# MEmoR: A Dataset for Multimodal Emotion Reasoning in Videos

Guangyao Shen*
Tsinghua University
thusgy2012@gmail.com

Xin Wang†
Tsinghua University
xin_wang@tsinghua.edu.cn

Xuguang Duan
Tsinghua University
duan_xg@outlook.com

Hongzhi Li
Microsoft Research
hongzhi.li@microsoft.com

Wenwu Zhu†
Tsinghua University
wwzhu@tsinghua.edu.cn

## ABSTRACT

Humans can perceive subtle emotions from various cues and contexts, even without hearing or seeing others. However, existing video datasets mainly focus on recognizing the emotions of the speakers from complete modalities. In this work, we present the task of multimodal emotion reasoning in videos. Beyond directly recognizing emotions from multimodal signals, this task requires a machine capable of reasoning about human emotions from the contexts and surrounding world. To facilitate the study towards this task, we introduce a new dataset, **MEmoR**, that provides fine-grained emotion annotations for both speakers and non-speakers. The videos in MEmoR are collected from TV shows closely in real-life scenarios. In these videos, while speakers may be non-visually described, non-speakers always deliver no audio-textual signals and are often visually inconspicuous. This modality-missing characteristic makes MEmoR a more practical yet challenging testbed for multimodal emotion reasoning. In support of various reasoning behaviors, the proposed MEmoR dataset provides both short-term contexts and external knowledge. We further propose an attention-based reasoning approach to model the intra-personal emotion contexts, inter-personal emotion propagation, and the personalities of different individuals. Experimental results demonstrate that our proposed approach outperforms related baselines significantly. We isolate and analyze the validity of different reasoning modules across various emotions of speakers and non-speakers. Finally, we draw forth several future research directions for multimodal emotion reasoning with MEmoR, aiming to empower high Emotional Quotient (EQ) in modern artificial intelligence systems. The code and dataset released on https://github.com/sunlightsgy/MEmoR.

## CCS CONCEPTS

• **Applied computing → Psychology**.

## KEYWORDS

dataset, emotion recognition, reasoning, multimodal

**ACM Reference Format:**
Guangyao Shen, Xin Wang, Xuguang Duan, Hongzhi Li, and Wenwu Zhu. 2020. MEmoR: A Dataset for Multimodal Emotion Reasoning in Videos. In *28th ACM International Conference on Multimedia (MM '20), October 12–16, 2020, Seattle, WA, USA.*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3394171.3413909

## 1 INTRODUCTION

Humans are naturally attuned to perceiving and understanding emotions in the surrounding environment [18]. In our daily life, humans can perceive subtle emotions from various cues and contexts without seeing expressions or hearing others directly. Beyond emotion recognition from observed behaviors, the capability of reasoning about others' emotional states from various contexts and knowledge is an essential aspect of emotional quotient (EQ). Many research areas, including human-computer interactions, multimedia analysis, and social emotion robots, would benefit a lot if modern artificial intelligence systems could be equipped with the reasoning ability to understand human emotions better.

Due to the complicated scenarios and insufficient signals in real-life videos, correctly understanding emotions requires the ability of reasoning. In this work, we present the task of multimodal emotion reasoning in videos. Given a video and an emotion moment, beyond direct recognition with the multimodal signals of target persons, an intelligent machine is expected to be capable of reasoning about human emotions from the environments and the world, including situation contexts, emotion propagation, and external knowledge. Currently, no systematic dataset is created for this challenging task. Existing datasets mainly focus on recognizing the utterance-level emotions of speakers in videos [2, 4, 32, 36, 51]. The speakers in these datasets are usually associated with complete modalities, making it much easier to directly recognize their emotions from multimodal signals. However, non-speakers, who also play indispensable roles in real life, are neglected by existing datasets. In fact, while speakers may be visually absent in real-life scenarios, non-speakers are often visually inconspicuous and always lack audio-textual signals, which urgently requires the capability of multimodal emotion reasoning beyond trivial recognition. Therefore, the existing datasets are inadequate to support understanding the emotions of speakers and non-speakers with incomplete modalities — not to mention developing robust multimodal emotion reasoning systems for general and practical applications.

*Beijing National Research Center for Information Science and Technology (BNRist).
†Corresponding authors.


*MM '20, October 12–16, 2020, Seattle, WA, USA.*

ACM ISBN 978-1-4503-7988-5/20/10…$15.00
https://doi.org/10.1145/3394171.3413909

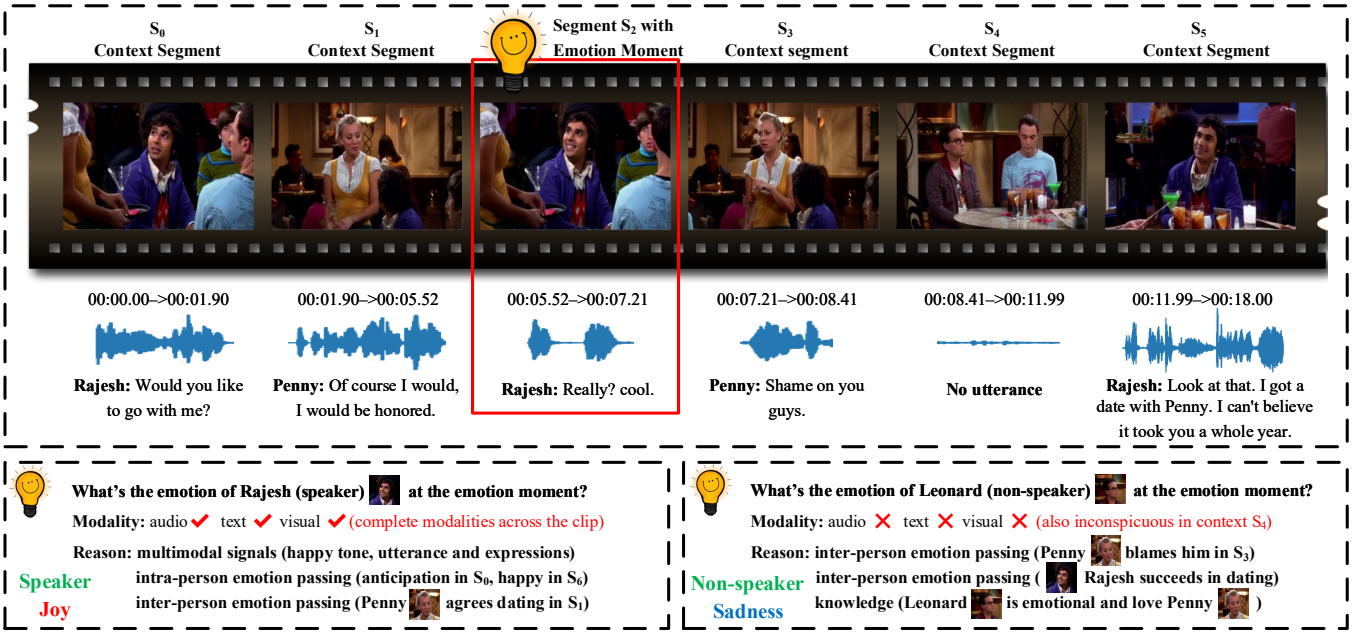



**Figure 1: An example from the *MEmoR* dataset. The top half is a video clip, which is split to semantic segments with aligned multimodal signals. While previous datasets focus on the emotions of speakers, we aim to reason about the emotions of both speakers (bottom left) and non-speakers (bottom right) at an emotion moment with short term contexts and external knowledge. Note that some modalities may be severely missing or inconspicuous in each segment and even across the contexts, especially for non-speakers. Thus, beyond direct recognition from multimodal signals of target persons, MEmoR urgently requires reasoning ability to model the situation contexts, emotion propagation, and external knowledge.**

To solve the above challenge, we propose **MEmoR**, a new dataset for **M**ultimodal **Emo**tion **R**easoning in videos. MEmoR is collected from the popular TV series *The Big Bang Theory* with 5,502 video clips and 8,536 data samples. Different from existing datasets, MEmoR is annotated in person-level with 14 fine-grained emotions from Plutchik's wheel [34] for both speakers and non-speakers. Given that videos in MEmoR are close to real life and persons in these videos may have incomplete signals, an intelligent system must possess emotion reasoning skills beyond direct recognition with complete modalities. In support of developing various reasoning approaches, MEmoR offers short term contexts around emotion moments and external knowledge such as personalities for the main characters in TBBT. Therefore, MEmoR is designed as a practical yet challenging testbed for multimodal emotion reasoning. Fig. 1 illustrates an example of a video clip with two data samples.

As a first attempt, we extract representative multimodal features and propose an attention-based reasoning method. In addition to the multimodal features, our approach reasons about emotions from intra-personal emotion contexts, inter-personal emotion propagation, and personalities. The key lies in the personality-guided self-attention mechanism across the persons and the contexts. We achieves the best performance across several multimodal emotion recognition baselines and conduct experiments to explore the roles of different components in emotion reasoning. Further, we take an in-depth analysis of the performance across different emotion categories for both speakers and non-speakers.

We summarize the contributions of our paper as fourfold: 1) We formalize the task of multimodal emotion reasoning in videos, which requires a deep understanding of the contexts and the world beyond directly recognizing emotions from multimodal signals. 2) We introduce MEmoR, a new dataset annotated with fine-grained emotions for both speakers and non-speakers in videos. The target persons involved in the videos may have incomplete multimodal signals, urgently requiring the capability of multimodal emotion reasoning. 3) MEmoR provides researchers with both short-term contexts and external knowledge in support of emotion reasoning. Besides, we also propose an attention-based approach and carry out extensive experiments to demonstrate the effectiveness of different reasoning strategies. 4) We summarize potential challenges and present several promising future directions for multimodal emotion reasoning with the proposed MEmoR dataset. We believe that MEmoR may help to push the research on affective computing from emotion recognition towards emotion reasoning.

## 2 RELATED WORK

### 2.1 Basic Emotion Theory

In basic emotion theory, humans are assumed to have an innate set of emotions. For example, William James in 1890 proposed 4 basic emotions [22], and Richard in 1996 suggest 15 emotions [25]. Ekman's six basic emotions [12] are the most commonly adopted emotion classification model. However, it is relatively simple and cannot reveal the intensities and relations among the emotions.

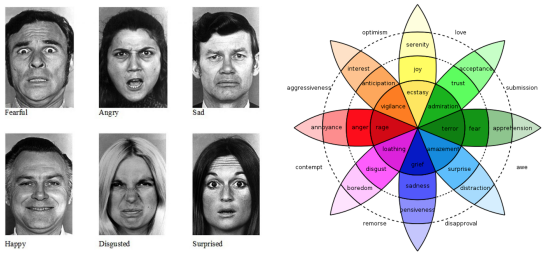



**Figure 2: Ekman's six basic emotions and Plutchik's wheel of emotions (the middle circle contains 8 primary emotions)**

**Table 1: Comparison between related datasets and MEmoR. "all$^+$" indicates both speakers and non-speakers. "?" indicates that the faces may be missing for target persons.**

| Dataset | Anno. lv. | Target | Face | Know. | #Emo |
|---|---|---|---|---|---|
| CMU-MOSEI | utter. | speakers | ✓ | × | 6 |
| OMG | utter. | speakers | ✓ | × | 7 |
| SEMAINE | utter. | speakers | ✓ | × | 7 |
| IEMOCAP | utter. | speakers | ✓ | × | 9 |
| MELD | utter. | speakers | ? | × | 7 |
| MEmoR | person | all$^+$ | ? | ✓ | 14 |

In this work, we leverage the Plutchik's wheel of emotions [34], including 8 primary and 24 fine-grained emotions, which also covers Ekman's emotions. The fine-grained emotions are of three different intensity scales for each of the primary emotions. Similar emotions lie close in the wheel. With Plutchik's wheel, we can make a more fine-grained emotion analysis for the emotion reasoning task.

## 2.2 Multimodal Emotion Recognition in Videos

Multimodal emotion recognition in videos aims to perceive human emotions from audio, text, and visual signals. The previous datasets mainly focus on recognizing the utterance-level emotions of speakers with full modalities. For example, **OMG-Emotion** dataset [2] and **CMU-MOSEI** dataset [51] collect user-generated monologue videos from YouTube and annotate Ekman's basic emotions in utterance level. Videos in these two datasets are required to include apparent front faces. With the support of these datasets, researchers have developed effective multimodal human emotion recognition methods, including multimodal fusion [8, 28, 29, 50], multimodal transfer learning [1], and multimodal attention mechanism [10, 16, 54]. In addition to the monologue videos, multimodal emotion recognition in conversations has attracted great interest in recent years. In conversations, two (dyadic) or more (multi-party) actors take turns to speak, which raises new challenges such as context modeling and the speaker's emotion shift. **IEMOCAP** [4] and **SEMAINE** [32] are two dyadic conversation video datasets that provide audio-visual signals by recording conversations with faces in front of fixed cameras. **MELD** [36], a recent multimodal dataset, provides wild multi-party conversation videos from TV shows. For these datasets, RNNs [30, 35], memory networks [19, 20] and graph neural networks [14] are applied to model the dynamic temporal emotions among the speakers.

However, all the exsiting datasets above focus on utterance-level emotions for the speakers, who are usually the key persons in videos with complete modalities. In contrast, MEmoR is more generic provides person-level annotations for both speakers and non-speakers. Because non-speakers lack audio-textual signals and are often visually inconspicuous, MEmoR requires emotion reasoning ability beyond direct recognition from multimodal signals. The comparison with related datasets is summarized in Tab. 1.

# 3 DATASET

## 3.1 Video Collection and Preprocessing

*3.1.1 Data source.* We aim to evaluate the performance of emotion reasoning models in real-life scenarios, so the video corpus should be generic and representative. We choose the popular TV show *The Big Bang Theory (TBBT)*, which has seven main characters[1] and many extras. Specifically, we utilize all the episodes from the first nine seasons of TBBT to construct our dataset. The videos are transcoded to 720p H.264 streams, while the audios are unified into 16-bit mono streams. As for the text data, we obtain the subtitles and download the episode transcripts from the Internet.

*3.1.2 Data Alignment.* We aim to align the audio, text, and characters for further annotation. As subtitles contain temporal information, and transcripts associate utterances with characters, We should align them accurately so that the utterances can be matched with their corresponding speakers and timestamps. However, there are many utterances grouped within identical timestamps in the subtitles. In order to locate the accurate timestamp for each utterance, we use a force-aligner tool *Gentle*[2] to make word-level speech-text alignments. In this way, we associate each utterance with its accurate timestamps and characters.

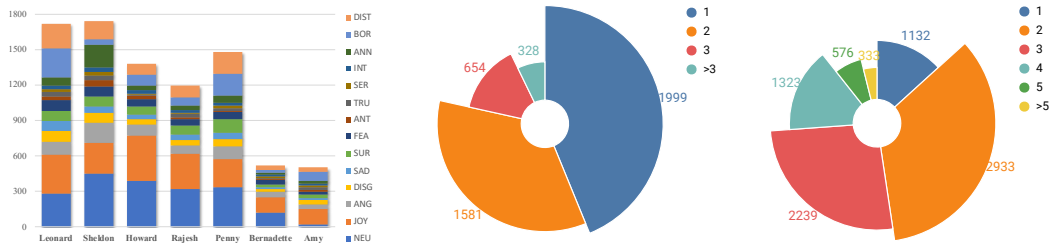
*3.1.3 Video Segmentation.* With the aligned data, we split the video into semantic segments (Fig. 1). First, we take all the utterances as initial segments, as they are aligned with non-overlap accurate timestamps. If a gap between the two utterance segments is shorter than 3 seconds, we merge the gap into the left nearest segment as the emotions are continuous during short intervals. Otherwise, we left the gap as a new visual segment. Thus, these non-overlap segments can cover the entire video sequences.

## 3.2 Annotation Process

We build web applications and design a two-step annotation process to reduce the complexity and improve the quality of the annotation process. We conduct on-site training for six annotators with bachelor degrees employed by a professional data labeling company. We adopt the Plutchik's wheel of emotions [34] with 8 primary and 24 fine-grained emotions, as discussed in Sec. 2.1.

*3.2.1 Emotion Moments Annotation.* The first step is to create video clips with potential emotions. As the continuous contexts are expected to be critical for emotion reasoning, we avoid interrupting the emotional experience of the annotators. They are merely asked to pause the video when the main characters present typical emotions on our web application, and the moments are recorded

[1] Leonard, Sheldon, Howard, Rajesh, Penny, Bernadette, and Amy.
[2] https://github.com/lowerquality/gentle




(a) Person-emotion Distribution (b) The number of annotations per clip (c) The number of segments per clip

**Figure 3: Emotions, annotations and segments distribution in MEmoR**

automatically. In this way, we get 6,128 emotion moments (29.89 per episode). For each emotion moment, we create a ten seconds time window centered around it. All the segments that cross this window are concatenated to an emotion video clip. We remove duplicate clips and retain the clips shorter than 30 seconds, which already provide enough contexts for emotion reasoning. Finally, we get 5,502 emotion video clips for next step annotation.

*3.2.2 Emotion States Annotation.* The second step is to annotate the emotions in each video clip. We ask annotators to label the primary and fine-grained emotions of each main character near the emotion moment. To ensure the quality, each clip is annotated by three annotators. For each character, they should first choose whether he or she has emotions. The emotion of the character would be further annotated if and only if he/she has a single emotion. The annotators are asked to give their reason (text, audio, visual, and context) to avoid too hasty decisions. Then, they should label one of the eight primary emotions. Finally, they are required to annotate one of the three corresponding fine-grained emotions. After this stage, we have 12,905 valid annotations each with a single emotion.

*3.2.3 Personality Annotation.* While emotions are very subjective experiences and would be different from person to person, researchers have found that personality has an important effect on daily life emotional processes [24]. In this work, we consider various models including 16PF [7], Big Five [37], and MBTI [3] to describe the personalities from different aspects. In real life, people take tests with tens of self-report questions to recognize their personalities for each model. Therefore, we provide a TBBT fan, who is very familiar with psychology and this show, lots of backgrounds and knowledge to mimic this process. For each of the seven main characters, he substitutes himself into the role and carries out the three tests, leading to a total of 26 dimensions vector to describe personality. The personalities can be seen as a kind of prior external knowledge, which provides more potential for emotion reasoning.

*3.2.4 Data Sample Creation.* We represent each sample as a tuple (clip, person, moment, emotion). To build a high-quality and consistent dataset for emotion reasoning, we only keep those samples in which at least two annotators agree that the character presents one specific fine-grained emotion. Besides, we find it necessary to add the *neutral* category. If a character appears in the video clip and all three annotators agree that he/she presents no emotion at the moment, we take it as a neutral sample. Finally, we create a total of 8,536 samples annotated with fine-grained emotions.

**Table 2: Summary of MEmoR dataset statistics.**

| MEmoR Statistics | Train | Test |
|---|---|---|
| # of different video clips | 4,081 | 1,488 |
| # of data samples | 6,829 | 1,707 |
| Avg./Max segment duration (s) | 5.000/31 | 5.043/30 |
| Total # of utt. segements | 18,218 | 4,514 |
| Total # of vis. segements | 1,009 | 285 |
| # of annos. in seg-level speakers | 2,078 | 502 |
| # of annos. in seg-level non-speakers | 4,751 | 1,205 |
| # of annos. in clip-level speakers | 4,314 | 1,086 |
| # of annos. in clip-level non-speakers | 2,515 | 621 |

*3.2.5 Emotion Sets Selection.* We provide two sets of emotion categories: **Primary Emotions** and **Fine-grained Emotions**. The primary 9 emotions are joy (JOY), anger (ANG), disgust (DISG), sadness (SAD), surprise (SUR), fear (FEA), anticipation (ANT), trust (TRU) from Plutchik's wheel as well as neutral (NEU). For fine-grained emotions, after checking the annotation results, we find it difficult for the annotators to tell the differences between some emotions. Besides, some emotions are not present or well-labeled in TBBT. Therefore, we merge and discard some categories (details in supplementary materials) to get the final 14 fine-grained emotions, which adds serenity (SER), interest (INT), annoyance (ANN), boredom (BOR), and distraction (DIST) to *Primary Emotions*.

*3.2.6 Train/Test Set Split.* We split the MEmoR dataset by randomly select 80% of the samples as the training set and the rest as the test set. We keep their label distributions to be mostly similar. It is worth mention that the samples are the same in primary emotions and fine-grained emotions label sets. The only difference is the choice of emotion sets for all samples.

## 3.3 Dataset Exploration

*3.3.1 Data Format.* Each sample in MEmoR dataset consists of a video clip, a target person, and an emotion moment. We provide primary and fine-grained emotion labels for each sample. A video clip contains: (1) semantic segments with accurate timestamps; (2) aligned audios and texts in utterance segments for the speakers; (3) the individuals (characters) appearing in this video clip. We show an example video clip along with two samples in Fig. 1.

*3.3.2 Dataset Statistics.* Similar to the previous datasets, emotions are imbalanced in MEmoR. Fig. 3(a) shows the distribution of fine-grained emotions for each character. While the emotion distributions are broadly similar across the main characters, the personalities and relations between characters have significant effects on certain emotions. For example, Leonard has much more boredom emotions than Sheldon. This is because Sheldon always centers himself and thinks less about others, which usually makes his roommate Leonard feel boredom. As MEmoR is annotated for all persons, there may be multiple annotations in one clip. We visualize the statistics of annotations per clip in Fig. 3(b). We further investigate two types of speaker statistics. If we treat the speakers as those who are speaking in the **target moments**, about 69.8% of the samples are non-speakers. If we treat speakers as those who are speaking during the **video clip**, about 36.7% of the samples are non-speakers. The high proportion of non-speakers brings the challenge of lacking audio-text signals, which requires emotion reasoning capability beyond direct recognition. While we provide an average of 2.8 context segments per sample on MEmoR, Fig. 3(c) shows that about 13% of the samples are presented as long monologues and most samples have less than five segments, which makes this dataset more diverse and challenging. Tab. 2 shows high-level summary statistics of the train/test set in MEmoR.

*3.3.3 Dataset Highlights.* We summarize the highlights of MEmoR as follows: 1) **Quality.** The videos in MEmoR are split into semantic segments with well-aligned modalities. We designed a two-stage annotation process with thorough on-site training. Finally, we adopt strict rules for sample creation and provide annotations for both primary and fine-grained emotion sets. 2) **Novelty.** While the related datasets only focus on utterance-level emotions for speakers, MEmoR provides person-level fine-grained annotation for both speakers and non-speakers, which is more generic for practical requirements. 3) **Challenge.** The annotated persons, especially for non-speakers, may suffer from severe modality missing, which is challenging and requires the ability of multimodal emotion reasoning. 4) **Potential.** MEmoR provides short-term contexts near emotion moments as well as each character's personalities. Actually, the commonsense knowledge and large-scale specific knowledge about *TBBT* from the Internet could also be used to enhance the reasoning ability further. Thus, MEmoR has great potential to support developing a variety of multimodal emotion reasoning techniques.

# 4 MULTIMODAL EMOTION REASONING

Following Sec. 3.2.4, we formally define a data sample as $(\boldsymbol{V}, P_m, \boldsymbol{S}_n, E_{m,n})$, where $\boldsymbol{V} = \left(\{P_i\}_{i=1}^{M}, \{\boldsymbol{S}_j\}_{j=1}^{N}\right)$ is the video clip containing $M$ persons and $N$ semantic segments (Sec. 3.1.3), $P_m \in \{P_i\}_{i=1}^{M}$ is the target person. $\boldsymbol{S}_n \in \{\boldsymbol{S}_j\}_{j=1}^{N}$ is the target segment where the annotated emotion moment lands inside, and $E_{m,n}$ is the labeled emotion for $P_m$ in $S_n$.

Given that the target person $P_m$ may have no visual signals or lack audio-textual signals (a non-speaker), $\boldsymbol{S}_n$ may miss one and even all the modalities for emotion recognition. Our goal is to reason about the target emotion $E_{m,n}$ of the target person $P_m$ in the target segment $S_n$ through utilizing the contextual information contained in $\boldsymbol{V}$ as well as external knowledge such as personalities.

## 4.1 Multimodal Feature Extraction

For person $P_i$ at segment $\boldsymbol{S}_j$, we extract the multimodal features for audio, text, visual and personality. When a modality is missing, we use a zero vector as the corresponding placeholder.

**Audio Features.** We use openSIMLE [13] to extract 6373-d audio features with the *IS13_ComParE* [40] config file, which is the common practice in the affective computing research community. We perform Z-standardization to the these features and obtain the final audio representation $\boldsymbol{a}_{i,j} \in \mathbb{R}^{6373}$.

**Text Features.** We obtain the text representation using BERT [11], which achieves outstanding performance in various multimedia tasks. Specifically, we use the PyTorch implementation [45] with a pretrained *bert-large-uncased* model. By averaging the sequence of hidden-states in the last layer for the input sequence, we obtain the textual representation $\boldsymbol{t}_{i,j} \in \mathbb{R}^{1024}$.

**Visual Features.** We split the visual features into three folds to describe the visual contents in frame-level. (1) **Facial features:** We use the pretrained MTCNN [53] to extract faces in all frames. To get facial identity features, we finetune a Facenet [39] model pretrained on VGGFace2 [5] for face recognition (FR) on the seven main characters. For facial emotion features, we train another Facenet model on FER2013 dataset [6] for facial expression recognition (FER). Finally, we concatenate the outputs before the final layer of the FR and FER models to obtain the overall 1024-d facial features. (2) **Object features:** We use the Detectron2 [46] library to detect 1230 object categories defined in LVIS dataset [17]. We pack them into a 1230-d feature vector, where each slot represents the number of one object category detected in the given frame. (3) **Environment features:** We feed the whole image into a pretrained Resnet152 model to get the 2048-d image-level feature. We average these features across frames and concatenate them to get the final visual representation $\boldsymbol{v}_{i,j} \in \mathbb{R}^{4302}$. If the person $i$ do not appear in the segment $\boldsymbol{S}_j$, we claims the modality is missing and $\boldsymbol{v}_{i,j}$ is set to zero vector.

**Personality Features.** First, we adopt the annotated 26-d personality features as discussed in Sec. 3.2.3. Based on the findings of [15], we also measure the individuals' personality traits with LIWC [43], where we feed all the utterances of each person and get the word frequencies of 92 psychological word categories. After Z-standardization, we finally concatenate the annotated features and LIWC features to get the personality representation $\boldsymbol{p}_i \in \mathbb{R}^{118}$.

## 4.2 Model Architecture

To reason about the person-level emotions, we propose an attention-based approach to model the intra-personal emotion contexts, inter-personal emotion propagation as well as the prior knowledge of the personalities. In this work, we adopt the scaled dot-product attention proposed in Transformer [44]: $Att(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V$, where $Q$, $K$, and $V$ are query, keys and values, respectively, $d$ is the size of queries and keys. Assuming a video clip $\boldsymbol{V} = \left(\{P_i\}_{i=1}^{M}, \{\boldsymbol{S}_j\}_{j=1}^{N}\right)$ has $N$ segments and $M$ persons, our goal is to recognize the emotion $\boldsymbol{E}_{m,n}$ of the target person $P_m$ in the target segment $\boldsymbol{S}_n$.

**Encoders:** We use three encoders $\mathcal{E}_a, \mathcal{E}_t, \mathcal{E}_v$ to encode the multimodal features $\{(\boldsymbol{a}_{i,j}, \boldsymbol{t}_{i,j}, \boldsymbol{v}_{i,j})\}_{i=1,j=1}^{M,N}$ into compact representations as $\{(\boldsymbol{f}_{i,j}^{(a)}, \boldsymbol{f}_{i,j}^{(t)}, \boldsymbol{f}_{i,j}^{(v)})\}_{i=1,j=1}^{M,N}$, where $\boldsymbol{f}_{i,j}^{(k)} \in \mathbb{R}^{256}$ for $k \in \{a, t, v\}$.

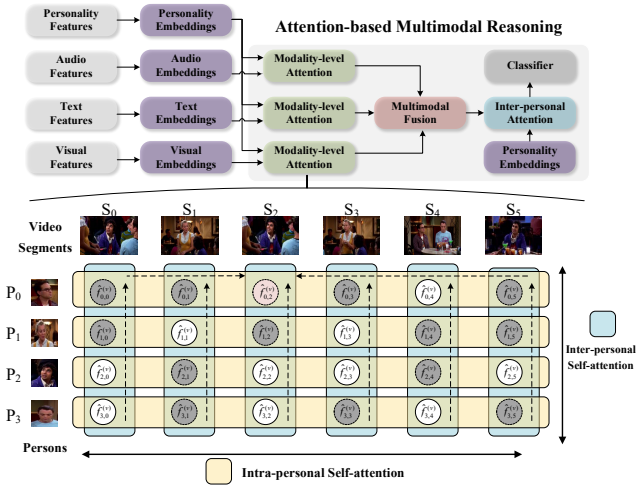



**Figure 4: Our attention-based multimodal reasoning model. For modality-level attention, we show the case of visual modality. The gray circles indicates modality missing, and the pink circle denotes the target features.**

**Personality Embedding:** Next, we adopt a two-layer perceptrons $\mathcal{E}_p$ to encode the personality features $\boldsymbol{p}_i$ as $\boldsymbol{f}_i^{(p)}$ for a person $i$. Served as global knowledge, the encoded personality embeddings are concatenated to enhance the multimodal features $\hat{\boldsymbol{f}}_{i,j}^{(k)} = \left[\boldsymbol{f}_{i,j}^{(k)}; \boldsymbol{f}_i^{(p)}\right]$ for each modality $k \in \{a, t, v\}$.

**Modality-level Attention:** Before fusing the multimodal features, we conduct personality-guided inter-personal and intra-personal attentions for each modality. Specifically, we perform inter-personal self-attention across the person-dimension at each segment $S_j$. $\boldsymbol{h}_{i,j}^{(k),\,1} = \hat{\boldsymbol{f}}_{i,j}^{(k)} + Att\left(\hat{\boldsymbol{f}}_{i,j}^{(k)}, \hat{\boldsymbol{f}}_{:,j}^{(k)}, \hat{\boldsymbol{f}}_{:,j}^{(k)}\right)$

Guided by personality embedding $\boldsymbol{f}^{(p)}$, paired person relation information are implicitly modeled. Due to the severe modality missing and slow convergence speed, we do not adopt RNNs to model the intra-personal emotion contexts. Instead, we use attention along the segment dimension for each person $P_i$ as: $\boldsymbol{h}_{i,j}^{(k),2} = \boldsymbol{h}_{i,j}^{(k),1} + Att\left(\boldsymbol{h}_{i,j}^{(k),1}, \boldsymbol{h}_{i,:}^{(k),1}, \boldsymbol{h}_{i,:}^{(k),1}\right)$. If the modality $k$ for $P_i$ in $S_j$ is missing, we will mask $\hat{\boldsymbol{f}}_{i,j}^{(k)}$ to zero in the modality-level attention.

**Multimodal Fusion:** We take an early fusion strategy for multiple modalities. After fusion, we concatenate the personality embeddings once again to leverage personal information in higher level. Therefore, we get the multimodal representations $\boldsymbol{h}_{i,j} = \left[\boldsymbol{h}_{i,j}^{(a),2}; \boldsymbol{h}_{i,j}^{(t),2}; \boldsymbol{h}_{i,j}^{(v),2}; \boldsymbol{f}_i^{(p)}\right]$ in person-level across all the segments.

**Person-level Inter-personal Attention:** For the target segment $\boldsymbol{S}_j$, we model the high level inter-personal emotion communications by a person-level self attention. Therefore, the final enhanced multimodal representations are obtained as: $\hat{\boldsymbol{h}}_{i,j} = \boldsymbol{h}_{i,j} + Att\left(\boldsymbol{h}_{i,j}, \boldsymbol{h}_{:,j}, \boldsymbol{h}_{:,j}\right)$.

Finally, the enhanced representation $\hat{\boldsymbol{h}}_{m,n}$ for target person $P_m$ in target segment $S_n$ is fed to a three-layer MLP for emotion classification. Fig. 4 sketches the overview of our architecture.

# 5 EXPERIMENTS

## 5.1 Experimental Setup

*5.1.1 Implementation Details.* For the encoders, $\mathcal{E}_a$ is a three-layer Multi-Layer Perception model (MLP) with hidden state sizes of 1024, 512, and 256. $\mathcal{E}_v$ and $\mathcal{E}_t$ are with hidden state sizes of 512, 384, and 256, $\mathcal{E}_p$ is with hidden state sizes of 512, 256. All the encoders use the Relu activation function. Our model is optimized using Adam [23] optimizer with a learning rate 0.00005, beta1 0.9, and beta2 0.999. The batch size is set to 8. For training with the unbalanced data on MEmoR, we adopt the weighted sampling strategies, where the weights are set to the proportion of samples of different emotions. We choose micro-F1, macro-F1, and weighted-F1 scores as the main evaluation metrics. Micro-F1, which equals to the precision and recall in micro-averaging case, measures the overall accuracy. As the dataset is unbalanced, we also report arithmetic mean (macro-F1) and weighted mean (weighted-F1) of F1 scores, which measures per-class performance. We stop training and report the best results when the micro-F1 performance fails to improve for 50 epochs.

*5.1.2 Comparison Methods.* We compares several multimodal emotion recognition baselines and our method on MEmoR.

- **MDL:** For multimodal deep learning architecture, we use an early fusion strategy, which concatenates the encoded features from different modalities for further classification.
- **MDAE:** Following [33], we design a multimodal deep autoencoder to reconstruct all modalities from shared representations for the modality missing setting.
- **BiLSTM+TFN:** Tensor Fusion Network (TFN) [50] performs multimodal fusion on unimodal, bimodal and trimodal components of the data. Before fusion, we summarized modalities using three Bi-LSTMs.
- **BiLSTM+LMF:** Low-rank Multimodal Fusion (LMF) [28] performs multimodal fusion robustly using low-rank tensors to improve efficiency. Before fusion, we summarized modalities using three Bi-LSTMs.
- **DialogueGCN:** DialogueGCN [14] leverages self and inter-speaker dependency of the speakers to model conversational context for multimodal emotion recognition.
- **AMER:** The proposed attention-based multimodal emotion reasoning method and three ablated variants.

## 5.2 Experimental Results

Tab. 3 shows the comparison results between the proposed AMER model and the baselines. The personality features are concatenated after multimodal fusion for each baseline method. MDAE performs worst with no intra-personal information, and the autoencoder fails to obtain good representations with insufficient full modality training data. As the modalities are summarized using three Bi-LSTMs, the multimodal fusion methods TFN and LMF get better performance in primary emotions but fail to capture the subtle emotion difference in fine-grained emotions. With inter-personal information, DialogueGCN achieves high macro-F1 scores, indicating the ability to recognize insignificant emotions. However, the performance is limited because DialogueGCN may aggregate too much invalid information from modality missing vertices. In contrast, our attention-based multimodal emotion reasoning approach

**Table 3: Experimental results of emotion reasoning in primary emotions and fine-grained emotions.**

| Methods | Modality | Primary Emotions | | | Fine-grained Emotions | | |
|---|---|---|---|---|---|---|---|
| | | Micro-F1 | Macro-F1 | Weighted-F1 | Micro-F1 | Macro-F1 | Weighted-F1 |
| MDL | A+V+T | 0.4083 | 0.2817 | 0.3990 | 0.3544 | 0.2126 | 0.3370 |
| MDL with Personality | A+V+T+P | 0.4294 | 0.3170 | 0.4228 | 0.3632 | 0.2171 | 0.3453 |
| MDAE | A+V+T+P | 0.4206 | 0.3034 | 0.4102 | 0.3626 | 0.2185 | 0.3412 |
| BiLSTM+TFN | A+V+T+P | 0.4704 | 0.3104 | 0.4539 | 0.3661 | 0.2068 | 0.3496 |
| BiLSTM+LMF | A+V+T+P | 0.4487 | 0.2943 | 0.4322 | 0.3638 | 0.1975 | 0.3508 |
| DialogueGCN | A+V+T+P | 0.4411 | 0.3100 | 0.4248 | 0.3726 | 0.2285 | 0.3725 |
| AMER w/o Personality | A+V+T | 0.4458 | 0.3391 | 0.4395 | 0.4007 | 0.2460 | 0.3788 |
| AMER w/o Intra-personal | A+V+T+P | 0.4001 | 0.2933 | 0.3933 | 0.3667 | 0.2183 | 0.3446 |
| AMER w/o Inter-personal | A+V+T+P | 0.4634 | 0.3086 | 0.4491 | 0.4030 | 0.2384 | 0.3860 |
| AMER Full | A+V+T+P | **0.4774** | **0.3534** | **0.4652** | **0.4188** | **0.2616** | **0.3996** |

**Table 4: Results for fine-grained emotions in different modality combinations on a full-modality subset.**

| Modalities | Micro-F1 | Macro-F1 | Weighted-F1 |
|---|---|---|---|
| Audio(A) | 0.2917 | 0.1244 | 0.2525 |
| Visual(V) | 0.2281 | 0.1067 | 0.2028 |
| Text(T) | 0.3236 | 0.1632 | 0.3074 |
| A+V | 0.3023 | 0.1296 | 0.2412 |
| T+V | 0.3209 | 0.1962 | 0.3110 |
| A+T | 0.3448 | 0.1696 | 0.3097 |
| A+V+T | 0.3501 | 0.1911 | 0.3188 |

**Table 5: Micro-F1 in different modalities on fine-grained emotions. "Missing" means at least one modality is missing.**

| Ablated Models | A+T | V | A+V+T | Missing |
|---|---|---|---|---|
| AMER w/o Personality | 0.3565 | 0.4134 | 0.3595 | 0.4167 |
| AMER w/o Intra-personal | 0.3217 | 0.3558 | 0.3464 | 0.3621 |
| AMER w/o Inter-personal | 0.3565 | 0.3851 | 0.3648 | 0.4121 |
| AMER full | **0.4086** | **0.4236** | **0.3753** | **0.4318** |

outperforms all the baseline methods with a significant margin in all the metrics on both primary and fine-grained emotions.

## 5.3 Further Analysis

*5.3.1 The Role of Multimodal Signals.* Here we validate the effectiveness of different modalities on a full-modality subset of MEmoR for fair comparisons across the modality combinations. Tab. 4 shows that all the modalities are effective in fine-grained emotions. The text features are the most representative, but audio and visual signals play a good supplementary role. While visual features are relatively weak, it improves the per-class performance because some emotions are highly vision-related.

*5.3.2 The Role of Personality.* We investigate the role of personality in two folds. In the first experiment, we add personality features to MDL as a new modality. In the second experiment, we conduct an ablation study by removing the personality embedding from the proposed AMER model. Tab. 3 shows that in these two experiments, we achieve better performance with the personality information.

**Table 6: Micro-F1 scores on different emotion categories. P and F denote primary and fine-grained emotions.**

| | NEU | JOY | ANG | DISG | SAD | SUR | FEA |
|---|---|---|---|---|---|---|---|
| P | 0.6358 | 0.5913 | 0.4731 | 0.2886 | 0.2857 | 0.3890 | 0.2411 |
| F | 0.6220 | 0.5840 | 0.3739 | 0.0775 | 0.2542 | 0.2784 | 0.2188 |
| | ANT | TRU | SER | INT | ANN | BOR | DIST |
| P | 0.2020 | 0.0741 | - | - | - | - | - |
| F | 0.1364 | 0.1154 | 0.0976 | 0.1818 | 0.2013 | 0.2411 | 0.2801 |

Furthermore, Tab. 5 shows that adding personalities benefits both the full and missing modality samples, especially when the visual modality is unavailable.

*5.3.3 The Role of Emotion Contexts and Propagation.* As the samples on MEmoR suffer from different degrees of modality missing in different segments, the intra-personal emotion contexts and inter-personal emotion propagation play an important role in emotion reasoning. Tab. 3 shows that: 1) the performance goes down when we remove the modality-level and person-level inter-personal attention modules; 2) the modality-level intra-personal attention is key to the AMER model, which improves the performance by a large margin. We can also see in Tab. 5 that the full AMER outperforms the ablated models in different scenarios of modalities, and intra-personal attention is the most effective component in recognizing modality missing samples.

*5.3.4 Performance across Different Emotion Categories.* First, we study the primary emotions in Tab. 6. Among all the categories, the performances on neutral, joy, and angry are much higher than others as they are the most distinguishable sentiments. Besides, Ekman's six basic emotions (Fig. 2) are easier to recognize than anticipation and trust, which may be the reason that previous datasets mainly focus on these significant emotions. Trust is most difficult because people express trust mainly by visual actions like nodding heads, which is hard to be captured by the current feature set. For fine-grained emotions, it is difficult to distinguish two emotions if they are close in the wheel. For example, as shown in Tab. 6, the recognition performances of disgust and serenity are really bad in fine-grained emotions. Indeed, most disgust samples are classified into boredom, and serenity, by definition a kind of "peaceful joy",

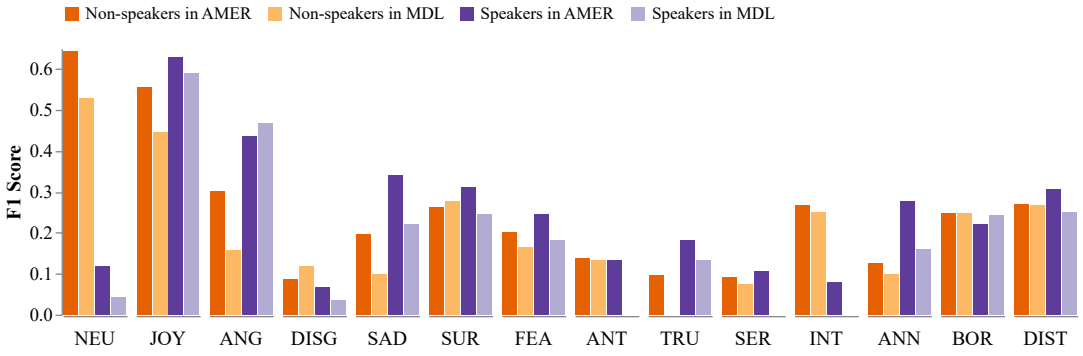



**Figure 5: Comparison between AMER and MDL on speakers and non-speakers across fine-grained emotions.**

is prone to be recognized as neutral or joy. While humans can distinguish the subtle differences of these fine-grained emotions, it is hard for current AI systems to develop the same ability.

*5.3.5 Detail Analysis for Speakers and Non-speakers.* As discussed in Sec. 3, while the previous datasets focus on speakers, MEmoR provides annotations for both speakers and non-speakers. Here we denote speakers as those speaking in the target emotion moments. Fig. 5 compares the performances of speakers and non-speakers between MDL and AMER for each emotion class. We can see that: (1) With the aforementioned reasoning abilities, the proposed AMER model outperforms MDL in most emotion categories for both speakers and non-speakers. Indeed, the overall micro-F1 scores are improved by a large margin (speakers ↑ 3.58%, non-speakers ↑ 7.64%). (2) AMER achieves higher performance for speakers than non-speakers in 9 of the 14 fine-grained emotions, indicating that the emotions of non-speakers are more difficult to recognize. (3) While non-speakers are naturally likely to be passive neutral listeners, speakers are much possible to convey vivid emotions. Therefore, it is not surprising that we can detect neutral emotion better from non-speakers than speakers. Overall, the reasoning abilities benefit the performance of all persons, especially for the non-speakers. More details can be seen in the supplementary materials.

## 6 CONCLUSION AND FUTURE WORK

In this paper, we introduce a challenging dataset MEmoR for the task of multimodal emotion reasoning in videos. While existing datasets mainly focus on recognizing the emotions of speakers, MEmoR is annotated with 14 fine-grained emotions for both speakers and non-speakers from sitcoms. The challenge to solve the emotion recognition problem in the real-world scenario is that the subjects are usually lack of audio-textual signals and visually inconspicuous. Thus, MEmoR requires the capability of emotion reasoning from the contexts and the prior knowledge. We describe the process of building this dataset and provide the multimodal feature sets. We further propose an attention-based reasoning approach and conduct extensive experiments to demonstrate the effectiveness of intra-personal emotion contexts, inter-personal emotion propagation, and prior knowledge. Finally, we take an in-depth analysis of the performance across different emotion categories for the speakers and non-speakers. Besides, the performances are unsatisfactory on some fine-grained emotions, indicating the great potential to further develop powerful techniques of multimodal emotion reasoning.

In addition, in the course of conducting this research, we identified some critical challenges that we believe are important to address in future research on multimodal emotional reasoning.

**Emotion Feature Design.** In this work, we use pre-trained features that are not explicitly designed for recognizing emotions and therefore have great potential for improvement with respect to various emotions. For example, elegantly representing visual cues is obviously very essential for designing emotion-specific features, including aesthetic features [21], human expressions [41], poses [47], actions [9, 52], and their interactions in videos.

**Multi-label Emotion Reasoning.** We create data samples with single label agreed by at least two annotators. However, one may feel sad and angry at the same time. Therefore, a future direction is multi-label emotion reasoning with MEmoR.

**Knowledge-enhanced Emotion Reasoning.** External knowledge has been successfully aggregated in visual reasoning applications [26, 27]. MEmoR can support emotion reasoning from both commonsense knowledge and specific knowledge. The commonsense knowledge can be obtained from knowledge bases like ConceptNet [42] and Atomic [38], and MEmoR could provide specific knowledge about the characters, stories, and backgrounds in TBBT. Thus, various techniques like neural-symbolic reasoning [31, 49], knowledge graph reasoning and graph neural network can be leveraged in multimodal emotion reasoning.

**Explainable Reasoning Procedure.** Some research areas like VQA has moved towards explainable reasoning [48]. While humans can easily point out how they understand emotions around them, the emotion reasoning procedure is still a "black-box" in this work. If we can make explainable reasoning over the causes, outcomes, and expressions of emotions, it will not only improve the performance but also reduce ethical risk in practical applications.

Finally, we believe the MEmoR dataset can push the community of computational emotion analysis from recognition to reasoning and empower high EQ in the modern intelligence systems.

## ACKNOWLEDGMENTS

This work was support by National Natural Science Foundation of China Major Project (No. U1611461) and National Key R&D Program of China under Grand No. 2018AAA0102000.

## REFERENCES


[1] Samuel Albanie, Arsha Nagrani, Andrea Vedaldi, and Andrew Zisserman. 2018. Emotion Recognition in Speech using Cross-Modal Transfer in the Wild. In *ACM International Conference on Multimedia*, Susanne Boll, Kyoung Mu Lee, Jiebo Luo, Wenwu Zhu, Hyeran Byun, Chang Wen Chen, Rainer Lienhart, and Tao Mei (Eds.). 292–301.

[2] Pablo Barros, Nikhil Churamani, Egor Lakomkin, Henrique Sequeira, Alexander Sutherland, and Stefan Wermter. 2018. The OMG-Emotion Behavior Dataset. In *IJCNN*. 1408–1414.

[3] Isabel Briggs-Myers and Peter B Myers. 1995. Gifts differing: Understanding personality type. (1995).

[4] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. *Language resources and evaluation* 42, 4 (2008), 335.

[5] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman. 2018. VGGFace2: A dataset for recognising faces across pose and age. In *International Conference on Automatic Face and Gesture Recognition*.

[6] Pierre-Luc Carrier and Aaron Courville. 2013. Challenges in representation learning: Facial expression recognition challenge. *Kaggle Competition, https://www. kaggle. com/c/challenges-inrepresentation-learning-facial-expression-recognitionchallenge/data* (2013).

[7] Raymond B Cattell and Heather E P. Cattell. 1995. Personality structure and the new fifth edition of the 16PF. *Educational and Psychological Measurement* 55, 6 (1995), 926–937.

[8] Chen Chen, Zuxuan Wu, and Yu-Gang Jiang. 2016. Emotion in Context: Deep Semantic Feature Fusion for Video Emotion Recognition. In *ACM International Conference on Multimedia*. 127–131.

[9] Peihao Chen, Chuang Gan, Guangyao Shen, Wenbing Huang, Runhao Zeng, and Mingkui Tan. 2019. Relation Attention for Temporal Action Localization. *IEEE Transactions on Multimedia* (2019).

[10] Shizhe Chen and Qin Jin. 2016. Multi-modal Conditional Attention Fusion for Dimensional Emotion Prediction. In *ACM International Conference on Multimedia*. 571–575.

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[12] Paul Ekman, Wallace V Friesen, Maureen O'sullivan, Anthony Chan, Irene Diacoyanni-Tarlatzis, Karl Heider, Rainer Krause, William Ayhan LeCompte, Tom Pitcairn, Pio E Ricci-Bitti, et al. 1987. Universals and cultural differences in the judgments of facial expressions of emotion. *Journal of personality and social psychology* 53, 4 (1987), 712.

[13] Florian Eyben, Felix Weninger, Florian Gross, and Björn Schuller. 2013. Recent Developments in OpenSMILE, the Munich Open-Source Multimedia Feature Extractor. In *ACM International Conference on Multimedia*. 835–838.

[14] Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander F. Gelbukh. 2019. DialogueGCN: A Graph Convolutional Neural Network for Emotion Recognition in Conversation. In *ACL*. 154–164.

[15] Haiqian Gu, Jie Wang, Ziwen Wang, Bojin Zhuang, and Fei Su. 2018. Modeling of User Portrait Through Social Media. In *ICME*. 1–6.

[16] Yue Gu, Xinyu Lyu, Weijia Sun, Weitian Li, Shuhong Chen, Xinyu Li, and Ivan Marsic. 2019. Mutual Correlation Attentive Factors in Dyadic Fusion Networks for Speech Emotion Recognition. In *ACM International Conference on Multimedia*. 157–166.

[17] Agrim Gupta, Piotr Dollar, and Ross Girshick. 2019. LVIS: A dataset for large vocabulary instance segmentation. In *CVPR*. 5356–5364.

[18] Paul L Harris. 1989. *Children and emotion: The development of psychological understanding*. Basil Blackwell.

[19] Devamanyu Hazarika, Soujanya Poria, Rada Mihalcea, Erik Cambria, and Roger Zimmermann. 2018. ICON: interactive conversational memory network for multimodal emotion detection. In *EMNLP*. 2594–2604.

[20] Devamanyu Hazarika, Soujanya Poria, Amir Zadeh, Erik Cambria, Louis-Philippe Morency, and Roger Zimmermann. 2018. Conversational Memory Network for Emotion Recognition in Dyadic Dialogue Videos. In *ACL*. 2122–2132.

[21] Zhiyuan Hu, Jia Jia, Bei Liu, Yaohua Bu, and Jianlong Fu. 2020. Aesthetic-Aware Image Style Transfer. In *Proceedings of the 28th ACM International Conference on Multimedia*.

[22] William James. 2007. *The principles of psychology*. Vol. 1. Cosimo, Inc.

[23] Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR*.

[24] Emma Komulainen, Katarina Meskanen, Jari Lipsanen, Jari Marko Lahti, Pekka Jylhä, Tarja Melartin, Marieke Wichers, Erkki Isometsä, and Jesper Ekelund. 2014. The Effect of Personality on Daily Life Emotional Processes. *PLOS ONE* 9 (10 2014), 1–9.

[25] Richard S Lazarus and Bernice N Lazarus. 1994. *Passion and reason: Making sense of our emotions*. Oxford University Press, USA.

[26] Guohao Li, Xin Wang, and Wenwu Zhu. 2019. Perceptual Visual Reasoning with Knowledge Propagation. In *Proceedings of the 27th ACM International Conference on Multimedia*. 530–538.

[27] Guohao Li, Xin Wang, and Wenwu Zhu. 2020. Boosting Visual Question Answering with Context-aware Knowledge Aggregation. In *Proceedings of the 28th ACM International Conference on Multimedia*.

[28] Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, AmirAli Bagher Zadeh, and Louis-Philippe Morency. 2018. Efficient Low-rank Multimodal Fusion With Modality-Specific Factors. In *ACL*. 2247–2256.

[29] Jia-Xin Ma, Hao Tang, Wei-Long Zheng, and Bao-Liang Lu. 2019. Emotion Recognition using Multimodal Residual LSTM Network. In *ACM International Conference on Multimedia*. 176–183.

[30] Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander F. Gelbukh, and Erik Cambria. 2019. DialogueRNN: An Attentive RNN for Emotion Detection in Conversations. In *AAAI*. 6818–6825.

[31] Jiayuan Mao, Chuang Gan, Pushmeet Kohli, Joshua B Tenenbaum, and Jiajun Wu. 2018. The Neuro-Symbolic Concept Learner: Interpreting Scenes, Words, and Sentences From Natural Supervision. In *International Conference on Learning Representations*.

[32] Gary McKeown, Michel François Valstar, Roddy Cowie, Maja Pantic, and Marc Schröder. 2012. The SEMAINE Database: Annotated Multimodal Records of Emotionally Colored Conversations between a Person and a Limited Agent. *IEEE transactions on Affective Computing* 3, 1 (2012), 5–17.

[33] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y. Ng. 2011. Multimodal Deep Learning. In *ICML*. 689–696.

[34] Robert Plutchik. 2001. The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American scientist* 89, 4 (2001), 344–350.

[35] Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. 2017. Context-Dependent Sentiment Analysis in User-Generated Videos. In *ACL*. 873–883.

[36] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations. In *ACL*. 527–536.

[37] S. Rothmann and E. P. Coetzer. 2003. The big five personality dimensions and job performance. *SA Journal of Industrial Psychology* 29, 1 (2003).

[38] Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. 2019. ATOMIC: An Atlas of Machine Commonsense for If-Then Reasoning. In *AAAI*. 3027–3035.

[39] Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *CVPR*. 815–823.

[40] Björn Schuller, Stefan Steidl, Anton Batliner, Alessandro Vinciarelli, Klaus Scherer, Fabien Ringeval, Mohamed Chetouani, Felix Weninger, Florian Eyben, Erik Marchi, et al. 2013. The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism. In *INTERSPEECH*.

[41] Guangyao Shen, Wenbing Huang, Chuang Gan, Mingkui Tan, Junzhou Huang, Wenwu Zhu, and Boqing Gong. 2019. Facial Image-to-Video Translation by a Hidden Affine Transformation. In *Proceedings of the 27th ACM International Conference on Multimedia*. 2505–2513.

[42] Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Thirty-First AAAI Conference on Artificial Intelligence*.

[43] Yla R Tausczik and James W Pennebaker. 2010. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of language and social psychology* 29, 1 (2010), 24–54.

[44] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NeurIPS*. 5998–6008.

[45] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *ArXiv* abs/1910.03771 (2019).

[46] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. 2019. Detectron2. https://github.com/facebookresearch/detectron2.

[47] Zijie Ye, Haozhe Wu, Jia Jia, Yaohua Bu, Wei Chen, Fanbo Meng, and Yanfeng Wang. 2020. ChoreoNet: Music to Dance Synthesis with Choreographic Action Unit. In *Proceedings of the 28th ACM International Conference on Multimedia*.

[48] Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B Tenenbaum. 2019. CLEVRER: Collision Events for Video Representation and Reasoning. In *International Conference on Learning Representations*.

[49] Kexin Yi, Jiajun Wu, Chuang Gan, Antonio Torralba, Pushmeet Kohli, and Josh Tenenbaum. 2018. Neural-symbolic vqa: Disentangling reasoning from vision and language understanding. In *Advances in neural information processing systems*. 1031–1042.

[50] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor Fusion Network for Multimodal Sentiment Analysis. In *EMNLP*. 1103–1114.

[51] AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *ACL*. 2236–2246.

[52] Runhao Zeng, Chuang Gan, Peihao Chen, Wenbing Huang, Qingyao Wu, and Mingkui Tan. 2019. Breaking Winner-Takes-All: Iterative-Winners-Out Networks for Weakly Supervised Temporal Action Localization. *IEEE Transactions on Image Processing* 28, 12 (2019), 5797–5808.

[53] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. 2016. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters* 23, 10 (2016), 1499–1503.

[54] Sicheng Zhao, Zizhou Jia, Hui Chen, Leida Li, Guiguang Ding, and Kurt Keutzer. 2019. PDANet: Polarity-consistent Deep Attention Network for Fine-grained Visual Emotion Regression. In *ACM International Conference on Multimedia*. 192–201.