A key challenge in the accurate prediction of viewers’ emotional responses to video stimuli in real-world applications is accounting for person- and situation-specific variation. An important contextual influence shaping individuals’ subjective experience of a video is the personal memories that it triggers in them. Prior research has found that this memory influence explains more variation in video-induced emotions than other contextual variables commonly used for personalizing predictions, such as viewers’ demographics or personality. In this article, we show (1) that automatic analysis of text describing their video-triggered memories can account for variation in viewers’ emotional responses, and (2) that combining such an analysis with that of a video’s audiovisual content enhances the accuracy of automatic predictions. We discuss the relevance of these findings for improving on state of the art approaches to automated affective video analysis in personalized contexts.

1 INTRODUCTION

The experience of specific feelings and emotional qualities is an essential driver for people to engage with media content, e.g. for entertainment or to regulate their mood [5]. For this reason, research on Video Affective Content Analysis (VACA) attempts to automatically predict how people emotionally respond to videos [7]. This has the potential to enable media applications to present video content that reflects the emotional preferences of their users [34], e.g. through facilitating emotion-based content retrieval and recommendation, or by identifying emotional highlights within clips. Existing VACA approaches typically base their predictions of emotional responses mostly on a video’s audiovisual data [68].

In this article, we argue that many VACA-driven applications can benefit from emotion predictions that also incorporate viewer- and situation-specific information as context for accurate estimations of individual viewers’ emotional responses. Considering context is essential, because of the inherently subjective nature of human emotional experience, which is shaped by a person’s unique background and current situation [30]. As such, emotional responses to videos can be drastically different across viewers, and even the feelings they induce in the same person might change from one viewing to the next, depending on the viewing context [62]. Consequently, to achieve affective recommendations that meaningfully match a viewer’s current desires, e.g., to see an “amusing” video, these will likely need some sensitivity for the conditions under which any potential video will create this experience specifically for him/her. Therefore, addressing variation in emotional responses by personalizing predictions is a vital step for progressing VACA research (see the reviews by Wang et al. [68] and Baveye et al. [7]). Despite the potential benefits, existing research efforts still rarely explore incorporating situation- or person-specific context to personalize predictions. Reasons for this include that it is both difficult to conceptualize context (identifying essential influences to exploit in automated predictions), as well as to operationalize it (obtaining and incorporating relevant information in a technological system) [33]. Progress towards overcoming these challenges for context-sensitive emotion predictions requires systematic exploration in computational modeling activities, guided by research from the social sciences [28].

Here, we contribute to these efforts by exploring the potential to account for personal memories as context in automatic predictions of video-induced emotions. Findings from psychology indicate that audiovisual media are potent triggers for personal memories in observers [9, 41], and these can evoke strong emotions upon recollection [37]. Once memories have been recollected, evidence shows that they possess a strong ability to elicit emotional responses, hence their frequent use in emotion induction procedures [44]. Moreover, the affect associated with media-triggered memories – i.e., how one feels about what is being remembered – has a strong connection to its emotional impact [6, 27]. These findings indicate that by accessing the emotion associated with a triggered memory, we are likely to be able to obtain a close estimate of the emotion induced by the media stimulus itself. Moreover, they underline the potential that accounting for the emotional influence of personal memories holds for applications. They may enable technologies a more accurate overall reflection of individual viewers’ emotional needs in affective video-retrieval tasks. However, they might also facilitate novel use-cases that target memory-related feelings in particular, such as recommending nostalgic media content [22].

One possible way to address memory-influences in automated prediction is through the analysis of text or speech data in which individuals explicitly describe their memories. Prior research has revealed that people frequently disclose memories from their personal lives to others [56, 67], likely because doing so is an essential element in social bonding and emotion regulation processes [11].
There is evidence that people share memories for similar reasons on social media [16], and that they readily describe memories triggered in them by social media content [22]. Moreover, predicting the affective state of authors of social media text [45], as well as text analysis of dialog content [17], are active areas of research. However, apart from the work of Nazareth et al. [50], we are not aware of specific efforts to analyze memories. Together, these findings indicate that it may be feasible to both (1) automatically extract emotional meaning from free-text descriptions of memories and (2) that such descriptions may be readily available for analysis by mining everyday life speech or exchanges on social media.

This second property may also make memory descriptions a useful source of information in situations where other data may be unavailable due to invasiveness of the required sensors – e.g., when sensing viewers’ facial expressions or physiological responses. Motivated by the potential of analyzing memories for supporting affective media applications, we present the following two contributions:

1. we demonstrate that it is possible to explain variance in viewers’ emotional video reactions by automatically analyzing free-text self-reports of their triggered memories, and
2. we quantify the benefits for the accuracy of automatic predictions when combining both the analysis of videos’ audio-visual content with that of viewers’ self-reported memories

2 BACKGROUND AND RELATED WORK

2.1 Video Affective Content Analysis

With respect to the use of context, existing work can be categorized into two types: (1) context-free VACA and (2) context-sensitive VACA.

Context-Free VACA: Works belonging to this type simplify the task of emotion prediction by making the working assumption that every video has quasi-objective emotional content [68]. Traditionally, researchers define this content as the emotion that a stimulus results in for a majority of viewers (i.e., its Expected Emotion [7]). The goal of VACA technology then is to automatically provide a single label for a video representing its expected emotional content, while ignoring variation in emotional responses that a video might elicit.

Existing technological approaches for this task primarily consist of machine learning models trained in a supervised fashion on human-annotated corpora [68]. The ground truth for expected emotions is formed by aggregating the individual emotional responses from multiple viewers for the same video (e.g., by taking the mean or mode across the distribution of their responses). These models can then be used to automatically label entire databases of videos with tags representing their emotional content. The whole process of prediction of an individual viewer’s emotional response using context-free VACA consists of two primary stages, as shown in the graphical overview displayed in Figure 1. (1) using a pre-trained VACA model to automatically label any video in a database of interest (with its expected emotion), and then (2) relying on the video label to be a plausible approximation for the specific emotion that any individual viewer experiences (i.e., his/her Induced Emotion).

There are two groups of technological approaches to automatically label videos in this way, differing in the information that they rely on as input for their predictions [68]. The first, and traditionally most widespread approach is Direct VACA (see region A in 1), which exclusively uses features derived from the audio and visual tracks of the video stimulus itself as the basis for predictions. The second is Indirect VACA, which is looking into automated approaches for analyzing the spontaneous behavior displayed by viewers to label videos without having to ask them for a rating. In essence, this approach uses measures of physiological responses or human behavior (e.g., [38, 63]) from a sample audience to predict the expected emotion of a population for a video (see region B in Figure 1). As such, this endeavor is closely connected to the broader research on emotion recognition from human physiology and behavior in affective computing (see D’mello and Kory [25] for a recent overview). Unfortunately, during the writing of this article, we found that the conceptual differences between the two research efforts are not necessarily clearly defined. Therefore, in this paper, we define emotion recognition to be approached using only measures of an individual’s behavior or physiology to predict his/her specific emotional response to a video (e.g., as in Shukla et al. [60]). In contrast to this, Indirect VACA methods (also known as Implicit Tagging [64]) collect behavioral or physiological data from multiple viewers in response to videos and use an aggregate of these measurements as input to label videos with their expected emotion automatically.

Context-free approaches intentionally blend out issues of subjectivity and variation. In essence, they rely entirely on the expected emotion-labels to be reasonable approximations for the emotional responses of an individual viewer to the video, independent of who he or she is, or in what situation they are watching the video. Because the prediction target is a video-specific aggregate, context-free affective content can only be valid if stimuli display only a small amount of within-video variation in responses. Creators of corpora for VACA modeling have typically enforced this by selecting only stimulus videos for which a low degree of induced emotional variation was already observed (e.g., [1, 38, 42]). Nevertheless,
In naturalistic viewing conditions, such strongly homogeneous responses are unrealistic [62]. We argue that variation in emotional impact is the norm rather than the exception. Consequently, VACA research needs to expand its notion of affective content from being quasi-objective to inherently subjective and context-dependent if it wants to make predictions that meaningfully reflect individual’s media experiences in the real world.

**Context-sensitive VACA:** We consider approaches as context-sensitive VACA when they (1) attempt to predict an individual viewer’s affective responses to a video (i.e. the Induced Emotion [7]), and (2) rely on both the analysis of a video’s audiovisual content and the context for this task.

Existing works can further be distinguished according to the type of context that they use:

- **Viewer-specific context** refers to properties and traits of individual viewers aimed at accounting for variation in video-induced emotions between different individuals (e.g. demographics, personality), while
- **Situation-specific context** denotes information that is temporarily relevant for predicting a viewer’s emotional responses. It covers influences that stretch over multiple videos viewed by the same person in succession (e.g. the type of social setting in which viewing takes place), or that may be specific to only a single instance of viewing (such as a triggered memory).

See Figure 2 for a schematic overview of context-sensitive VACA.

According to this coarse categorization, initial efforts have explored the impact of viewer-specific context, capturing their personality and cultural background [32, 59]. Other research has touched on situational properties, such as mood or the time of day at which someone watches a video [62]. It has also explored the impact of viewing in a group compared to being alone [71]). Finally, measurements of an individual viewer’s behavior or physiology while watching a video can be considered fine-grained information about their current situation and is explored extensively in existing computational approaches (e.g., [38, 69]).

### 2.2 Representing Video-Induced Emotions

Researchers in psychology do not universally agree upon a single system to formally categorize the various states and experiences that make up human emotion. Instead, there exist competing schemes and taxonomies, many of which have been used to represent video-induced emotion [7]. These fall broadly into two distinct groups: categorical and dimensional schemes. Categorical schemes describe human emotion in terms of a set of discrete emotional states, such as happiness or anger. In contrast to this, dimensional frameworks describe subjective emotional experiences as points in a continuous, multidimensional space. Each dimension in the scheme intends to capture a vital characteristic to differentiate emotional experiences from one another.

Computational work on video-induced emotions has traditionally favored categorical representation schemes [68], as does the field of automated emotion recognition in a broader sense [25]. However, psychological research has substantially criticized the theories underpinning prominent categorical representations (e.g., [3, 52]). Moreover, encoding emotions with a limited amount of categories may fail to capture differences at a level that is desired by applications, making dimensional schemes an attractive alternative for affective computing [31]. Perhaps the most prominent scheme for psychological investigations of media-induced emotions is the dimensional PAD framework [43]. It describes affective experiences in terms of the dimensions **pleasure (P)** (is an experience pleasant or discomforting?), **arousal (A)** (does it involve a high or low degree of bodily excitement?), and **dominance (D)** (does it involve the experience of high or low control over the situation?). PAD has been used to code several prominent datasets for VACA research (e.g. DEAP [38] and MAHNOB-HCI [63]), as well as stimuli that are widely used for emotion elicitation in psychology (e.g. the IAPS image corpus [39]). As such, PAD may form a particularly good basis for incorporating insights from psychological research into VACA technology. While VACA practitioners sometimes discard the more comprehensive PAD scheme in favor of the simpler Pleasure-Arousal (PA) scheme [57], there are sound reasons for including dominance when modeling emotional responses, e.g., because it relates to central emotional appraisals [14].

Overall, there exists no consensus in VACA research about a particular way to represent emotional responses, resulting in substantial variation in approaches. While not a problem in itself, this makes it challenging to compare the psychological and technological implications of different computational research projects. Because of its relatively widespread use in VACA and the high degree of compatibility with psychological research, we adopt the PAD dimensional scheme for our data collection and modeling.

### 3 THE MEMENTOS DATASET

In this section, we describe the creation and relevant elements of a crowd-sourced dataset that we have collected for our experiments. The Mementos Dataset contains detailed information about viewers’ responses to a series of music videos, including self-reports about their emotional responses and free-text descriptions of the memories that were triggered while watching them.
3.1 Data Collection Procedure
We recruited 300 crowd-workers via Amazon Mechanical Turk, each receiving 6 USD for their participation. After providing their informed consent about participation and data usage, subjects filled in a survey with additional information about themselves and their situation (Additional Context Measures). Then each participant viewed a random selection of 7 stimuli from our pool of 42 music videos. After each video, we requested ratings for the emotions it had induced (Induced Emotion), followed by a questionnaire asking whether watching the video had triggered any personal memories. If this was the case, viewers were required to describe each of these memories with a short text (Memory Descriptions) and rate the feelings that they associate with these recollections (Memory-Associated Affect). This procedure resulted in a total of 2098 unique responses from the participants (49 to 50 for each video clip). Out of these, 978 responses from 260 unique participants involved the recollection of at least one personal memory. For the modeling activities reported in this article, we focus only on the subset of data that triggered personal memories. See Table 1 for summary statistics.

3.2 Video Stimuli
The dataset contains a selection of 42 music video segments from among a set of 150 that the researchers creating the DEAP corpus have previously evaluated for their induced affect [38]. We opted for these videos because of two reasons. The first one is the strong capacity of musical stimuli to trigger personal memories [37]. The second one is that the creators of the DEAP corpus collected PAD-ratings from multiple viewers for each evaluated video. We used these stimulus-wise ratings to balance the sample of videos that we selected for our study to produce both low and high variation responses across all three affective dimensions of the PAD-space. For example, for each video where viewers in the DEAP validation study displayed a high agreement in the pleasure dimension, we selected another one where agreement on pleasure was low).

3.3 Self-Report Measures
Induced Emotions: We took self-reports for emotions that participants experienced while watching the video in the form of pleasure-, arousal- and dominance-ratings on a continuous scale in the interval of [−1, +1]. Participants provided them using the AffectButton instrument – a 2d-widget displaying an iconic facial expression that changes in response to users’ mouse or touch interactions. Participants can select the facial expression that best fits their affect (see [15] for a detailed description and validation).

Memory-Associated Affect: We asked participants to rate the affective associations with each personal memory that a video has triggered using the AffectButton instrument. Participants could report and rate as many memories per video as they had experienced. However, only 51 out of 978 responses from viewers involved recollections of 2 or more memories. For these instances, we retained only the single memory for our modeling activities with the most intense associated affect in terms of PAD-ratings.

Memory Descriptions: Participants were requested to describe personal memories triggered in them with a short free-text description. We requested a response in English with a minimum length of three words. After filtering for multi-recollection responses (see the previous paragraph), we retained a 978 memory description.

Additional Context Measures: Participating viewers filled in a survey that provided us with additional person- and situation-specific information. This demographic information about viewers’, i.e. their age, gender and nationality, as well as their personality traits in terms of the 6-factor HEXACO-model (Honesty-Humility, Emotionality, eXtraversion, Agreeableness, Conscientiousness, and Openness to experience) collected with a brief questionnaire [23]. Finally, viewers’ reported their mood at the time of participation (PAD ratings with the AffectButton).

4 Influence of Personal Memories on Video-Induced Emotions
In a previous study based on the Mementos dataset [27], we have demonstrated that (1) videos create more intensive and positive emotional responses when triggering personal memories in a viewer, and that (2) the occurrence of and affect associated with memories explains more variability in induced emotions than theoretically relevant viewer-specific measures, e.g., personality traits.

To further illustrate the relevance of personal memories as context for predicting individual viewers’ experiences, we extend these earlier investigations here with a focused statistical analysis of only those responses that involve memories. Concretely, we quantify
the variance explained by memory-associated affect in comparison to that of viewer-specific context variables captured in the dataset (see Demographics, Personality, and Mood in [1]). For this purpose, we conduct a regression analysis with nested linear mixed-effects models targeting the emotions induced in viewers (one model per affective dimension: Pleasure (P), Arousal (A), and Dominance (D)). Figure 3 shows the differences in explained variance by these models for (1) a baseline model (Vid), predicting only the video-specific average (i.e. the Expected Emotion in context-free VACA), (2) the combined contribution of Demographics, Personality, as well as Mood as additional predictors in the model (+ (D + P + M)), and finally (3) the additional effect of memory-associated affect (+Ma).

We see that the combined viewer-specific measures account for a significant share of additional variance over the baseline (P: $\Delta R^2_m = .052, F(13, 213.55) = 4.05, p < .001$; A: $\Delta R^2_m = .088, F(13, 219.21) = 5.75, p < .001$; D: $\Delta R^2_m = .036, F(13, 213.68) = 2.85, p < .001$). However, information about memory-associated affect explains a large share of unique variation on top of that (P: $\Delta R^2_m = .364, F(16, 268.76) = 48.1, p < .001$; A: $\Delta R^2_m = .332, F(16, 278.06) = 31.86, p < .001$; D: $\Delta R^2_m = .301, F(16, 263.22) = 263.22, p < .001$). In fact, the average amount of additional variance explained by memory-associated affect for affective dimensions is about six times higher. Together, these findings provide extremely compelling evidence for the exploitation of video-triggered memories as a form of situation-specific context for predictions in VACA.

5 Predictive Modeling

In this section, we outline the framework that we use as a proof of concept for integrating descriptions of personal memories triggered by a video as contextual information. We have made essential design choices, in how we collected our data to stay true to recognized affective theory and its relation to the task at hand. As a consequence, no state-of-the-art baseline exists from which we can compare our approach to others. We do not intend to provide significant extensions to the technological state-of-the-art. Instead, we provide a state-of-the-art baseline to investigate the contribution of memory descriptions such that we can better expose the nature of this novel task approach. Concretely, we describe the machine learning models and multimodal fusion strategies that we deploy for this task.

5.1 Overview

In line with previous work on VACA ([68]), we address modeling viewers’ emotional responses as a regression problem. An important aspect of context-sensitive VACA technology is the integration of information from different sources into a single multimodal prediction model, i.e., video content and context features. Previous VACA work has repeatedly relied on early fusion to combine the information provided by different sources. Support Vector Machines are a popular type of machine learning algorithm for this purpose (e.g., [42, 59], especially when predictions take place in a regression setting (see [68]). We also consider the state of the art approaches for emotion prediction from text since we intend to exploit free-text descriptions of memories as context in predictions. Interestingly, state-of-the-art results for predicting the author’s affective states from short texts (i.e., tweets) have been achieved by decision-level fusion. For instance, Duppada et al. used random forest regressors with a regularized linear model as a meta regressor different text feature-sets [29]. Motivated by this, we explore both early and late fusion approaches in our experiments. Figure 4 provides a graphical overview of the two machine learning pipelines that we deploy: an Early Fusion-Approach, consisting feature-level fusion of the different information sources, and a Late Fusion-Approach, using a meta-regressor to combine the predictions from modality-specific models via stacking. For all machine learning algorithms, we used the implementation provided by the Scikit-Learn python library [53] in Version 0.22.2.

Early Fusion-Approach: This pipeline consists of a series of pre-processing and feature extraction steps that are specific for memory, visual, and audio sources. The resulting feature vector is then concatenated, and fed to a Support Vector Regressor (SVR) with Radial Basis-Function (RBF) kernel for predictions.

Late Fusion-Approach: Here, we also perform the same modality-specific pre-processing and feature extraction operations. We then rely on two separate SVRs with RBF kernel for predictions based on the audio or visual features, while, in line with [29], we use a Random Forest Regressor for predictions based on memory descriptions. We combine the output of the modality-specific models with an L2-regularized linear regression model (Ridge regression).

5.2 Stimulus Video Processing

To represent the audio content of the music videos, we used the software openSMILE in the configuration ‘âĂIJemobase2010’ for feature extraction. It derives low-level descriptors from raw audio signals in a windowed fashion and aggregates them statistically, resulting in a 1582-dimensional feature vector (see [58] for a detailed description).
This feature set is widely used for audio representation in VACA research as a baseline approach in benchmarking challenges [24].

For visual representation of the stimulus video, one frame is extracted every second. For each frame, we then extract three types of visual features:

**Theory-inspired Descriptors:** Work on affective visual content analysis has developed features that were specifically engineered to capture the affective properties of images. These descriptors are inspired by findings from psychological research or art-theoretic concepts. We use the set of descriptors developed by Machajdik & Hanbury [40], as well as those of Bhattacharya et al. [10] to characterize each of the extracted video frames (resulting in a 271-dimensional feature vector). This combination has been used in context-sensitive VACA work before [59].

**Deep Visual Descriptors:** Deep neural networks form an essential part of modern approaches to visual content analysis and computer vision. Instead of engineering descriptors, these models learn effective and reusable representations for prediction tasks directly from visual training data. We use the activation of the FC1-layer of a pre-trained VGG16 network [61] from the Keras framework for python [20] for this purpose (resulting in a feature vector with 4096 dimensions). This representation has been used for visual representation in VACA research as a baseline in benchmarking challenges [24].

**Visual Sentiment Descriptors:** Classifiers that identify the presence of Adjective-Noun Pairs (ANPs) in images have been successfully used as high-level descriptors in VACA approaches (e.g., [42, 59]). ANPs consist of labels denoting objects or persons identified in an image, coupled with an affective attribute (e.g., "beautiful house"). We use the class-probabilities assigned by the DeepSentiBank Network [18] for any of the ANPs in its ontology as descriptors for the memory descriptions (e.g. “can’t” is transformed into “cannot”). To model the affective impact of personal memories we rely on features that have proven successful in state-of-the-art work modeling emotional numbers with 0 terms (e.g. “that year” or “that decade”). Additionally, we replace any specific years or decades (e.g. “1990”, or “the 90s”) with generic terms (e.g. “that year” or “that decade”).

To explore whether this decrease is a result of our modeling choices, we investigate how well humans can infer affective information from the memory descriptions in our dataset. For this purpose, we leverage two pre-trained word embedding-models to represent each word in the memory description texts as a real-valued feature vector: (1) a Word2Vec-model pre-trained on the Google News dataset, resulting in a 300-dimensional feature vector when applied to a word, and (2) a GloVe-model [55] pre-trained on the Wikipedia 2014 and the Gigaword 5 corpora. It encodes individual words as a 200-dimensional feature vector. For both implementations we rely on the Gensim-library for python [72]. To generate a representation for the entire memory description from these word-level embeddings, vectors of both types are concatenated and averaged for the entire description, resulting in a single 500-dimensional feature vector for each memory.

### 6.2 Results and Analysis

#### 6.2.1 Experiment 1 – Using Memory Descriptions for Prediction

We conduct two experiments in an ablation setting of our early and late fusion approaches when predicting the video-induced emotions of viewers in the Mementos Dataset. In Experiment 1, we assess the feasibility of predicting viewers’ emotional responses to a video from text describing the memories that it has triggered in them. Experiment 2 quantifies the relative contribution of memory descriptions when used for predicting emotional responses alongside the audiovisual content of videos.
purpose, two raters manually annotated a random selection of 150
descriptions with two kinds of affective evaluations for pleasure,
arousal, and dominance: (1) – Perceived Conveyed Affect (PCA) of
the text, and (2) the Inferred Affective Experience (IAX) of the author.
For PCA ratings, the annotators answer the question “What feelings
does this text express?”. We instruct them only to consider emotions
or feelings that are explicitly described by the authors, e.g., by us-
ing emotion words like “love” or “hate”. Performance on this task
will provide us with insights about how explicit authors describe
their emotions in the text. In the case of IAX ratings, annotators
answer the question “How do you think the person describing this
memory feels about it? Put yourself into their situation”. The motiva-
tion for the different task formulation is to encourage annotators
to draw on their cultural background and experience to infer im-

cplicit emotional meaning from the descriptions. Such inferences
are a vital component of human emotion perception [4], and per-
formance on this task shows the degree to which the texts facilitate
them. Raters provide PAD annotations with the widely adopted and
validated Self-Assessment Manikin instrument [12]. Based on this
information, we assess the correspondence of the two annotators’
ratings with viewers’ self-reported ratings for memory-associated
affect (MA), and the reliability with which they were able to do so.
For this, we calculate Pearson correlations between the average
PCA/IAX ratings of both annotators and viewers’ MA ratings to
measure correspondence. Similarly, we calculated the correlations
between the PCA/IAX ratings between both raters as a measure of
reliability and agreement. Results for both are listed in Table
2. Our findings show that annotators’ ratings for pleasure in the
PCA and IAX tasks both highly agree with viewers’ experienced
emotions, as well as each other. This pattern is still present – albeit
less strongly pronounced – for dominance. However, for arousal,
analysts’ judgments correspond much less with viewers’ MA
ratings. Moreover, annotators also tend to disagree much more
with each other. On average, both correspondence and reliability of
judgments are higher in the IAX task than for the PCA task. This
result confirms our hypothesis that simulating the authors’ state
of mind helps human annotators. However, levels of performance
do not still reach those displayed for pleasure or dominance. This
finding points towards an inherently greater difficulty for recogn-
izing arousal from text, rather than to a particular weakness in
our modeling approach. A likely explanation is that descriptions
contain few explicit or implicit expressions of an author’s arousal,
making it challenging for both humans and automatic text analysis
to perform well.

6.2.2 Experiment 2: Combining Direct VACA and Memory Context.
Figure 6 shows the performance displayed by our approaches when
having access to either (1) only videos’ audiovisual data (AV), or
(2) a combination of both audiovisual data and memory descriptions
(AVM). In addition, we also list the performance of a model that
predicts the video-wise mean for affective dimensions of induced
emotions from a sample (AV�). This measure indicates the best-case
performance that a context-free VACA model can provide for the
current dataset if it always makes the correct prediction of the
“Expected Emotion” for a video.

Our analysis shows that memory descriptions can provide sub-
stantial additional information about viewers’ affective responses,
independent of fusion strategy. Overall, improvements when giv-
ing models access to memory descriptions are most pronounced
for predictions of pleasure (Early Fusion: ∆AvgR² = +.059; Late
Fusion: ∆AvgR² = +.092). This evidence is consistent with the
findings of experiment 1, which demonstrates the greater ability
of our models for predicting pleasure. Similarly, the performance
for arousal remains poor across the board, even when using audiovisual features only. Despite the comparably simple modeling approach that we have deployed, the performance of our models using only audiovisual features (AV) approximates that offered by an ideal context-free VACA model (AV+1) (absolute differences in performance averaged across affective dimensions – Early Fusion: ΔAVqR² = .003; Late Fusion: ΔAVqR² = .006). Finally, we observe that the late fusion approach displays a two times greater increase in performance when provided with memory descriptions than the early fusion approach (difference in performance gains averaged across affective dimensions – Early: ΔAVqR² = +.021; Late: ΔAVqR² = +.054). This result highlights that despite the efficacy of early fusion in classic VACA, research should not rule out late fusion for context-sensitive approaches.

7 DISCUSSION

Our empirical investigations demonstrate that it is both feasible to use viewers’ self-reported memory descriptions for predicting emotional experience and that doing so provides a valuable source of context for personalized predictions in video affective content analysis. Particularly for pleasure, the automatic analysis of memory descriptions explains variation that is substantially above and beyond that of a video’s audio and visual content. Surprisingly, none of our models provided substantial insights into viewers’ arousal. Our findings that even humans struggle to reliably and accurately infer arousal from memories offer a possible explanation. Nevertheless, the weak performance for arousal when using stimulus features is surprising, since previous research achieved performance of arousal predictions comparable to pleasure or dominance (e.g., [8, 38, 70]). A reason for this might be that we did not filter our stimuli based on the variability of responses they elicited before modeling. Especially for arousal, our viewers reported widely different levels of arousal to the same video. This fluctuation might make it harder for algorithms to learn directly from a video’s audiovisual content. Similarly, the stimulus-specific mean as “Expected Emotion” might not be a good approximation for these cases. This last finding underlines the limited capacity of purely context-free predictions to reflect viewers’ individual experiences of video stimuli accurately.

Overall, the studies demonstrate that descriptions of video-triggered personal memories are a viable and useful resource for making personalized predictions. However, there are some limitations to our findings. First, the memory descriptions that we analyzed may differ from those that viewers would create in-the-wild, such as on social media platforms, or during a conversation. Second, we have collected them from paid crowd-workers, who provided them with the full knowledge that their identity remains protected. Under these circumstances, participants may have been willing to provide more detailed and candid accounts of their memories than they otherwise would have. Future work could expand on our findings by compiling corpora that capture how people describe media-evoked recollections under more natural conditions in the wild. An effective way to achieve this might be to elicit recollections over more extended periods with specifically developed applications, e.g., via social media [21], or through conversations with interactive intelligent agents [54]. Moreover, research could explore technological efforts to identify and extract memory descriptions from generally available data – e.g., social media posts, – automatically.

However, relying on the availability of explicit descriptions can only be a first step for improving predictions in real-world settings. For many use-cases of VACA, such descriptions may not be available at prediction time, as viewers can only describe the memories triggered in them by a video after they have already been exposed to it, but not before. To anticipate the effect of memories on viewers beforehand – e.g., to personalize recommendations of unseen content –, models will have to estimate when memories are triggered and what their content will be. While these are undoubtedly challenging tasks, progress towards achieving them in automated systems seems feasible. A first step could be to explore how well the use of already obtained text-based memory descriptions generalizes to new, but related stimuli, e.g., music videos from the same artist. Additionally, Dudzik et al. [26] argue that existing research from ubiquitous computing and cognitive modeling offers numerous starting points for modeling memory processes in adaptive media technology. For example, they propose that modeling a user’s attentional focus from sensor data might be a way to identify situations and stimuli that are likely to trigger memories. Given the substantial role that personal memories play for induced emotions, such research topics should be actively pursued in context-sensitive VACA. Naturally, this requires access to rich corpora of data that facilitate such investigations. In this spirit, we plan to publish the Mementos dataset as a resource for future work on modeling affective memory processing during video consumption.

8 SUMMARY AND CONCLUSION

Video-affective Content Analysis (VACA) has traditionally operated under the assumption that videos possess a more or less objective emotional meaning, existing across different viewers and independent of the situation under which they are experienced. However, the emotional impact of videos in the real world is highly subjective and varies greatly with the situation. Consequently, not accounting for context limits the ability of predictions to reflect emotional responses accurately. Contemporary video content is consumed by an increasingly diverse, global-spanning community and in a broad variety of circumstances. Given these developments, research on context-sensitive approaches to VACA seems vital for predictions to be of use to media applications.

Personal memories that are triggered in viewers form a highly situation-specific form of context, shaping individual emotional responses to the video. The two empirical investigations that we describe in this article show the feasibility of using free-text descriptions to account for this influence in automated predictions. Moreover, our findings demonstrate that combining this approach with an analysis of a video’s audiovisual content can provide significant performance benefits. As such, when memory descriptions are available, they offer a powerful form of context for personalizing predictions in VACA. Because people tend to talk about their memories with each other, automatic approaches can feasibly extract such descriptions from communications on social media or face-to-face interactions. Nevertheless, the investigations described in this article only form a first step towards accounting for the influence of memories in automatic predictions. Future investigations should more generally explore modeling the occurrence (when?), content (what?), and influence (what does it do?) of personal memories in VACA predictions to support media applications in the real world.
