Grange. Progressive multi-source domain adaptation for personalized facial expression recognition. *CoRR*, 2025.

D. Zeng, Z. Lin, X. Yan, Y. Liu, F. Wang, and B. Tang. Face2exp: Combating data biases for facial expression recognition. In *CVPR*, 2022.

C. Zhang, H. Liu, Y. Deng, B. Xie, and Y. Li. Tokenhpe: Learning orientation tokens for efficient head pose estimation via transformers. In *CVPR*, 2023a.

S. Zhang, Z. Zhao, and C. Guan. Multimodal continuous emotion recognition: A technical report for abaw5. In *ABAW workshop at CVPR*, 2023b.

Y. Zhang, R. Zhao, W. Dong, B.-G. Hu, and Q. Ji. Bilateral ordinal relevance multi-instance regression for facial action unit intensity estimation. In *CVPR*, 2018.

S. Zhao, G. Jia, J. Yang, G. Ding, and K. Keutzer. Emotion recognition from multiple modalities: Fundamentals and methodologies. *IEEE Signal Processing Magazine*, 38(6):59–73, 2021.

C. Zheng, M. Mendieta, and C. Chen. POSTER: A pyramid cross-fusion transformer network for facial expression recognition. In *ICCVw*, 2023.

Appendix

Table of Contents

A To Reviewers: Please Read Before Downloading BAH Dataset

You can directly download BAH for review purposes only: Via our newly installed private and anonymous server at our university through this link `https://142.137.245.13/index.php/s/MyY2GyzBwjNXFLq`. For the first time usage, please accept the security certificate on your internet browser before proceeding – we are working to fix that. If this fails, please use the second option.

Please use this password to unzip the file "BAH_DB-shared-public-no-comp-ICLR2026.zip" (and "data.zip"): *@ICRL_oSY5QhGTHH5ckAf3qKCF_2026_Brazil*
For access issues, please reach out to the ICLR organizers so we can help you.

To ensure that we are compliant with our ethical requirements we are asking reviewers to read the "0_Read_before_downloading_data.pdf" (which is in the root of the files pointed by the download link of BAH dataset) before downloading the dataset. If you agree to the terms, please proceed to the dataset download. Here is the content of the file "0_Read_before_downloading_data.pdf":

" `To be consistent with the ethical requirements of the BAH dataset, by downloading the BAH dataset, you are agreeing to the following terms:`
`o Purpose of the access: You are accessing the BAH dataset as a part of a blind review process for the ICLR 2026 purposes.`
`o Use of the data: You will not redistribute, republish, or disseminate the BAH dataset.`
`o Duration of the access: The dataset must be securely destroyed after the review process has been completed.`
`If you have any concerns regarding these terms, please contact the ICLR 2026 organisers.` "

The download link provides access to the dataset itself for review purposes. Additionally, we included other materials such as anonymized EULA, and request forms. We also include a presentation to our data collection platform www.aerstudy.ca.

B Related Work

This section provides works in affective computing related to behavioural science.

a) Affect Recognition using Machine Learning:

Basic Emotions. An important line for ML research in affective computing is discrete emotion recognition in facial image (facial Expression Recognition – FER) (Bonnard et al., 2022; Liu et al., 2024a; Kollias et al., 2025; Lee et al., 2023; Mao et al., 2024; Wang et al., 2024; Wu & Cui, 2023; Xue et al., 2021; Zeng et al., 2022; Zheng et al., 2023). This usually involves classifying facial images into one of seven or eight basic emotions, such as ‘Happy’, ‘Sad’, and ‘Surprised’. Other works focus on videos (Liu et al., 2023a; 2021a;b; 2023b) as well. There has also a recent interest in designing robust FER methods that are interpretable (Belharbi et al., 2024a;b; Wang & Kawka, 2024; Xue et al., 2022). They typically produce a heat map that points to relevant regions used by a model to perform a prediction. This is usually formulated as an attention map or a Class-Activation Map (CAM) (Choe et al., 2022; Murtaza et al., 2025). Other work aims to predict ordinal levels (i.e, ordered labels), including pain and stress estimation (Aslam et al., 2024; Chaptoukaev et al., 2023; Zeeshan et al., 2024; Nasimzada et al., 2024); a task that can be extremely useful in healthcare applications. Some datasets such as `BioVid` (Walter et al., 2013) rely on advanced and expensive modalities such as bio-signals to predict pain for instance. Dimension recognition of emotions, typically aims to estimate continuous valence and arousal values linked to emotions (Dong et al., 2024; Praveen & Alam, 2024a; Praveen et al., 2023; 2021). Finally, another task related in emotion recognition is Action Units (AUs) detection (Jacob & Stenger, 2021; Luo et al., 2022). It aims to predicting active AUs in the face under a multi-label classification framework. Other works go further to estimate the intensity of AUs (Fan et al., 2020; Zhang et al., 2018), or both (Sánchez-Lozano et al., 2018), a much more challenging task.

Compound Emotions. Real-world scenarios often present complex emotions that combine basic ones. There has been recent interest in building affective computing models to predict compound emotions, a case where a mixture of basic emotions are expressed (Kollias, 2023; Richet et al., 2024). These are show in several practical real-world application since such complex emotions occur in daily interactions. However, they are more difficult to recognize as they are subtle, ambiguous, and resemble basic emotions. A recent specialized video-based dataset named `C-EXPR-DB` (Kol-

Dataset	**Affect**	**Modalities**	**Subject-based**	**Num. of participants**	**Num. of samples**	**Environment**	**Annotation**
`RAF-DB` (Li et al., 2017)	Basic/compound emotions	Images	No	–	15,339 images	Wild	Image-level
`AffectNet` (Mollahosseini et al., 2019)	Basic emotions	Images	No	–	450k images	Wild	Image label
`Aff-wild2` (Kollias & Zafeiriou, 2019)	Basic emotions, Valence/Arousal, Action Units	Video, audio	No	–	564 videos	Wild	Frame-level
`MELD` (Poria et al., 2019)	Basic emotions	Video, audio	No	–	13000 utterances	Actors/TV-show	Frame-level
`C-EXPR-DB` (Kollias, 2023)	Compound emotions	Video, audio	No	–	400 videos	Wild	Frame-level
`UNBC-McMaster` (Kollias & Zafeiriou, 2019)	Pain estimation	Frames	Yes	25	200 videos	Lab	Frame-level
`BioVid` (Walter et al., 2013)	Pain estimation	Frames, biomedical signals (GSR, ECG, and EMG at trapezius muscle)	Yes	90	18017 samples	Lab	Frame-level
`RECOLA` (Ringeval et al., 2013)	Apparent Emotional Reaction Recognition	video, audio, physiology (electrocardiogram, and electrodermal activity)	Yes	46	46 videos	Lab	Frame-level
`SEWA` (Kossaifi et al., 2019)	Apparent Emotional Reaction Recognition	video, audio	Yes	398	1,990 videos	Wild	Frame-level
`WEMAC` (Miranda Calero et al., 2024)	Discrete, dimensional emotions	Physiology (blood volume pulse, galvanic skin response, and skin temperature), audio	Yes	100	100 records	Lab	Self-reported
`StressID` (Chaptoukaev et al., 2023)	Stress	EDA, ECG, Respiration, Face video, Speech	Yes	65	587 videos	Lab	Frame-level
`SchiNet` (Bishay et al., 2019)	Estimation of Symptoms of Schizophrenia	video	Yes	91	91 videos	Wild	Video-level
`MESC` (Chu et al., 2024)	Emotional Support Conversation	video, audio, text	Yes	–	1,019 dialogues	Wild	Utterance-level
`IEMOCAP` (Busso et al., 2008)	Improvisations of scripted scenarios for basic emotions	video, audio, text	Yes	10 actors	–	Lab/Actors	Frame-level
`BAH` (ours)	Ambivalence/Hesitancy	Video, audio, transcript	Yes	300	1,427 videos	Wild	Video-level, Frame-level, A/H cues

Table 7: Common affective computing datasets for emotion modelling in health contexts.

lias, 2023) has been constructed for the design/evaluation of models. The dataset accounts for the difficulty of the task as different modalities are required to better recognize compound emotions.

Despite the recent progress in affect modelling, Ambivalence/Hesitancy recognition is still unexplored in ML. A possible reason is the lack of specialized dataset for training and evaluation of ML models. As it is implicated in healthcare and interventions, A/H is a common topic in behavioural science (Hohman et al., 2016; Manuel & Moyers, 2016). A/H recognition is related to compound emotion recognition task where intention and attitudes are conflicted or in a in-between state, between willingness and resistance (MacDonald, 2015), or positive and negative affect (Armitage & Conner, 2000). This can manifest in how an individual expresses them self and can be recognized (Hayashi et al., 2023) in their facial expression, tone, verbal, and body language. As a result, A/H exhibits a multimodal nature that comes as the result of subtle interconnection between different cues. Unfortunately, such discord is extremely difficult to spot; a task that requires human training. This is a tedious and expensive procedure, leading to ineffective and less scalable eHealth interventions under limited resources. Assisting healthcare providers with automatic, reliable and inconspicuous tools to help them recognize A/H can have a major impact in improving eHealth interventions.

Our `BAH` dataset fills in the gap in the literature, and to provide an important resource to design/evaluate ML models for A/H recognition task. It is a video Q&A dataset from which we extract audiovisual information with transcripts, offering multiple modalities. The dataset is fully annotated by behaviour science experts at video- and frame-level. In addition, cues used by annotators to recognize A/H at each segment are provided. This includes facial and vocal expressions, body language, language in addition to highlighting where there is inconsistency between the modalities. As shown in Table 7, our `BAH` dataset is competitive compared to existing affective computing datasets in terms of modalities, number and diversity of participants, and annotations. While no dataset matches the specific focus on A/H in digital health interventions, datasets like `MESC` (Chu et al., 2024), `SchiNet` (Bishay et al., 2019), and `IEMOCAP` (Busso et al., 2008) contain videos from interviews with psychological relevance. Therefore, `BAH` provides an important asset for the ML community to begin research in A/H recognition.

b) Behavioural Science:

Health-Related Behaviour Change and Non-Communicable Diseases. High-risk health behaviours, such as tobacco use, physical inactivity, unhealthy diets, and harmful alcohol consumption, are responsible for the vast majority of non-communicable diseases (NCDs), which include cardiovascular disease, type 2 diabetes, cancer, and chronic respiratory illnesses. According to the World Health Organization (WHO) (Ortiz et al., 2025), NCDs account for approximately 74% of global deaths, and these outcomes are disproportionately influenced by modifiable behavioural factors. Evidence suggests that around 80% of chronic disease risk is attributable to these high-risk behaviours.

Consequently, health-related behaviour change has become a primary target for preventive and therapeutic interventions. Traditional methods, such as motivational interviewing (MI) and cognitive behavioural therapy (CBT), rely on face-to-face clinical interviews, which remain foundational to behavioural health practice (O'Donnell et al., 2019). These interactions provide unique opportunities for clinicians to detect ambivalence, hesitancy, and other complex affective states, often through subtle verbal and nonverbal cues (Hall et al., 1995). Despite the growing shift toward digital platforms, clinical interviews remain the gold standard for eliciting meaningful emotional and cognitive responses, insights that are essential to tailoring behaviour change strategies. Efforts to change health behaviours over the long term are inherently complex. Individuals often experience ambivalence and hesitancy, understood as fluctuating between intention and resistance, when attempting to adopt healthier lifestyles. In traditional healthcare contexts, providers rely on both verbal communication and non-verbal cues (e.g., tone, gestures, facial expressions) to recognize and address such motivational conflicts. This in-person interaction allows for nuanced support that can adapt to a patient's readiness for change (Davidson & Scholz, 2020). The purpose of developing multimodal A/H recognition systems is to capture and replicate this nuanced understanding of patient behaviour within digital health interventions, thereby supporting clinicians and scaling behavioural health care.

Multimodal Cues and the Detection of Complex Emotions. Identifying complex emotional states such as ambivalence, resistance, or hesitancy is crucial for tailoring behavioural interventions. Research in psychology and human-computer interaction has shown that complex emotional states, such as ambivalence, uncertainty, or defensiveness, are communicated through a combination of facial expressions, body posture, vocal tone, speech patterns, and physiological responses (Guo et al., 2018; Pantic & Rothkrantz, 2003). In digital contexts, however, the absence of physical presence makes this task more difficult. Recent research in psychology and computer science has focused on the use of multimodal cues, such as facial expressions, voice tone, body posture, and physiological responses, as proxies for emotional and motivational states (Kraack, 2024; Yan et al., 2024). These cues can reveal underlying emotional conflict or uncertainty that might not be captured by self-report alone. Studies have shown that combining multiple input channels (e.g., audio-visual data) can enhance the accuracy of emotion recognition systems. For instance, multimodal datasets are being used to train models that detect affective states like confusion, frustration, and mixed emotions, which are highly relevant in contexts such as education, mental health, and behaviour change. By incorporating these data streams, researchers can better approximate the nuanced human capacity for reading emotions, paving the way for emotionally aware systems (He et al., 2020; Zhao et al., 2021).

Affective Computing and Personalized Digital Health Interventions Affective computing, a subfield of artificial intelligence (AI) focused on recognizing, interpreting, and responding to human emotions, holds promise for advancing personalized digital health interventions. By leveraging emotion-aware algorithms, digital platforms can better understand users' psychological readiness and tailor support accordingly (Lokhande et al., 2024; Vairamani, 2024). For example, interventions that dynamically respond to detected signs of resistance or disengagement may improve user retention and behavioural outcomes. Incorporating affective computing into digital health technologies also allows dynamic tailoring of content based on users' real-time affect, responsive dialogue, mimicking the adaptability found in face-to-face interactions. Recent advancements in conversational agents, voice analysis, and facial expression recognition have made it possible for digital interventions to adapt content delivery based on real-time emotional assessments (Khanna et al., 2022). This not only improves user engagement but also enhances intervention effectiveness by ensuring messages are delivered in an emotionally congruent and contextually appropriate manner (Hornstein et al., 2023).

C BAH Dataset Limitations

While BAH dataset offers a novel contribution to emotion recognition for digital behaviour change, several limitations should be considered.

1) Data collection constraints. The web-based platform occasionally experienced technical issues, preventing some participants from completing all seven videos. As participants used their own

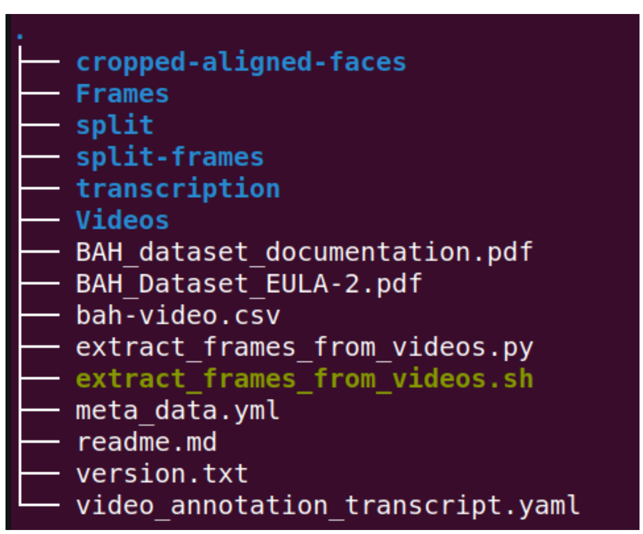


Figure 4: File structure of the shared BAH dataset.

devices in home settings, video and audio quality varied significantly despite clear instructions and testing. Response length was participant-determined, leading to high variability in content. Some environmental noise or visual distractions (e.g., background conversations, movement) were present in a subset of recordings.

2) Participant representation. Although participants were recruited from nine Canadian provinces with diverse age and ethnic backgrounds, individuals from under-resourced areas or without reliable internet access were likely underrepresented. Gender identity was collected but not used in sampling, and no data on socioeconomic status was recorded. Digital literacy and access may have biased participation toward more tech-savvy individuals.

3) Multimodal and data balance issues. The expressiveness of cues (facial, vocal, bodily, verbal) varied widely by participant, complicating consistent multimodal analysis. Though the dataset is balanced at the video level, frame-level imbalance exists (fewer A/H frames than non-A/H). Training strategies that account for class imbalance should be considered.

D BAH DATASET FILE STRUCTURE

Figure 4 shows the file structure of the shared BAH dataset. The file "BAH_dataset_documentation.pdf" contains the detailed documentation about all files/directory, including annotation structure.

E BAH DATA COLLECTION WEB-PLATFORM

Alongside the dataset files, we include a slide presentation of our "Automatic Expression Recognition" (AER) web-based platform (www.aerstudy.ca). The presentation is in the file "AER-web-platform.pdf". Figure 5 shows an example of the platform.

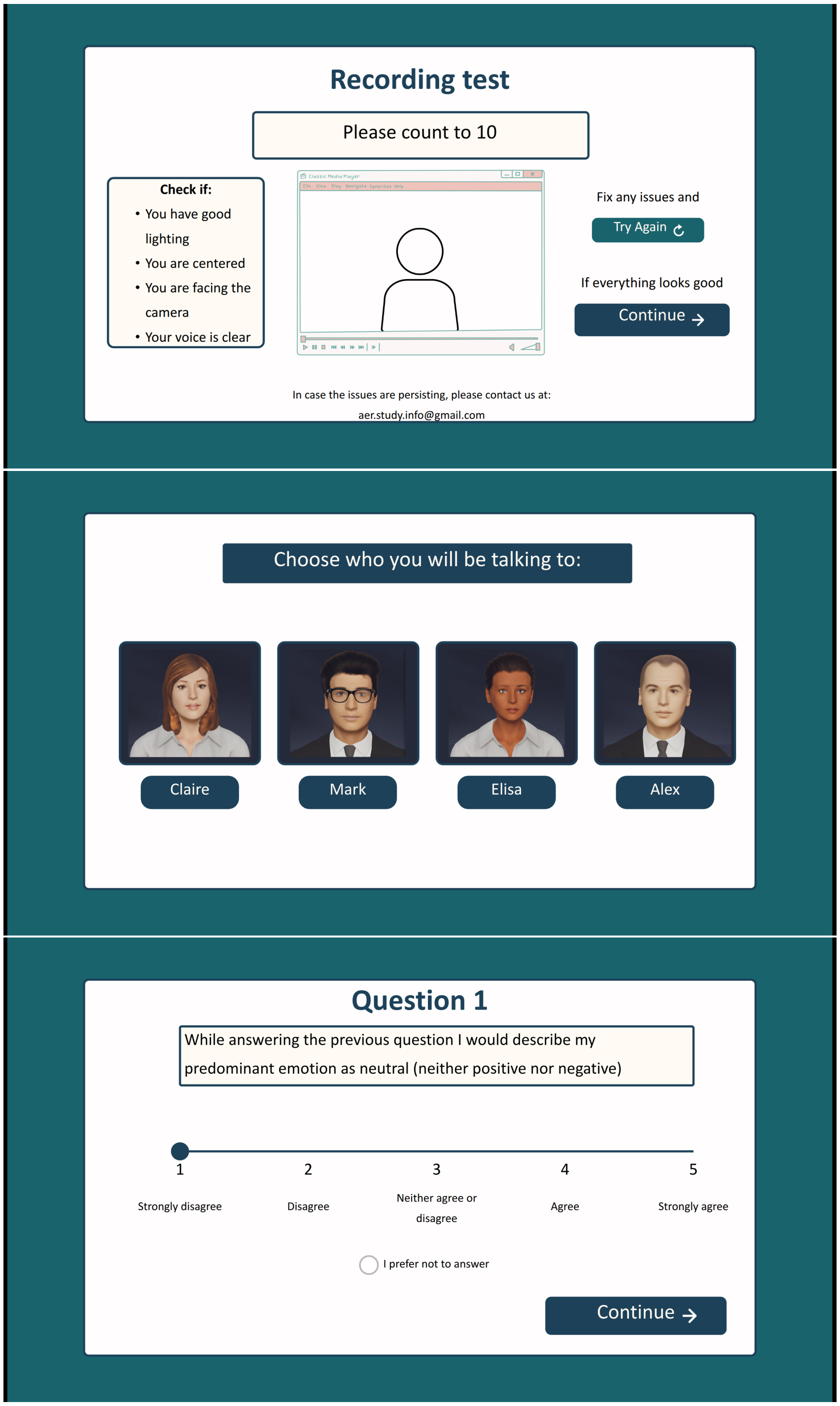


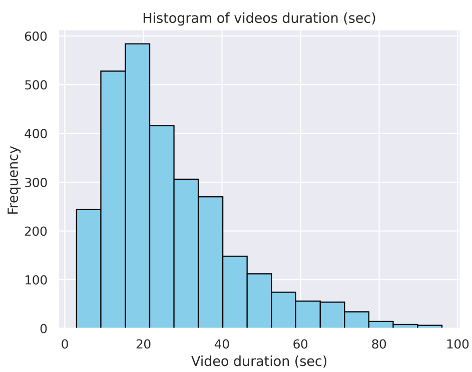
Figure 5: Examples taken from the platform to present our ”Automatic Expression Recognition” (AER) web-based platform (www.aerstudy.ca).

F BAH Dataset Diversity

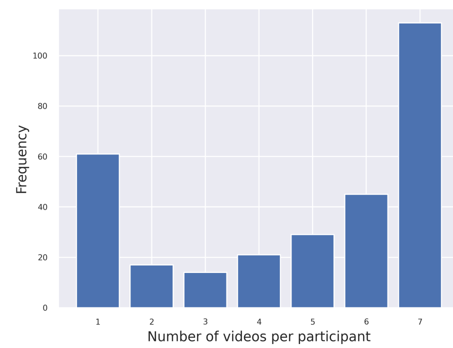
This section includes more statistics about BAH dataset to highlight its diversity. Figure 11 shows a general overview via a nutrition label. Overall, BAH dataset has significant diversity. It covers different Canadian provinces, age range, ethnicities, and male/female presence. It has a large number of videos (1,427) where 778 videos contain A/H. Most asked questions elicited A/H, especially question-4 (Ambivalent). In addition, since we have less control over the participants, and their environment, the dataset is considered in-the-wild. On top of video and audio modality, we provide audio transcript which has shown to be an important modality for A/H recognition. `BAH` is fully annotated at video- and frame-level. Moreover, annotators report the used cues to recognize A/H at each segment. All these properties make our dataset a realistic and relevant asset to design ML model for the task of A/H recognition in videos.

We include the following general information:

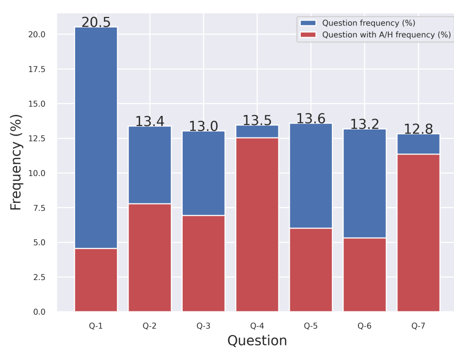
- Videos durations distribution (Figure 6a).
- Videos per participants distribution. (Figure 6b).
- Questions and A/H distribution (Figure 6c).
- A/H segments duration (Figure 7).
- Participants sex distribution (Figure 6d).



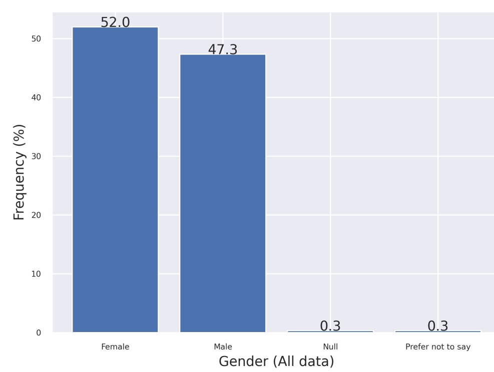
(a) Videos duration histogram.



(b) Distribution of number of videos per participant.

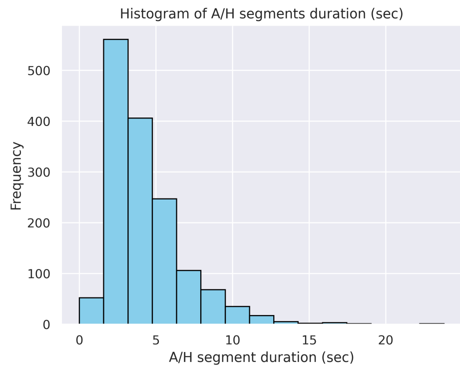


(c) Distribution over 7 questions: Num. videos per-question (blue), Num. videos with A/H (red).



(d) Sex distribution.

Figure 6: Video duration (a), and videos/participant (b), question distribution (c), and sex distribution (d) over `BAH` dataset.



(a) Distribution of A/H segment duration in seconds.

(b) Distribution of A/H segment duration in frames.

Figure 7: Distribution of A/H segment duration in seconds (a), and frames (b) over `BAH` dataset.

In addition, more demographics statistics are included as well:

- Participants' age distribution (Figure 8).
- Participants' age range distribution (Figure 9a).
- Distribution of Canada provinces where participants live (Figure 9b).
- Participants' simplified ethnicity distribution (Figure 10a).
- Participants' student-status distribution (Figure 10b).
- Participants' consent to use their data in challenges distribution (Figure 10c).
- Participants' consent to use their data in publications distribution (Figure 10d).
- Participants' birth country distribution (Table 8).

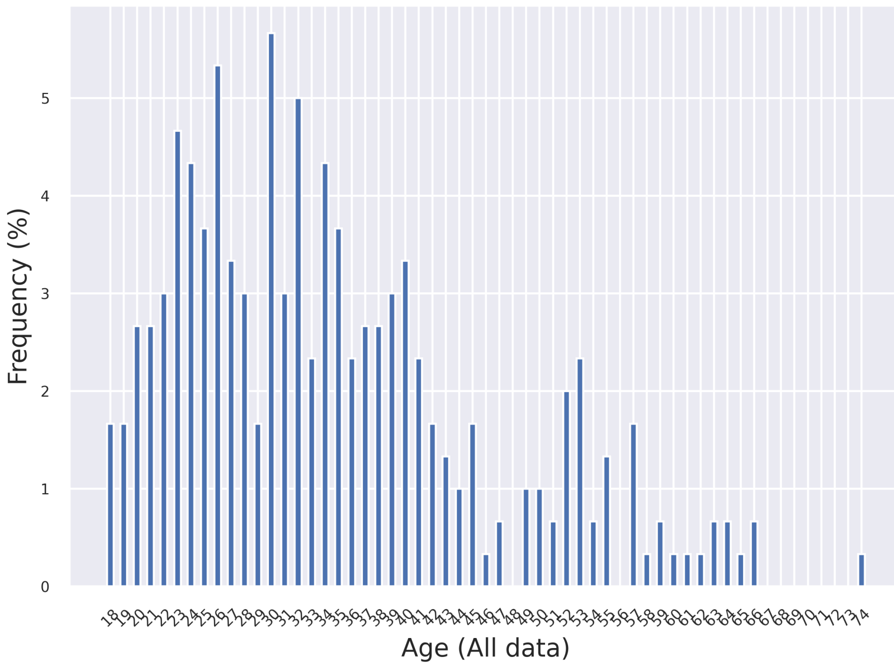
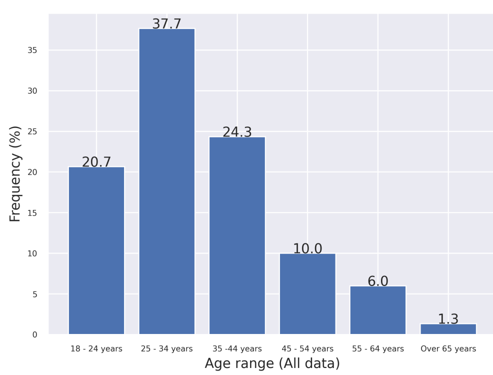
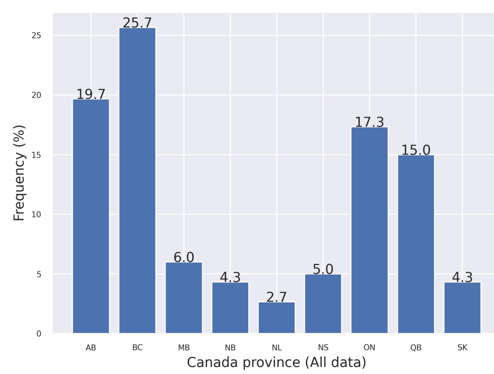


Figure 8: Participants' age distribution in BAH dataset.

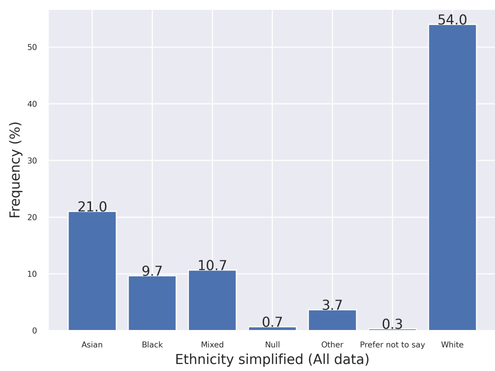


(a) Participants' age range distribution.

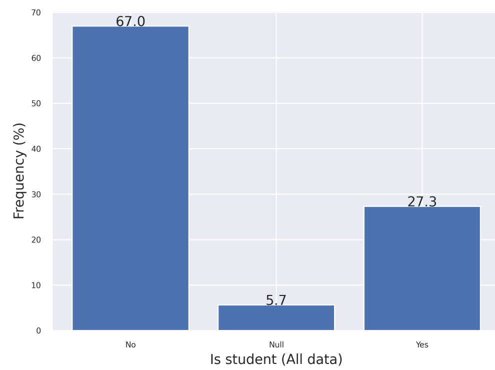


(b) Distribution of Canada provinces where participants live.

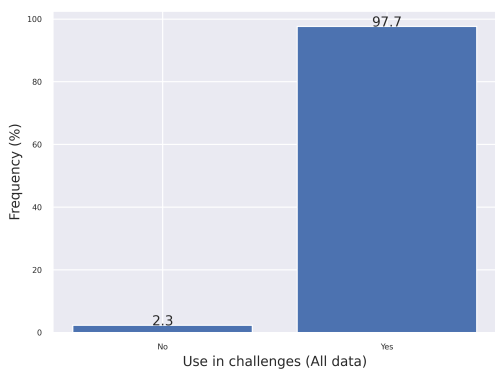
Figure 9: Participants' age range (a), and where the provinces where they live (b) over BAH dataset. Name of provinces: 'Manitoba (MB)', 'Alberta (AB)', 'Nova Scotia (NS)', 'Newfoundland and Labrador (NL)', 'Saskatchewan (SK)', 'New Brunswick (NB)', 'Ontario (ON)', 'Quebec (QC)', 'British Columbia (BC)'.



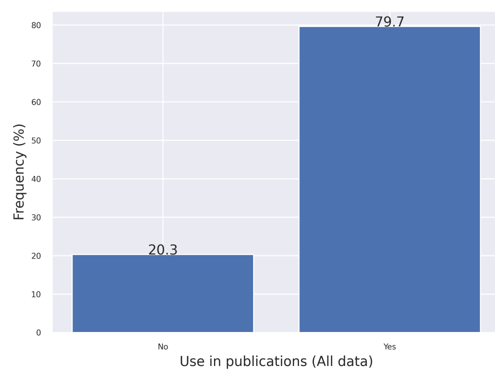
(a) Participants' simplified ethnicity distribution in `BAH` dataset.



(b) Participants' student-status distribution in `BAH` dataset.



(c) Participants' consent to use their data in challenges distribution in `BAH` dataset.



(d) Participants' consent to use their data in publications distribution in `BAH` dataset.

Figure 10: Distribution of participants' simplified ethnicity (a), their student-status (b), their consent to use their data in challenges (c), and publications (d) over `BAH` dataset.

Birth country	Number of participants
Algeria	1
Australia	1
Bangladesh	2
Belize	1
Bulgaria	1
Canada	**127**
China	**8**
Colombia	1
France	1
Germany	3
Ghana	1
Hong Kong	1
India	5
Japan	1
Kenya	1
Macedonia	1
New Zealand	1
Nigeria	**9**
Null	1
Peru	1
Philippines	6
Russian Federation	1
Saint Lucia	1
South Africa	1
Sri Lanka	2
Taiwan	1
Thailand	1
Tunisia	1
Turkey	2
United Arab Emirates	1
United Kingdom	6
United States	3
Vietnam	1

Table 8: Distribution of participants' birth country in BAH dataset. Top-3 countires are in bold.

BAH Dataset Facts

Dataset BAH (Behavioural Ambivalence/Hesitancy – A/H)
Nature of Dataset A Dataset for Ambivalence/Hesitancy recognition in videos for participants recruited in Canada
Participants Country Canada
Number of provinces in Canada 9
Provinces in Canada 'Manitoba (MB)', 'Alberta (AB)', 'Nova Scotia (NS)', 'Newfoundland and Labrador (NL)', 'Saskatchewan (SK)', 'New Brunswick (NB)', 'Ontario (ON)', 'Quebec (QC)', 'British Columbia (BC)'.
Number of participants 224
Number of videos 1,118 where 638 videos contains A/H
Video length 26.58 ± 16.36 (seconds) with a minimum and maximum duration of 3 and 96 seconds
Total duration 8.26 hours where A/H duration is 1.5 hours
Total number of frames 714,005 where 131,103 frames contains A/H
Total number of A/H video segments 1,274
Length A/H video segment 4.25 ± 2.47 seconds or 102.92 ± 59.16 frames. The minimum and maximum A/H segment is 0.01 seconds (1 frame), and 23.8 seconds (572 frames)
Data capture web-platform www.aerstudy.ca

Motivation

Summary Behavioural Ambivalence/Hesitancy (BAH) is a dataset collected for subject-based multimodal recognition of A/H in videos. It contains videos from 224 participants captured across 9 provinces in Canada, with different age, and ethnicity. Through our web platform, we recruited participants to answer 7 questions, some of which were designed to elicit A/H while recording themselves via webcam with microphone. BAH amounts to 1,118 videos for a total duration of 8.26 hours with 1.5 hours of A/H. Our behavioural team annotated timestamp segments to indicate where A/H occurs, and provide frame- and video-level annotations with the A/H cues. Video transcripts and their timestamps are also included, along with cropped and aligned faces in each frame, and a variety of participants meta-data.
Original Authors Redacted for anonymity reasons.

Metadata

URL Redacted for anonymity reasons.
Keywords Ambivalence/hesitancy, eHealth, digital health intervention, video, Deep Learning, Benchmark
Available participants meta-data Age, birth country, Canada province where the participant lives, ethnicity, ethnicity simplified, sex, student status, consent to use recordings in publications
Video format *.mp4
Ethical Review Redacted for anonymity reasons.
License Custom - for research purposes only.
How to request the data? Fill in this form, sign, and upload the EULA - Redacted for anonymity reasons.
First release 2025

Annotation

Annotators 3 experts in behavioural science
Video- and frame-level Label “1” for presence of A/H, “0”, its absence
Cues provided by annotators for each A/H segment Facial expressions, body language, audio and language in addition to highlighting where there is inconsistency between the modalities

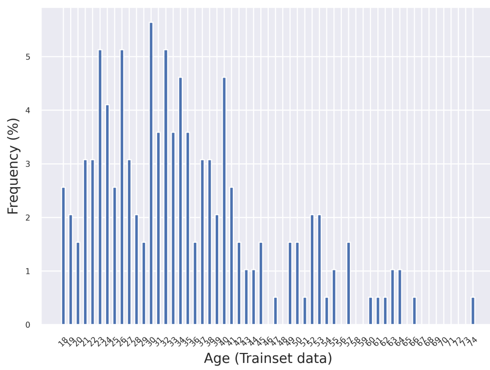
Data size

All files zipped (*.zip) 8.3 GB

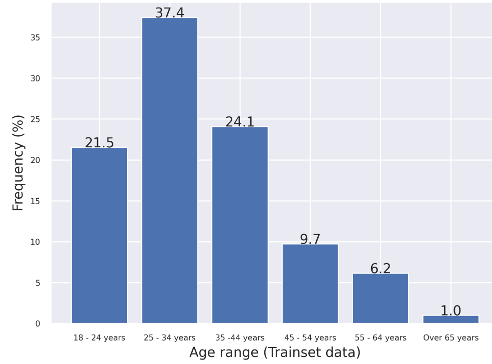
Figure 11: A data card styled (nutrition label) for BAH dataset.

G DEMOGRAPHICS ANALYSIS OF TRAIN, VALIDATION, AND TEST SETS

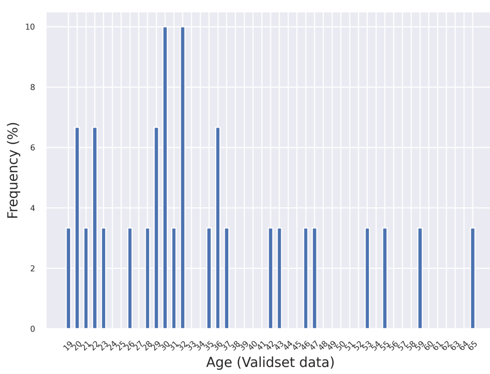
We provide in Figure 12 and 13 relevant distributions of the splits train (195 participants), validation (30 participants) and test (75 participants) sets in BAH dataset. The split participants-based where videos of participant belong exclusively to one split. We ensure that all splits have similar distribution in terms of 3 main factors that are deemed important in A/H variation across the population: gender, age, and ethnicity. We also include provinces distribution in Figure 14.



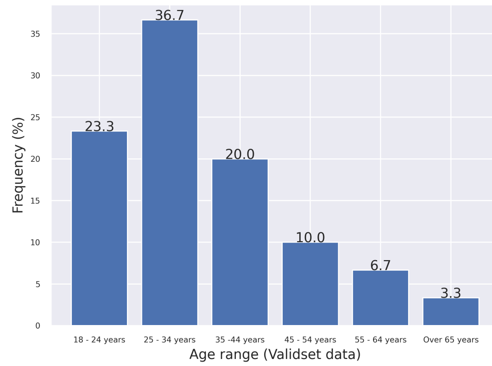
(a) Train set split participants age distribution.



(b) Train set split participants age-range distribution.

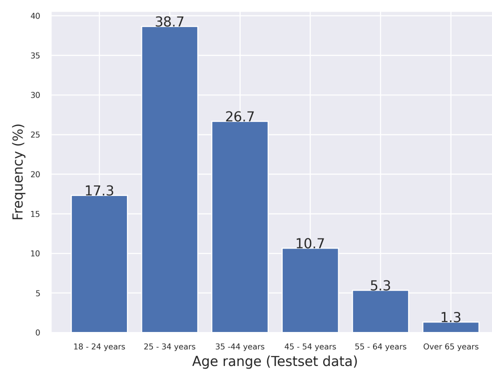


(c) Validation set split participants age distribution.



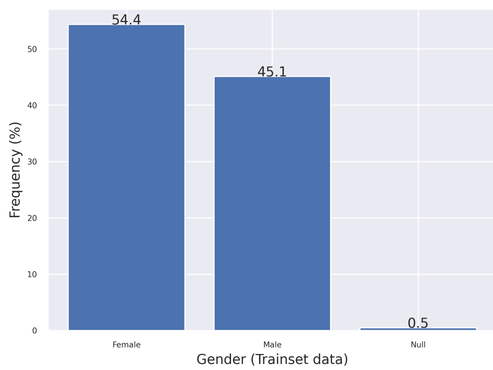
(d) Validation set split participants age-range distribution.

(e) Test set split participants age distribution.

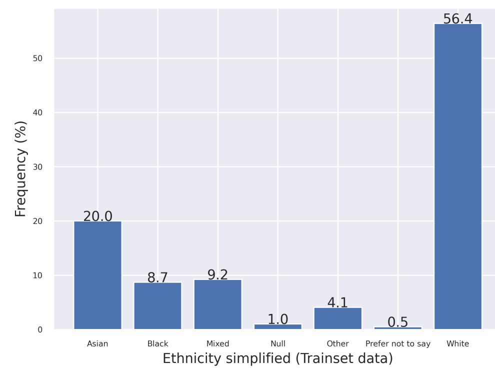


(f) test set split participants age-range distribution.

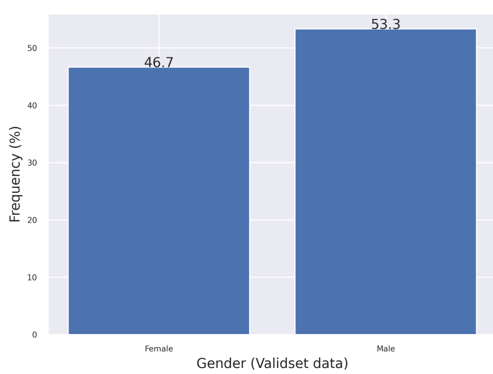
Figure 12: Participants age and age-range distribution across all splits (train, validation, and test) in BAH dataset.



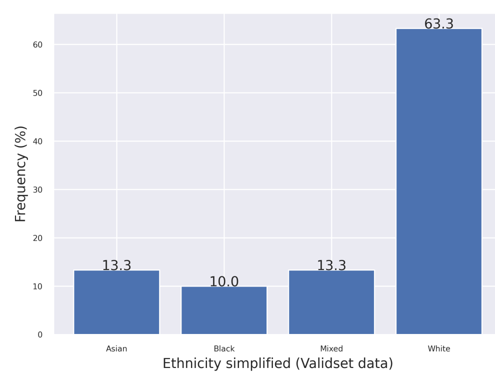
(a) Train set split participants sex distribution.



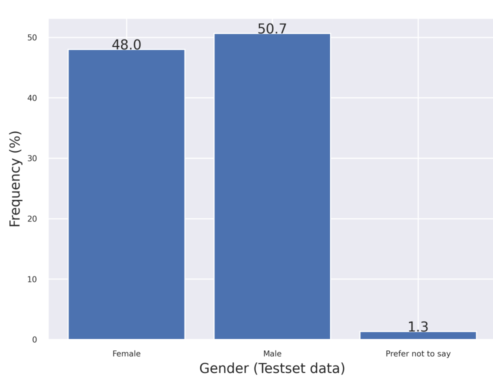
(b) Train set split participants simplified ethnicity distribution.



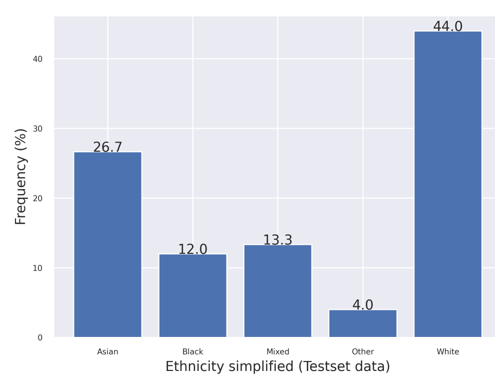
(c) Validation set split participants sex distribution.



(d) Validation set split participants simplified ethnicity distribution.

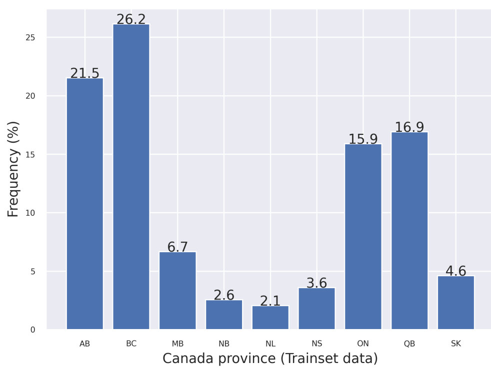


(e) Test set split participants sex distribution.

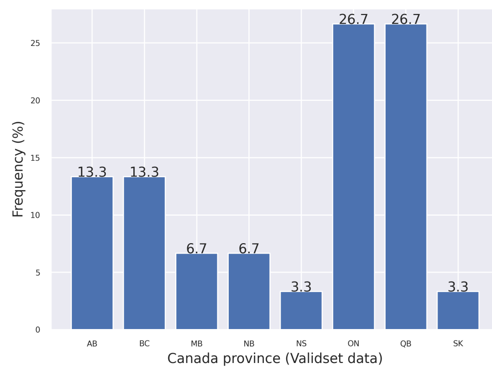


(f) test set split participants simplified ethnicity distribution.

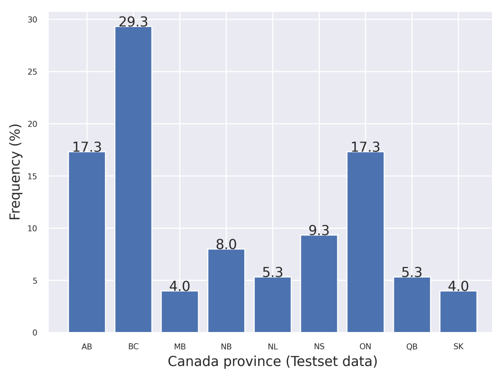
Figure 13: Participants sex and simplified ethnicity distribution across all splits (train, validation, and test) in BAH dataset.



(a) Train set split participants provinces distribution.



(b) Valid set split participants provinces distribution.



(c) Test set split participants provinces distribution.

Figure 14: Participants provinces distribution across all splits (train, validation, and test) in BAH dataset.

H ANALYSIS OF ANNOTATORS CUES

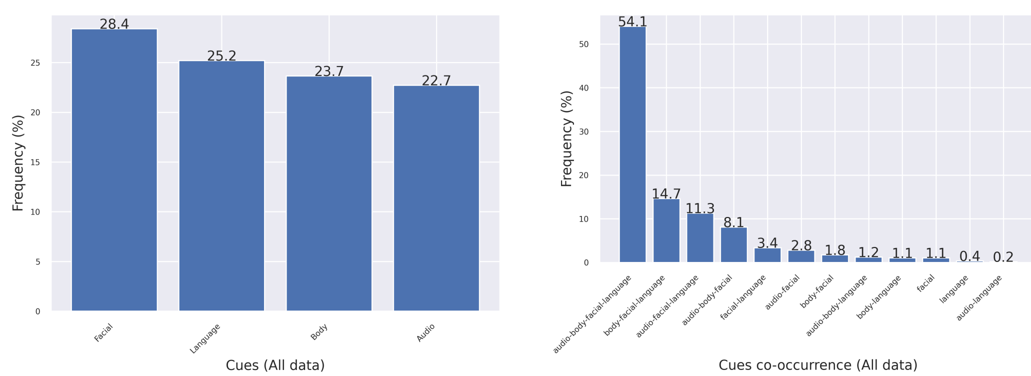
Figure 15a shows the distribution of modalities used by annotators to assess the presence of A/H. Overall, the four modalities contribution almost at the same rate in detecting A/H. Facial cues lead, followed by language, then body and audio. This suggests that the four modalities are equally important for A/H recognition. When looking to co-occurrence of these cues, we find that the four cues at once dominates by 54.1%, followed by "body-facial-language" with 14.7%, and "audio-facial-language". This suggest an adapted fusion of these four modalities is necessary to be able to assess there is an affect between them.

We include in Figure 16, 17 the distribution of cues used per modality. It shows that "Pause" is the dominate cues used in audio modality; "Filler sound" for language; "Gaze" for facial; and "Shake" for body.

We also analyzed the inconsistencies between modalities in Figure 18a, 18b. The inconsistencies between facial and language cues seem to be the lead by 42.6% of cases followed by language and body with 26.6%. This could be used as priors to recognize the case of A/H cross-modalities. When looking to co-occurrence, similarly, facial and language dominates alone with 21.4% of cases, followed by facial-language and language-body with the same rate. Leveraging these statistics by using them as constraints could help designing better A/H frameworks.

A/H can also be present within a single modality alone making it detection much more difficult compared to the case of cross-modality. Figure 19a shows that a large part of A/H cases fall into the case of within-modality, which is also spread across all questions (Figure 19b). A dedicated per-modality temporal modelling could help detect these cases.

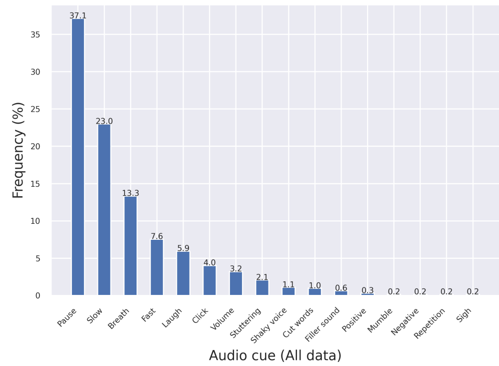
These aforementioned statistics show that recognizing A/H requires considering all four modalities, and being able to detect affect conflict across pairs and higher combinations of modalities for the case of cross-modality. Additionally, a better temporal modelling is required to detect A/H in the case of within modality.



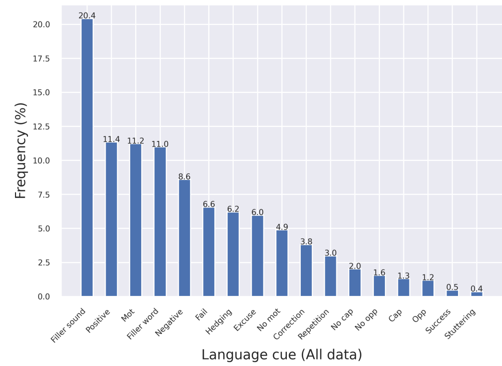
(a) Modalities cues distribution.

(b) Modalities cues co-occurrence distribution.

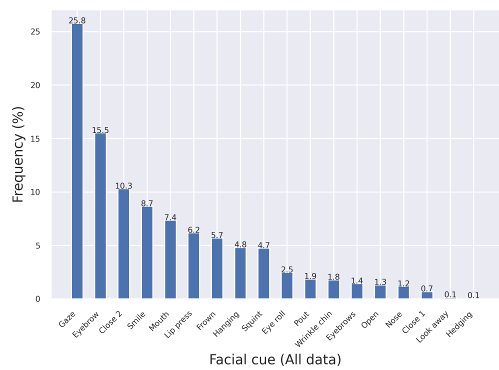
Figure 15: Modalities distribution used by annotators in BAH dataset.



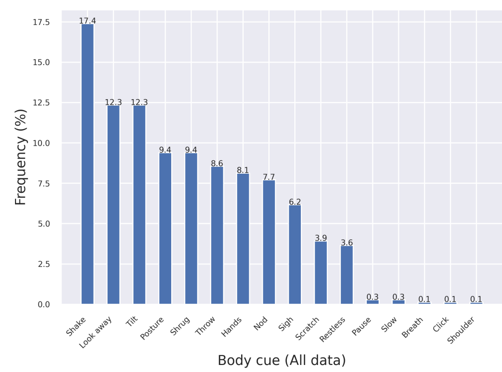
(a) Audio modality cues distribution.



(b) Language modality cues distribution.

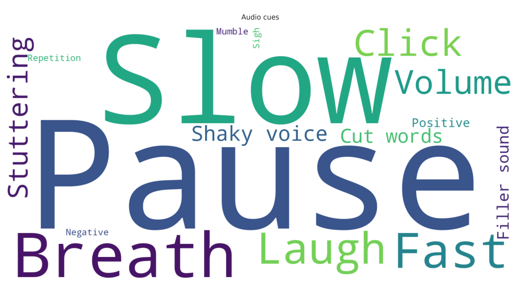


(c) Facial modality cues distribution.

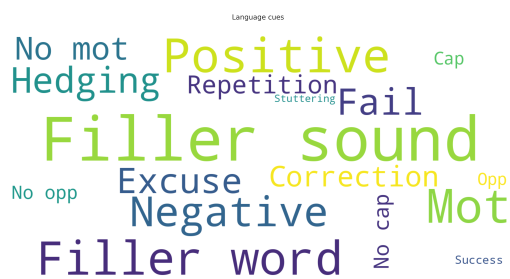


(d) Body modality cues distribution.

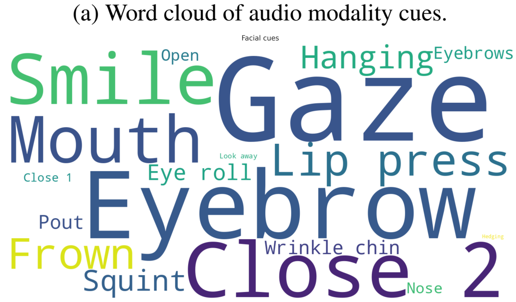
Figure 16: Per-modality cues distribution used by annotators in BAH dataset.



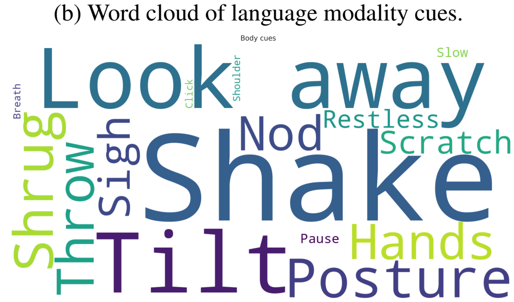
(a) Word cloud of audio modality cues.



(b) Word cloud of language modality cues.

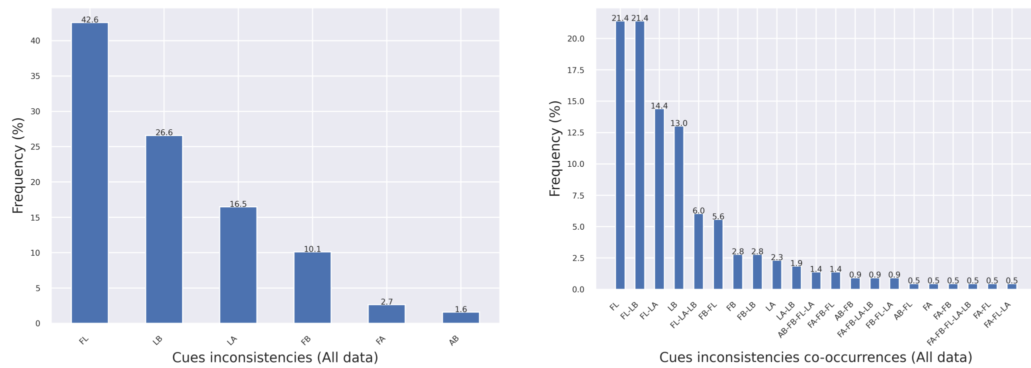


(c) Word cloud of facial modality cues.



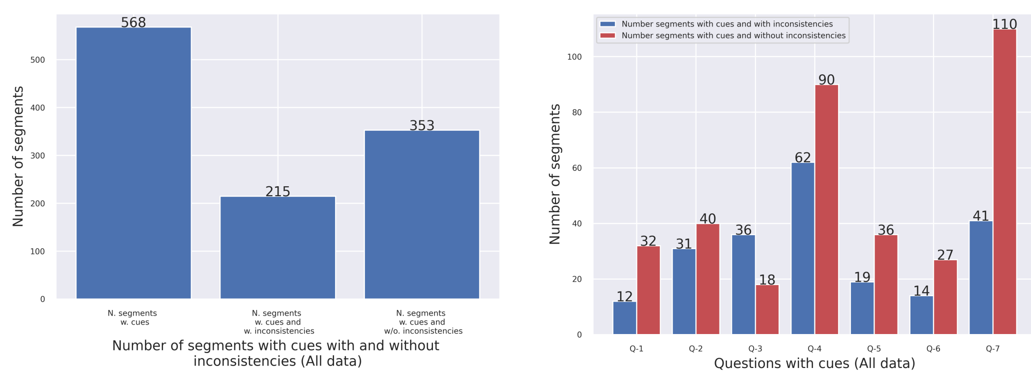
(d) Word cloud of body modality cues.

Figure 17: Word cloud of per-modality cues used by annotators in BAH dataset.



(a) Cues inconsistencies distribution. (b) Cues inconsistencies co-occurrence distribution.

Figure 18: Cues inconsistencies and their co-occurrence distribution used by annotators in BAH dataset.



(a) Cues distribution with and without cross-modality inconsistencies over segments. (b) Cues distribution with and without cross-modality inconsistencies over questions.

Figure 19: Distribution of cues with and without cross-modality inconsistencies over segments; in addition to question-based distribution used by annotators in BAH dataset.

I BAH DATASET ANNOTATION CODEBOOK

This section contains relevant information regarding our designed annotation codebook for A/H recognition. We provide the definitions of A/H and the types of cues (Table 9), as well as a more detailed description of the most relevant cues in each modality used to detect A/H, which include facial cues (Table 10), language cues (Table11), audio cues (Table 12), body language cues (Table 13), and cross-modal inconsistency cues (Table 14). The codebook is a working document that continues to evolve in response to relevant insights emerging from expert annotations and contributions from behaviour change experts. Updates on the codebook will be made available and communicated upon request. This iterative approach aligns with established qualitative research practices, where coding frameworks are refined throughout the analysis process to better reflect the complexity and richness of the data (Bradley et al., 2007).

Term	Definition
Ambivalence/Hesitancy	The simultaneous presence of competing positive and negative feelings, ideas, thoughts, or emotions towards one same object or goal. A state in which a person has not entirely made up their mind about doing something; when they aren't fully decided on how to act (towards a behaviour or object; not necessarily the goal behaviour; excluding towards language or answering questions)
Facial Cues	Different motions of the muscles in the face. Facial expressions commonly occur around the mouth and eyes, including changes in a person's gaze. They can be used to assess a person's emotional state.
Language Cues	Includes verbal/speech-based expressions of ambivalence or hesitancy. Some common verbal expressions can include the use of 'I want to… but…', 'mmmm', among others.
Audio Cues	Changes in a person's non-verbal language, such as changes in tone, speed and pitch.
Body Cues	Non-verbal signals that include gestures, body posture and movements. Some of the cues that can be annotated as body language are hand movements, head tilts, shoulders shrugging and sighs (chest movement).
Cross-modal inconsistency Cues	Simultaneous incompatibility between two or more modalities or different types of cues. For example, this could be represented by someone saying 'yes' while shaking their head side to side.

Table 9: `BAH` dataset annotation codebook: definitions.

Facial cues	Definition
Upper Region	
Close 2	A change in the frequency with which someone blinks or closing one's eyes for a longer period (e.g., either keeping them closed, or blinking for a long time). This excludes normal blinking, it is annotated when there is a difference compared to the participants baseline. Closing of both eyes; "blinking" (with both eyes).
Close 1	Closing one eye at the time; includes winking. The duration of the wink is not relevant, it can be a quick wink or a longer one.
Squint	Partially closing one or both eyes. Significant or identifiable changes or contractions in the muscles around the outer or inner corners of the eyes. It might involve some changes in the eyebrows, forehead and cheeks. Includes squinting eyes, muscles contracting around the eyes.
Frown	To bring your eyebrows together (inner eyebrow) so that there are lines on your face above your eyes. Frown, forehead fold, small frown, tensed forehead, wrinkled forehead, furrowing brows, lowering inside corners of eyebrow
Eyebrow	Lowering/raising external parts of eyebrow(s) (or full eyebrow(s)). [i.e., one or both eyebrows]
Gaze	Changes in the direction of the gaze by moving the eyes. Moving eyes (not face) to look down, up, to the side.
Eye roll	Eye-rolling is a transitory gesture in which a person briefly turns their eyes upward, often in an arcing motion from one side to the other. The eyes do not set on anything in particular and go back to their previous position.
Open	Opening the eyes, looks like an increase in awareness. Eyes look slightly bigger. Engagement of the eyelids, contracting the eyelid muscles to make them look wider.
Lower Region	
Smile	Ends of the mouth/lips curve up, often with the lips moving apart. Includes: Smile, smirking, half a smirk, fake smile, raising both sides of the mouth, side smile, half smile
Pout	Pushing one's lips or one's bottom lip forward; or turning the outer sides of the lips downwards. Pouting, pursed lips, "frowning" with one's mouth
Lip press	Contracting or pressing lips without pushing them forward. Includes: pressed lips, pressing lips together, putting lips together. Excludes pressing lips to pout/purse.
Hanging	Leaving one's mouth open for an extended time (e.g., hanging mouth, gaping mouth).
Mouth	Any other movements of the mouth that (1) are not captured by smile, pout, or pressing lips, and (2) is not a result of the baseline speech patterns. Opening mouth, opened mouth, raising upper lip, rising one side of the mouth, taking the mouth corners back and lower them
Wrinkle Chin	Moving the muscles around the chin to create identifiable lines, folds, ridges or furrows in the chin. Usually seen as a contraction of the chin muscles creating creases around or on the chin.
Nose	Changes in the movement or looks of the nose. Includes significant movements on the nostrils, the tip of the nose, scrunching the nose, or any other muscle movement that would create a change in the nose.

Table 10: BAH dataset annotation codebook: facial cues.

Language cues	**Definition**
Filler sound	Sound made during a pause in speech signalling the person isn't done taking. Examples: "mmm", "umm", "hum", "emmm", "err", "uh", "ah"
Filler word	Words used that do not contain substantive content, but are used as fillers to fill in space while the person thinks (or to signal they are not done talking, or that they are about to talk): "like", "you know", "I mean", "okay", "so", "actually", "basically"
Hedging	Words/expressions used to express ambiguity about what one is saying (about to say or just said). Examples: "somewhat"; "I'm not an expert, but..."; "… right?"; "… isn't it?"; "I do not know…"; "all I know"; "I think…"
Correction	Corrects something they said. This focuses on the content of what is said, not on a syntax-based error, or speech error.
Repetition	Emphasizing a phrase by repeating, or repetition of a word, might be related to trying to find the right word or expression
Com-B Constructs	
Positive	Statement of positive feelings towards a behaviour or action.
Negative	Statement of negative feelings towards a behaviour or action.
Excuse	Statement where the participant mentions an excuse, a reason or justification for something that has happened or hasn't happened. It can also be an expression of regret for doing/not doing something. Use of 'but'. Shows avoidance or lack of responsibility
Success	Statement of success with goal (focused on the behaviour)
Fail	Statement of failure with goal (focused on the behaviour)
Cap	Mentions having the capability to change their behaviour. Includes physical capability (e.g., balance, dexterity) or psychological capability (e.g., knowledge, skills, memory).
No cap	Mentions NOT having the capability to change their behaviour. Includes physical capability (e.g., balance, dexterity) or psychological capability (e.g., knowledge, skills, memory).
Mot	Mentions having motivation to change their behaviour. Includes reflective motivation (e.g., making plans, having positive attitudes/beliefs) or automatic motivation (e.g., desires, habit, feelings)
No mot	Mentions NOT having motivation to change their behaviour OR motivation NOT to change their behaviour. Includes reflective motivation (e.g., making plans, having positive attitudes/beliefs) or automatic motivation (e.g., desires, habit, feelings).
Opp	Mentions having opportunity to change their behaviour. Includes physical opportunities (e.g., access to financial resources, location, time) or social opportunity (e.g., support/encouragement from others; norms)
No opp	Mentions NOT having opportunity to change their behaviour. Includes physical opportunities (e.g., access to financial resources, location, time) or social opportunity (e.g., support/encouragement from others; norms)

Table 11: BAH dataset annotation codebook: language cues.

Audio cues	Definition
Pause	Briefly interrupting a sentence by having silent pauses in between words or ideas that differ from the usual pace of how the participants speaks. It includes silent pauses, paused speech or ideas.
Cut words	Ending speaking a word before completing the utterance of the word (e.g., say “exer...” instead of “exercise”). Breaking the words or interrupting the words while they are being spoken. Might involve correcting syntax/speech
Slow	Reducing the speed of speaking. There is a perceptible change in the speed while someone is talking, making it slower or de-accelerated. It differs from paused speech or cutting off words since the words, phrases or ideas are not cut off or left in the middle, there are no significant silences in the answers. It can include elongating syllables or words. Speed change is determined in comparison to the person’s own baseline.
Fast	Changes in the speed of the answers, making it faster. Information is given quickly, briskly or lively. Speed change is determined in comparison to the person’s own baseline.
Volume	Changes in volume of speech. Differences in how loud or quiet an answer is shared, or there can be differences in the volume of specific words or syllables. Includes: Raising volume, lowering volume, high volume, low volume, and mumbling. Volume change is determined in comparison to the person’s own baseline.
Shaky voice	When there is an rapid fluctuation or trembling of rhythm or tone (i.e., there is instability) to the way someone is speaking. It includes voice shaking, quivering in voice. Excludes case when shaking is due to laughing
Breath	Audible breath, inhaling or exhaling, it can be while the person is talking or before/after a phrase. It includes changes in the breathing rhythm, intensity or deepness of the person (compared to the baseline) that create a sound. Includes sigh, deep breath
Click	Quick sound made by pressing the tongue against the roof of the mouth or back of the teeth and snapping it downward. It often signals disapproval, unsureness or impatience. The sound resembles a “tsk” or “tsk-tsk.”
Laugh	Engaging in laughter, or variations thereof (e.g., snicker, chortle, giggle)
Stuttering	Involuntary repetition of sounds while speaking. This can be seen as a disruption or blocking of the speech by prolongation sounds or by struggling to say a word or a part of a word. Even though the stutter might cut off a word or phrase it is different since the person will finish the word or idea. Includes stammer, stumble.

Table 12: BAH dataset annotation codebook: audio cues.

Body cues	Definition
Look away	Moving the orientation of the head away from the baseline position such that eyes or the gaze will look away. Includes the head facing down, head facing up, looking down, looking up, looking from side to side, lowering head, raising head.
Shake	Turning the head from side to side, it can be done with repetitive head movements or with a slight turn of the head to one or both sides. Includes shaking head ”no”. Rotation is on the horizontal plane
Tilt	Angling the head to the side without focusing on something else, and holding the position. Changing the position of the head so it is in a sloping position. It can be accompanied by changes in the gaze but not necessarily. Includes head tilting up and down, tilting head to the side, tilted head. Includes bobbling head.
Throw	Throwing the head in a rapid movement in a particular direction.
Sigh	Movements of the chest, shoulder or head that accompany a sigh or a deep breath. It includes long sigh, deep breath, sigh, big sigh. Noticeable bringing the chest or diaphragm muscles up and down. Change determined in comparison to the person’s own baseline.
Nod	Moving the head up and down. Lowering and raising the head, it can be done by slight or clearly marked movements. Includes movements such as back and forward or a single small nod.
Shrug	Raising of the shoulders, it can be a momentary or slight rise or a longer movement where one or both shoulders is raised. It includes shrugging shoulders, shrugs
Hands	Movements or placement of the hands that differs from baseline
Posture	Movements in the overall positioning of the spine, body or arms (independent from the head). The changes are determined by each person’s baseline. It includes movements like readjusting in the seat, sloughing, turning to the sides. Needs to involve more than just the head. Excludes shrugging.
Scratch	Movements in the hands and arms to scratch or caress another part of the body or face. It includes scratching head, scratching neck, scratching eyes, scratching chin
Restless	Rhythmic and repeated movements. Can be swaying, shaking, being jittery.

Table 13: BAH dataset annotation codebook: body cues.

Cross-modal inconsistency	Definition
FL	Face and language/speech do not match. E.g., looking uncomfortable while saying yes, looking annoyed or uncomfortable while saying they are happy, smiling while saying they are worried.
FA	Face and audio do not match. E.g., speaks in a sad, energetic tone while smiling.
FB	Face and body do not match. E.g., Nodding while looking afraid or concerned, showing disgust but leaning forward
LA	Language/speech and audio do not match. E.g., speaks in a sad, energetic tone while saying they are happy.
LB	Language/speech and body do not match. E.g., seems like they are about to say something but do not, nod is discrepant with verbal speech, shaking head while saying yes
AB	Body and language/speech do not match. E.g., unengaged tone while nodding (in agreement)

Table 14: BAH dataset annotation codebook: cross-modal inconsistency cues - occurring simultaneously.

J BASELINE RESULTS

Visual features	AVGF1	AP
Cropped faces	**0.5028** ± 0.0078	**0.1923** ± 0.0054
Full frame (body)	0.4781 ± 0.0209	0.1774 ± 0.0063
Head-pose	0.4321 ± 0.0213	0.1915 ± 0.0021

Table 15: Impact of visual modality type on model performance on test set of BAH at frame-level classification. ResNet18 backbone is used. We report the average and standard deviation of 5 repetitions with different seeds.

J.1 VISUAL MODALITY: CROPPED FACES VS FULL FRAME VS HEAD-POSE FEATURES

While cropped faces are the de facto visual modality used in affective computing, other visual features could be also relevant in A/H recognition. We explore here two additional visual features that are related to cues used by annotator to detect A/H. In particular, we consider the full frame which covers the body which is an important cue. In addition, we use head-pose estimation which covers the body cue "Look away". To estimate the head pose, we use the pretrained model TokenHPE (Zhang et al., 2023a). The backbone is frozen to yield head-pose features, which are followed with an A/H classifier layer. Results obtained in Tab.15 show that while cropped faces and full frame yield competitive results, head-pose features have poor results in terms of AVGF1. This suggests that using only head-pose is not enough as many other cues are lost. However, it could be used to model context. Additionally, cropped faces results are more stable. We note also that the full frame case contains noisy background which may have contributed in decreasing performance. Segmenting the body alone could yield better performance. In all our experiments, we use cropped faces for the visual modality.

J.2 SUPERVISED LEARNING CASES (FOLLOWUP FROM MAIN PAPER)

Training details. For both cases, pretraining visual models on RAF-DB (Li et al., 2017), AffectNet (Mollahosseini et al., 2019), and Aff-wild2 (Kollias & Zafeiriou, 2019), their fine-tuning, and final training with multimodal setup, we used a learning rate between 0.0009 to 0.001 with multiplying coefficient of 10. When training on BAH, and in the case of using context, we used a window size between 24 (1 second) to 2880 (2 minutes) with a step of 1 second (24 frames). In this case, we use a mini-batch size in $\{2, 4, 8, 16, 32\}$, where a sample in the mini-batch is a window of frames; and a single-GPU training. In all trainings, we used a weight decay of 0.0001. All our experiments were conducted on a server with 4 NVIDIA A100 GPUs with 40 GB of memory, AMD EPYC 7413 24-Core Processor, and 503GB of RAM. We present in Table 16 the computation time of the multimodal case.

Ablation over the window length. We conduct an ablation to study the impact of the context (window length) on the performance of recognizing A/H. To this end, we use a window length from 24 to 3264 with a step of 1 second (24 frames). Figure 20 shows the obtained results. By considering WF1, performance improves with the increase of the context where it can reach above 0.825. However, F1 does not necessarily improve with the increase of the context. On the other hand, AP prefers small context. Using a small context of few seconds could be a good compromise for all the metrics. Note that the average length of an A/H segment is around 4 seconds (96 frames).

Video-level classification. To obtain video level predictions, we resort to a simple post-processing of frame-level predictions in the main paper. In particular, we follow (De-la Torre et al., 2015), where a sliding window averages the probability of the positive class at each frame. The case where high probability in a window suggests the presence of A/H in the video. In our case, we consider a context of 1 second (24 frame). The maximum probability across all windows is considered the video probability to have the positive class. The probability of the negative class is the complement of the probability of the positive one. We then proceed to measure the same performance metrics AVGF1, and AP. Similar to the main paper, we report the performance at video-level for different cases: using visual backbone only without and with context (Table 17), multimodal (Table 18),

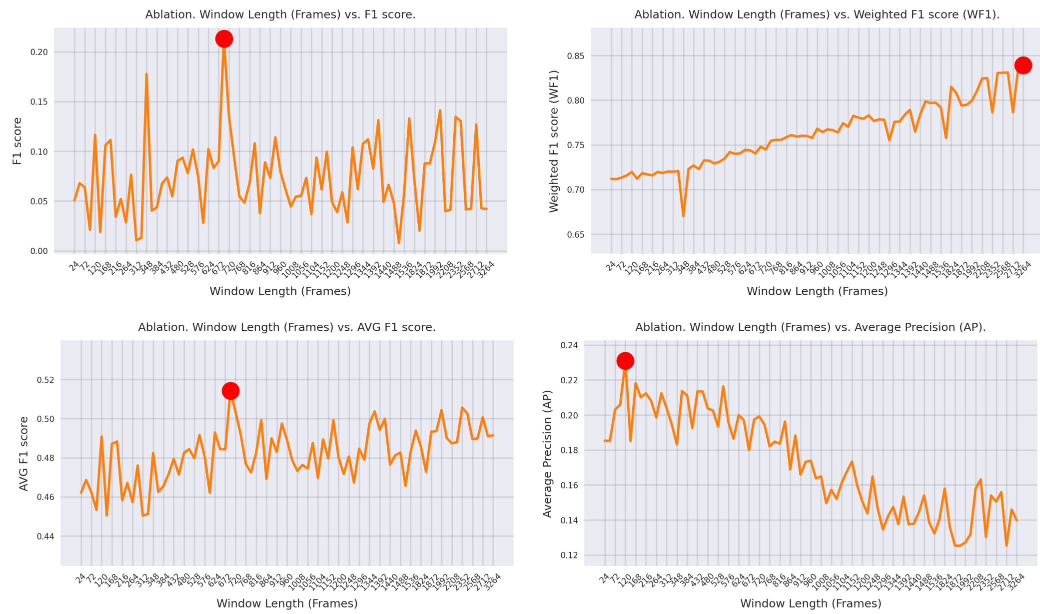


Figure 20: Impact of context (window) length on the performance of frame-level classification when using visual modality alone (ResNet18): `F1`, `WF1`, `AVGF1`, and `AP`. Best performance is indicated in red dot.

Case	Value
Train time 1 epoch	$\sim$ 5mins
Inference time per-frame	$\sim$ 0.12ms
Total n. params.	$\sim$ 223M
N. learnable params.	$\sim$ 5M
N. FLOPs	$\sim$ 1.87 TFLOPs
N. MACs	$\sim$ 938 GMACs

Table 16: Computation time, number of parameters, number of FLOPs/MACs for multimodal case with visual, audio and text (with ResNet152 for visual backbone). Visual backbone is frozen, while audio and text backbones are used to extract features offline and store them.

	Without context		With context (TCN)	
Backbone	`AVGF1`	`AP`	`AVGF1`	`AP`
APViT (Xue et al., 2022)	**0.5871**	0.2882	**0.6134**	**0.3703**
ResNet18 (He et al., 2016)	0.4445	0.0993	0.4425	0.1437
ResNet34 (He et al., 2016)	0.4419	0.1598	0.4448	0.2188
ResNet50 (He et al., 2016)	0.4405	0.1456	0.4442	0.1247
ResNet101 (He et al., 2016)	0.4236	0.1315	0.4448	0.1622
ResNet152 (He et al., 2016)	0.4715	**0.1852**	0.4316	0.2053

Table 17: Visual modality performance on test set of `BAH` at video-level classification: impact of architecture and context.

and fusion (Table 19). Similar to frame-level results, using context and multimodal yields better performance. In addition, FAN fusion yields the highest performance over both metrics.

Modalities	AVGF1	AP
Visual	0.4448	0.1622
Audio	0.4546	**0.5306**
Text	0.5589	0.4731
Visual + Audio	0.4428	0.1964
Visual + Text	0.5614	0.3407
Audio + Text	0.4366	0.3148
Visual + Audio + Text	**0.5934**	0.3219

Table 18: Multimodal models performance on test set of BAH at video-level classification. For visual modality, ResNet152 backbone is used.

Fusion type	AVGF1	AP
LFAN (Zhang et al., 2023b) *(cvprw,2023)*	**0.5934**	**0.3219**
CAN (Zhang et al., 2023b) *(cvprw,2023)*	0.4011	0.3206
MT (Waligora et al., 2024) *(cvprw,2024)*	0.4448	0.1069
JMT (Waligora et al., 2024) *(cvprw,2024)*	0.4448	0.0993

Table 19: Feature fusion performance on test set of BAH at video-level classification.

J.3 Zero-shot Inference: Multimodal Large Language Models (M-LLMs)

Multimodal LLMs (M-LLMs) have gained significant attention in the affective computing space due to their ability to infer cross-modal dynamics across the visual, aural, and textual modalities. The problem of detecting ambivalence and hesitancy in videos is inherently multimodal as it requires also capturing the cross-modal inconsistency. To get out-of-the-box performance of existing SOTA M-LLM, we performed zero-shot inference using the 'Video-LLaVA-7B-hf' (Lin et al., 2024). Since the performance of an M-LLM or LLMs in general can be heavily influenced by the query prompt, we experiment with different variations of the prompts. Table 20 summarizes the different prompt variations used for zero-shot inference.

	Prompt
Simple	'Classify the emotion in the video as either '*Non-Ambivalen*t' or '*Ambivalent*'.' Respond with only one word: '
Definition 1	'Definition: Ambivalence is the state of having contradictory or conflicting feelings or attitudes towards something or someone simultaneously. Classify the emotion in the video as either '*Non-Ambivalent*' or '*Ambivalent*'. Respond with only one word: '
Definition 2	'Definition: Ambivalence and Hesitancy is understood as the simultaneous experience of desires for change and against change. Classify the emotion in the video as either '*Non-Ambivalent*' or '*Ambivalent*'. Respond with only one word: '
Transcript + Def 1	'Video transcript: {transcript}. Definition: Ambivalence is the state of having contradictory or conflicting feelings or attitudes towards something or someone simultaneously. Classify the emotion in the video as either '*Non-Ambivalent*' or '*Ambivalent*'. Respond with only one word: '
Transcript + Def 2	'Video transcript: {transcript}. Definition: Ambivalence and Hesitancy understood as the simultaneous experience of desires for change and against change. Classify the emotion in the video as either '*Non-Ambivalent*' or '*Ambivalent*'. Respond with only one word:'

Table 20: Summary of prompt variations used for zero-shot inference.

J.3.1 FRAME-LEVEL PREDICTION

For frame-level prediction, we adopt a segment-wise strategy, where the entire video is divided into 8-frame chunks and passed through the model using a sliding window. This way, the model sees all the frames in each video. A single prediction is obtained for the window, which is replicated for the segment to match the total number of frame labels in each video. The model's output, 'Non-Ambivalent' or 'Ambivalent', is mapped to 0 and 1 respectively to match the ground truth.

Prompt	AVGF1
Simple	0.4416
Definition Only 1	0.4456
Definition Only 2	0.1651
Transcript + Def 1	**0.4535**
Transcript + Def 2	0.1849

Table 21: Frame Level Prediction using M-LLM.

Table 21 shows the results obtained for frame-level predictions using different prompts. The best results for frame-level prediction are obtained using the 'Transcript + Def 1' prompt, where the actual transcript of the video is also provided, along with a straightforward definition of A/H. The model performs better with a more straightforward definition of the concept of ambivalence, with Definition 1.

In the segment-wise approach applied with a sliding window of 8 frames, the model essentially sees every frame, but this approach limits the context window to be 1/3 of a second, which may not be enough to capture the temporal dependencies in the visual modality. We investigate the effect of various lengths context windows on the overall performance.

Context Window	AVGF1
24 Frames	0.4539
48 Frames	0.4542
80 Frames	0.4535
120 Frames	**0.4546**
192 Frames	0.4540

Table 22: Performance comparison with increasing size context window for frame-level prediction.

We selected the best-performing query prompt from the first experiment (Table 21) to perform the ablation on the context window size. Table 22 shows the results with different lengths of context window for the visual modality. 24 frames represent a one-second context window. Increasing the context window size to 120 frames (5 seconds) only marginally improves the overall performance of the model, and it plateaus at 120 frames and then starts to drop which is an indicator that the visual encoder starts losing information with a longer context window.

J.3.2 VIDEO-LEVEL PREDICTION

For video-level prediction, the entire video is fed to the model, and the transcript is embedded in the prompt. The model selects 8 uniformly spaced frames from the video and predicts a single output. Similar to frame-level predictions, the model's output is mapped to 0 and 1, and the performance metrics are calculated.

Table 23 presents the video-level prediction results. Similar to frame-level predictions, the 'simple' prompt without any context on the definition or the transcript performs the worst and predicts all samples to be 'Non-Ambivalent'. A similar trend is also observed here, i.e., adding the definition and the transcript substantially affects the model performance.

Prompt	AVGF1
Simple	0.2827
Definition Only 1	0.3326
Definition Only 2	0.3772
Transcript + Def 1	**0.6341**
Transcript + Def 2	0.3945

Table 23: Video Level Prediction using M-LLM.

J.3.3 Analysis and Discussion

The performance of M-LLM with zero-shot inference is substantially influenced by the query prompt. As observed from tables 21 and 23, simply asking the model to predict emotion based on the visual modality only performs the worst, whereas adding only the definition of A/H in the query prompt helps the model better identify the positive(A/H) class. Best results in all cases are obtained with the introduction of the text transcript of the video in the query prompt. We conjecture that this happens for two reasons: i) the textual modality serves a significant role in the identification of the A/H class, and ii) the current M-LLMs' performance is heavily reliant on the textual modality. This coheres with the overall structure of traditional M-LLMs that are built upon well-trained LLMs with the addition of a visual encoder like ViT, which is used to encode the visual information that is fed to the LLM for downstream tasks. Intuitively, the performance should increase with careful fine-tuning on the BAH dataset.

Further, the idea of textualizing the aural and visual modalities explored in (Richet et al., 2024) can be well-suited for a task like this where the audio and visual modalities essentially summarize the cues detected in the corresponding modalities. Particularly for tasks like subtle emotion recognition or the detection of A/H, where cross-modal inconsistency has to be considered. Textualizing the aural and visual modalities can be done to adequately exploit the reasoning abilities of SOTA LLMs.

J.4 Personalization using Domain Adaptation

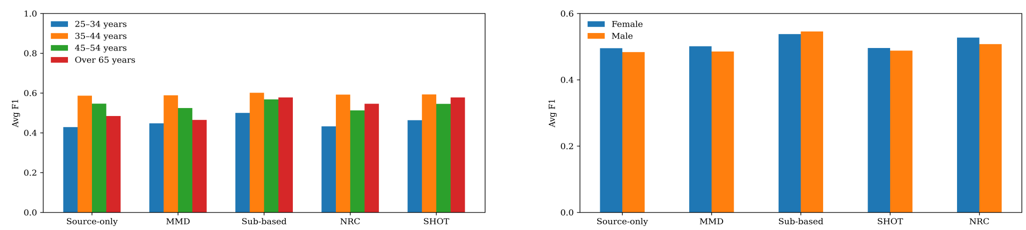
Domain adaptation (DA) (Han et al., 2020; Li & Deng, 2018) has emerged as a promising approach for personalized expression recognition, where the model is trained on diverse labeled source data to generalize to unlabeled target domains representing individual users. Recent research emphasizes on subject-based domain adaptation (Sharafi et al., 2025a;b; Zeeshan et al., 2024; 2025), where each individual is defined as a distinct domain. DA will be employed to personalize ML models by considering each participant in the test set as a separate target domain.

Experimental Protocol. For personalized in BAH, we adopt the standard protocol from prior work (Zeeshan et al., 2024; Sharafi et al., 2025b), which involves partitioning the data of each target individual into train, validation, and test sets. Given the class imbalance in the BAH dataset, we ensure a balanced representation of positive and negative samples within each split. We establish the following baseline methods to evaluate the effectiveness of personalized BAH recognition: **Source-only**: The model trained on the source data is directly evaluated on the target individual test set without any adaptation. This assesses the generalization capability of the source model. **Unsupervised Domain Adaptation (UDA)**: Source data with labels is utilized to adapt the model to each target individual using unlabeled data. This explores the potential of leveraging source knowledge for personalization in the absence of target labels. **Source Free Unsupervised Domain Adaptation (SFUDA)**: Adaptation is performed solely using the unlabeled data from the target individual, without access to the source data. This examines the feasibility of personalization when the source data is unavailable. **Oracle**: The model is fine-tuned using the labeled data from the target individual during training. This provides an upper-bound performance, representing a fully supervised model.

Visual backbone. We employ a ViT-based model for personalization, leveraging its superior performance over ResNet-based architectures for visual tasks without contextual information. In all our experiments, we utilize a ViT-based model pre-trained on the source data.

Methods	AVGF1	AP
Source-only	0.4894 ± 0.0999	0.3565 ± 0.1841
UDA (MMD) (Sejdinovic et al., 2013)	0.4931 ± 0.0943	0.3589 ± 0.1831
UDA (Sub-Based) (Zeeshan et al., 2024) *(fg,2024)*	$\mathbf{0.5417 \pm 0.0728}$	$\mathbf{0.3739 \pm 0.1789}$
SFUDA (SHOT) (Liang et al., 2020) *(icml,2020)*	0.4919 ± 0.1056	0.3520 ± 0.1656
SFUDA (NRC) (Yang et al., 2021) *(neurips,2021)*	0.5174 ± 0.1041	0.3688 ± 0.1487
Oracle	0.5864 ± 0.0751	0.4181 ± 0.1750

Table 24: Performance of UDA and SFUDA with Source-only and Oracle on BAH. Results are reported as the average over all target subjects (mean $\pm$ standard deviation).



(a) Comparison between different age groups on DA methods.

(b) Comparison between different DA methods based on participant Sex.

Figure 21: Comparison between different age groups and sex on DA methods.

J.4.1 Unsupervised Domain Adaptation:

We investigated two unsupervised domain adaptation (UDA) approaches for personalized BAH recognition: (i) a discrepancy-based method using Maximum Mean Discrepancy (MMD) (Sejdinovic et al., 2013) to minimize the domain gap and improve performance on the target subject, and (ii) a subject-based (Zeeshan et al., 2024) method using self-supervision that trains the model by generating pseudo-labels for the target domain, followed by reducing the domain shift using MMD that aligns source and target.

Implementation detail. In UDA experiments, we optimize our model using Stochastic Gradient Descent (SGD) (Sutskever et al., 2013) with a learning rate of $2e-4$, momentum of 0.9, weight decay of $5e-4$, and a cosine annealing scheduler (Loshchilov & Hutter, 2017) with a minimum learning rate of $2e-5$. We set the batch size to 64 and run each target adaptation for 10 epochs. For the subject-based method, we introduce a hyperparameter $\gamma_3 = 0.01$ to weight the target loss, computed using pseudo-labels generated by the Augmented Confident Pseudo-Label (ACPL) technique (Zeeshan et al., 2024). This weighting is essential for mitigating noise in the pseudo-labels, in conjunction with a confidence threshold of 0.95 that is updated every 4 epochs.

J.4.2 Source Free Unsupervised Domain Adaptation

Two source-free unsupervised domain adaptation (SFUDA) approaches were explored for personalized BAH recognition: (i) a representation learning strategy inspired by hypothesis transfer (Liang et al., 2020), where information maximization was used to adapt the model to the target domain, and target-specific prototypes guided pseudo-labelling for class-level alignment, and (ii) a neighbourhood-based method (Yang et al., 2021) in which label consistency was encouraged among target features and their reciprocal nearest neighbours, while expanded neighbourhoods were used to aggregate local structure and reduce the impact of noisy supervision through self-regularization and affinity-weighted loss.

Experimental protocol. We optimize the model using SGD with a learning rate of $1e-4$, momentum of 0.9, and weight decay of $1e-3$. The model is trained for 30 epochs with a batch size of 64. For NRC-based adaptation, we maintain memory banks of target features and predictions to retrieve $K = 3$ nearest neighbours and $M = 2$ expanded neighbors.

J.4.3 Result and Analysis

The average performance across target participants, along with the standard deviation, is presented in Table 24. All reported results are based on evaluations of the respective target test sets. Our analysis demonstrates the effectiveness of domain adaptation for personalized detection of the A/H class in the BAH dataset. All tested methods surpass the Source-only baseline in `AVGF1` and `AP` positive-class metrics. Notably, Sub-based achieves the highest `AVGF1` (0.5417) and `AP` (0.3739), outperforming other domain adaptation techniques. While MMD and SHOT yield only modest gains, they still lag behind the Sub-based method, underscoring the advantage of pseudo-labeling for boosting minority-class recall and precision. In contrast, NRC achieves a more substantial improvement of around 3 percentage points over Source-only, indicating that even with some sensitivity to domain shifts, it more effectively exploits target-domain structure. The Oracle upper bound (`AVGF1`: 0.5864) underscores the substantial potential for further advancements in positive-class detection within this context. Even slight degradations in negative-class performance disproportionately impact the overall `AP` score. For example, Sub-based method emphasizes enhancing positive class identification, likely incurs a cost in precision or recall on the more frequent negative class, a necessary compromise to effectively detect the A/H class.

Age-wise analysis. Figure 21a illustrates the varying impact of age on DA methods. Overall performance tends to peak in the *35–44 years* group across most methods, while the *25–34 years* group generally shows the lowest `AVGF1`. **Sub-based** consistently achieves the highest or near-highest performance in all age groups, with clear gains over Source-only, and is only slightly outperformed by **SHOT** in the *Over 65 years* group. **SHOT** remains competitive across all ages and, together with Sub-based, yields the largest improvements for the youngest (*25–34 years*) and oldest (*Over 65 years*) subjects. In contrast, **MMD** and **NRC** provide only modest gains over Source-only in the *25–34 years* and *35–44 years* groups and can even underperform Source-only for *45–54 years*, highlighting their sensitivity to age-related domain shifts. Finally, **Source-only** has consistently lower Avg F1 across all age groups, confirming the benefit of DA, particularly for the more challenging age ranges.

Sex-wise analysis. In the Figure 21b, we can observe that the *female* category generally exhibits higher values across most methods compared to the *male*. Specifically, *female* shows the highest values in **NRC** (0.0.52), **SHOT** (0.0.49), and **MMD** (0.50). It can also be noted that the number of female subjects is equal to the male subjects.