Temporal Attention and Consistency Measuring for Video Question Answering

Lingyu Zhang
Rensselaer Polytechnic Institute
Troy, New York
zhangl34@rpi.edu

Richard J. Radke
Rensselaer Polytechnic Institute
Troy, New York
rjradke@ecse.rpi.edu

ABSTRACT

Social signal processing algorithms have become increasingly better at solving well-defined prediction and estimation problems in audiovisual recordings of group discussion. However, much human behavior and communication is less structured and more subtle. In this paper, we address the problem of generic question answering from diverse audiovisual recordings of human interaction. The goal is to select the correct free-text answer to a free-text question about human interaction in a video. We propose an RNN-based model with two novel ideas: a temporal attention module that highlights key words and phrases in the question and candidate answers, and a consistency measurement module that scores the similarity between the multimodal data, the question, and the candidate answers. This small set of consistency scores forms the input to the final question-answering stage, resulting in a lightweight model. We demonstrate that our model achieves state of the art accuracy on the Social-IQ dataset containing hundreds of videos and question/answer pairs.

CCS CONCEPTS

• **Computing methodologies** → **Neural networks**; **Activity recognition and understanding**.

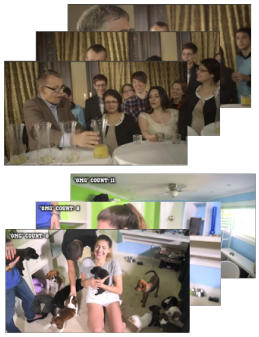
KEYWORDS

Video question answering, multimodal machine learning, attention, human conversation

ACM Reference Format:
Lingyu Zhang and Richard J. Radke. 2020. Temporal Attention and Consistency Measuring for Video Question Answering. In *Proceedings of the 2020 International Conference on Multimodal Interaction (ICMI '20), October 25–29, 2020, Virtual event, Netherlands.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3382507.3418886

1 INTRODUCTION

Automatically understanding human activity in video has made substantial progress, from detecting and classifying behaviors like jumping and waving [3, 14] to automatically producing sentence-level descriptions of clips such as "The man starts dancing after hearing the women playing piano" [5, 27]. This paper addresses the problem of answering questions about the events in a video, such as "Why the man is not happy when the woman brings him a birthday cake?". This is a challenging problem that involves fine-grained actions and behavior, causal connections between events, and multimodal data streams involving multiple people.



How come the man in brown is pouring juice?
A1: He is pouring juice to show a visual metaphor for how vaccines are administered.
A2: He is pouring juice because he wants his guests to have something to drink.

Why does the woman in the gray cut short her conversation with the woman in blue?
A1: She is distracted by the puppies.
A2: She dislikes the woman in blue and wishes to stop talking with her.

Figure 1: Example videos, questions, and answers in the Social-IQ dataset [31]. Green answers are correct.

Making headway on this problem requires appropriate datasets involving natural multi-human interaction. While several group interaction datasets [22, 25, 33] have been constructed to study and predict emotion [28, 33], intention [15], leadership style [16], coordination patterns [24] or collaborative quality [8], most are not suitable for the video question answering problem for several reasons. The main one is that participants are usually seated in the same configurations with a fixed, known camera perspective. Often the number of participants is fixed in a meeting, and high-quality per-participant video (e.g., from a frontal-facing camera) and audio (e.g., from a dedicated microphone) are collected.

In contrast, here we work with a recently proposed human interaction dataset called Social-IQ [31] designed for research on video question answering (VQA) tasks. Social-IQ features diverse topics and environments, varying numbers of people and camera angles in each clip, and unstructured human actions and conversations. Each clip is accompanied by a set of questions relating to the causes of events and intentions and mental states of the participants, as well as corresponding candidate answers that are both true and false. Figure 1 shows two examples of videos and question/answer pairs (QA-pairs) in the dataset, which is discussed in more detail in Section 3.

The Social-IQ VQA dataset presents several challenges compared to existing social signal processing datasets, including:

ICMI '20, October 25–29, 2020, Virtual event, Netherlands

ACM ISBN 978-1-4503-7581-8/20/10…$15.00
https://doi.org/10.1145/3382507.3418886

- The video recordings are not frontal-facing, nor are they individually focused. This makes it difficult to apply facial or body feature extraction algorithms for estimating the Visual Focus of Attention (VFOAs) or expression of each participant.
- Videos frequently cut between different camera perspectives containing different people.
- The transcripts and the audio signals are single-channel, not segmented for each visible participant, making it difficult to extract features such as turn-taking patterns.
- The questions and answers relate not to apparent behaviors, but to hidden signals such as mental states, emotions, and intentions, which requires a fine understanding of the order and timing of the interaction dynamics.
- Both the video clips and the text of the questions and answers are longer than in typical VQA datasets.

In this paper, we present a novel neural network to attack the VQA problem with two key contributions. The first is **temporal attention**: our model processes the video, audio, and transcript streams, as well as the questions and answers, to highlight moments and words that are particularly relevant. The second is **consistency measuring**: we quantify the feature similarity between the multimodal streams, the question, and the candidate answers as the evidence for making the final decision. Our approach is inspired by the way humans approach reading comprehension tasks involving long passages with irrelevant information, i.e., keeping the question and candidate answers in mind and then skimming through the reading material to find possible supporting evidence to get to the right answer quickly and accurately. Figure 2 overviews our overall approach.

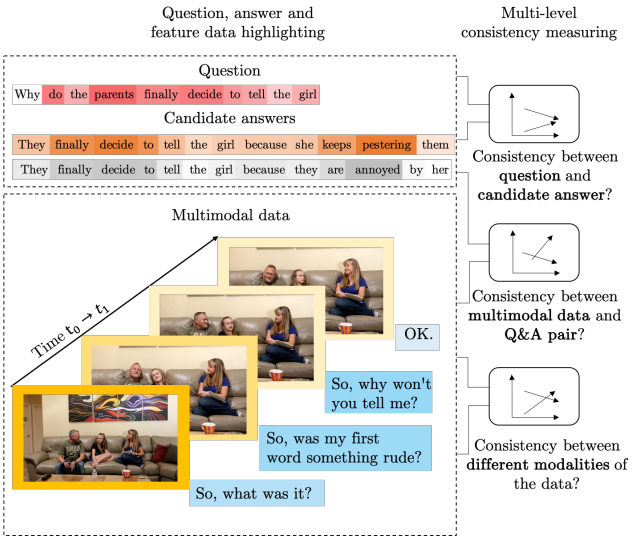


Figure 2: An illustration of two key mechanisms in our approach. Darker colors indicate higher temporal weights automatically extracted by our model.

Since a large intermediate feature map is reduced to a small set of consistency scores before input to the final network, our model is computationally extremely lightweight, and is extensible to general video question answering. We demonstrate state-of-the-art accuracy on the Social-IQ dataset, taking a step towards correctly answering challenging "why" and "how" questions in social signal processing.

2 RELATED WORK

2.1 Multimodal social signal estimation

A considerable body of literature exists on social signal estimation using multimodal machine learning. For example, the ELEA corpus [22] was used to train a Support Vector Machine (SVM) for personality trait [20] and emergent leadership [2] prediction, in which hand-crafted multimodal features such as visual focus of attention and speaking turns are used as behavior cues. Zhang *et al.* [34] proposed a Long-Short Term Memory (LSTM)-based temporal fusion mechanism to model group interactions to capture the co-occurrent and successive behaviors in multimodal recordings for dynamic social status classification. Zadeh *et al.* [32] constructed a tensor fusion network to explicitly aggregate unimodal, bimodal and trimodal dynamics for sentiment analysis in online videos.

2.2 Video captioning

Several algorithms have been proposed for video captioning, the automatic textual description of events in video. Xu *et al.* [29] designed a joint event detection and description network to generate sentences from video. Video features are extracted using 3D convolutional layers. Candidate video segments for event proposals are generated, followed by an LSTM-based language-text fusion and captioning module. Wang *et al.* [27] proposed a deep reinforcement learning model for video captioning with fine-grained action description. Video frames are first processed by a convolutional neural network (CNN) for feature extraction and then fed into a set of LSTM layers for context encoding. A manager-worker module is designed to do video captioning at both a higher-level for goal-setting and a lower-level for actual word generation. Rohrbach *et al.* [21] constructed a description generation model that jointly localizes the subjects in the video clip based on the relationship between the visual appearance and the text description in a semi-supervised way.

2.3 Video question answering

Several algorithms for VQA tasks have recently been proposed. Zhao *et al.* [35] proposed an encoder-decoder based framework that learns temporal features via gated recurrent units [4]. A hierarchical reasoning process was applied for progressive understanding of the video content. The algorithm was evaluated using questions and answers built from the TGIF dataset [19] consisting of 200K GIFs and corresponding descriptions. Lei *et al.* constructed the TVQA dataset [18] from multiple TV shows and proposed a multi-stream Recurrent Neural Network (RNN) with a context matching module to jointly model the question, answer and data. Wang *et al.* [26] constructed frame-level representations using regional features extracted by CNNs and generated clip-level content on top of the frame representations. Ye *et al.* [30] learned video features with an LSTM and augmented them using detected key objects appearing in each frame such as "dog" or "plate" to obtain a finer

video content representation for answering questions constructed on the YouTube2Text dataset [10].

3 THE SOCIAL-IQ DATASET

Zadeh et al. [31] constructed the Social-IQ dataset from Youtube videos. The dataset contains 1250 videos with durations of 30 to 60 seconds. 1015 videos have been publicly released with 888 designated as a training set and the remaining 127 designated as the testing set. The additional 235 videos have been retained by the original authors for future use and are not publicly available.

The topics and environmental settings of Social-IQ include political debates, outdoor entertainment, video blogs about daily life, talk shows, and movie clips. The videos, audio sound signals, and subtitles (transcripts) are provided in the dataset. For each video, 6 different questions are given with lengths ranging from 5 to 25 words, asking about topics including the feelings of the people involved, their attitudes towards a person or an event and the manner of expression for such attitudes, the personalities of the participants, and the relationships and social statuses in a group of people. Most questions start with "Why", "How", "What", or "Does"; several example questions in different categories are shown in Table 1.

Table 1: Example of the questions in different categories in the Social-IQ dataset [31].

Categories	Example question
Reaction interpretation	Why doesn't the woman want to eat any more food?
Attitude	What is the woman's attitude towards her grandmother?
Agreement	Are the two men in agreement?
Feeling	How confident was the woman in the mint suit during her speech?
Manner of expression	How did the people seated on the blue chairs react to the questions the woman in the mint suit asked?
Atmosphere	Do the men appear to get along?

For each question, there are 4 correct answers and 3 incorrect answers. These can be combined into 12 different correct/incorrect answer pairs resulting in 72 different question/candidate answer sets for each video. During the experiment, given a video, a question, and an answer pair with a correct and incorrect response in an arbitrary order, the objective is to select the correct answer by predicting the position of it in the given answer pair.

4 APPROACH

Selecting the correct answer to the challenging questions requires a fine understanding of the details in the question and candidate answers, the comprehensive fusion of the multimodal data features, and the accurate extraction of the critical information without producing a huge model. We call the algorithm we propose and describe in this section TACO-Net, which stands for Temporal Attention and COnsistency.

The whole framework is shown in Figure 3. It consists of (1) a sequence preprocessing module for extracting features from the raw video, audio, and transcript data, (2) a temporal encoding module using an LSTM to reveal temporal dependencies of information at different positions in the sequence, (3) a temporal highlighting module to apply more weight to more important positions in the sequence, (4) a multi-level consistency measuring module to reduce the large feature map to a small number of similarity values, and (5) a multi-step reasoning module for final comprehension and decision making. The details of the specific modules in our network are described in the following sections.

4.1 Sequence preprocessing module

4.1.1 Feature extraction and alignment. We use the same preprocessed multimodal features that are provided in the Social-QA dataset [31]. In particular, the feature set includes the visual, audio and transcript components extracted by the following processes:

Visual features: The video frames are sampled at 1 frame per second and fed into a pre-trained DenseNet161 [13] model to obtain a 2208-dimensional feature vector representing the visual content of the image frame.

Audio non-verbal features: The audio signals are processed using the COVAREP toolbox [6] sampled in roughly 10ms windows to produce 74-dimensional feature vectors including rhythm features such as MFCCs and parabolic spectral parameters.

Transcript features: The transcript of each video is divided into multiple segments of around 4 seconds, and the transcript of each segment is projected to a 768-dimensional vector using the BERT word embedding model [7].

The visual and audio features are aligned to the transcript features based on averaging. For example, if L is the total number of subtitle segments, the transcript feature is a vector of dimension $L \times 768$. Given one such segment extending from time t_0 to t_1, there are $(t_1 - t_0)$ video frames extracted during this period that are averaged to produce the final visual feature. Similarly, the final audio feature is calculated as the average of all the original audio features during this time period. Therefore, after the alignment, the final visual, audio, and transcript features have the dimensions $L \times 2208$, $L \times 74$, and $L \times 768$ respectively.

Question and answer features: Similarly, the question and the candidate answers are processed using the BERT model [7] to represent the word strings as 768-dimensional vectors.

4.1.2 Temporal encoding. The input visual, audio, and transcript sequences are fed into a bi-directional LSTM [12, 23] with n_0 hidden nodes to learn the temporal dependencies in the data stream. The contextual information between the past time step $(t-1)$ (or the future time step $(t+1)$ in the backward direction) and the current time step t are then fused in the output hidden state $\overrightarrow{h}^t$ (or $\overleftarrow{h}^t$ in the backward direction). By concatenating the hidden states in the last layer of the LSTM along two directions, we obtain the encoded sequence $R = [\overrightarrow{h}^{1:L}, \overleftarrow{h}^{1:L}]$. We also extract the hidden states at the last timestamp as $M_h = [\overrightarrow{h}^L, \overleftarrow{h}^L]$. Similarly, a bi-directional LSTM with n_1 hidden nodes is designed for contextual information

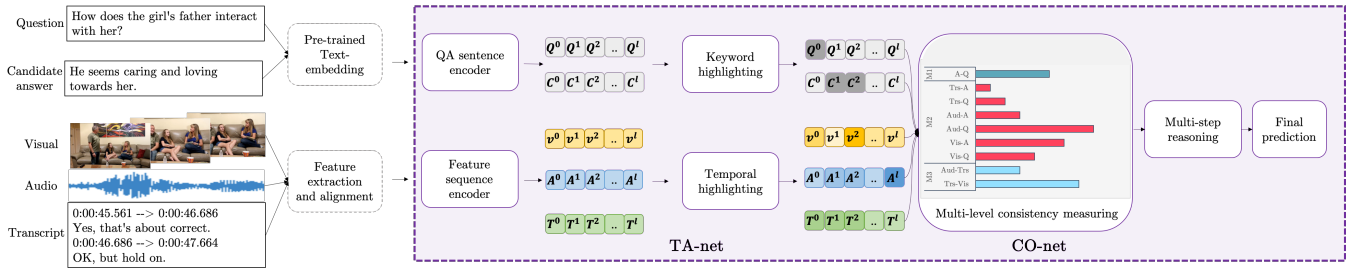


Figure 3: The framework of the proposed TACO-Net algorithm.

understanding in the question and the candidate answers. In all the experiments reported here, we used $n_0 = 150$ and $n_1 = 75$.

4.2 Temporal highlighting module

The next module consists of two pieces. The first is a temporal highlighting piece to make the model pay more attention to key regions in the multimodal sequences. The second is a content comprehension piece for finer understanding of the weighted multimodal sequence. The network details are illustrated in Figure 4.

To make the network focus more on critical moments during the interaction and key words in the question and answer text, we apply a dot-product attention mechanism, in which weights are calculated indicating the importance level of different parts. We first merge the hidden states in the forward direction and the backward direction and then calculate the weight as the similarity between the encoded sequence after LSTM and the merged hidden states. Specifically, the weights of different time steps in the input sequences are calculated as:

$$W = \text{softmax}(R^T M_h) \tag{1}$$

where R denotes the output of the bi-directional LSTM, M_h represents the concatenated hidden states, and the softmax function is used to normalize the weights.

We then apply the weights to different positions of the encoded sequence, and calculate the highlighted sequence as

$$S_h = \begin{bmatrix} R^0 W^0 & R^1 W^1 & \dots & R^L W^L \end{bmatrix} \tag{2}$$

An example of this automatic highlighting of the QA pair and the multimodal sequences is illustrated in Figure 5 in Section 5.4.

The highlighted multimodal data streams are fed into an additive-aggregation layer for dimension reduction resulting in a multimodal context vector with dimension 300 and a QA context vector with dimension 150. As multimodal material contains more complicated information that is hidden and requires deeper understanding, the multimodal context vectors are then fed into an extra set of fully connected layers to generate a final representation with dimension 150. A ReLU activation function is applied at each layer to increase the nonlinearity of the network.

4.3 Multi-level consistency measuring module

The second key innovation of our algorithm is a consistency measuring module that considers three aspects:

- **M1.** Are the question and the candidate answer consistent with each other?
- **M2.** Are the multimodal features consistent with the question/answer pair?
- **M3.** Are the multimodal features self-consistent?

In this module, we measure the consistency scores for the QA pair and between the multimodal feature sets. We denote the intermediate encoded feature vectors from the visual, audio, and transcript data streams as $\kappa_{vis}, \kappa_{aud}, \kappa_{trs}$, respectively, and the feature vectors from the question and the candidate answer as κ_q and κ_a. Then Table 2 maps the three consistency measures to pairs of features. We use ϕ to denote the function for consistency score measurement.

Table 2: Three-level consistency measuring.

M1	M2	M3
$\phi(\kappa_q, \kappa_a)$	$\phi(\kappa_q, \kappa_{vis})$ $\phi(\kappa_q, \kappa_{trs})$ $\phi(\kappa_q, \kappa_{aud})$ $\phi(\kappa_a, \kappa_{vis})$ $\phi(\kappa_a, \kappa_{trs})$ $\phi(\kappa_a, \kappa_{aud})$	$\phi(\kappa_{trs}, \kappa_{vis})$ $\phi(\kappa_{aud}, \kappa_{vis})$

In our approach, we select ϕ to be the cosine similarity score:

$$\phi(\alpha, \beta) = \text{sim}(\alpha, \beta) = \frac{\alpha \cdot \beta}{\|\alpha\| \cdot \|\beta\|} \tag{3}$$

In this way, the large intermediate feature maps from the multimodal data are reduced to a vector φ of 9 consistency scores that capture the relationships between the multimodal data, the question, and the candidate answer, which is the input to the final reasoning stage discussed next. We note that while **M3** could also include the similarity between the audio data and the transcript data, we found this not to improve performance in our experiments.

4.4 Multi-level reasoning module

We then feed the similarity scores φ into a multi-level reasoning module that consists of a set of fully connected layers. A ReLU layer is applied after each fully connected layer for increased nonlinearity. In the experiments, we found that 4 fully connected layers with 30 nodes at each layer followed by a ReLU activation function and a dropout layer achieved the best performance. This module generates a scalar regression value representing the final joint consistency measurement of the input candidate answer with the given question and the multimodal materials. The final decision is made by comparing the regressed values corresponding to the two candidate answers fed into the network.

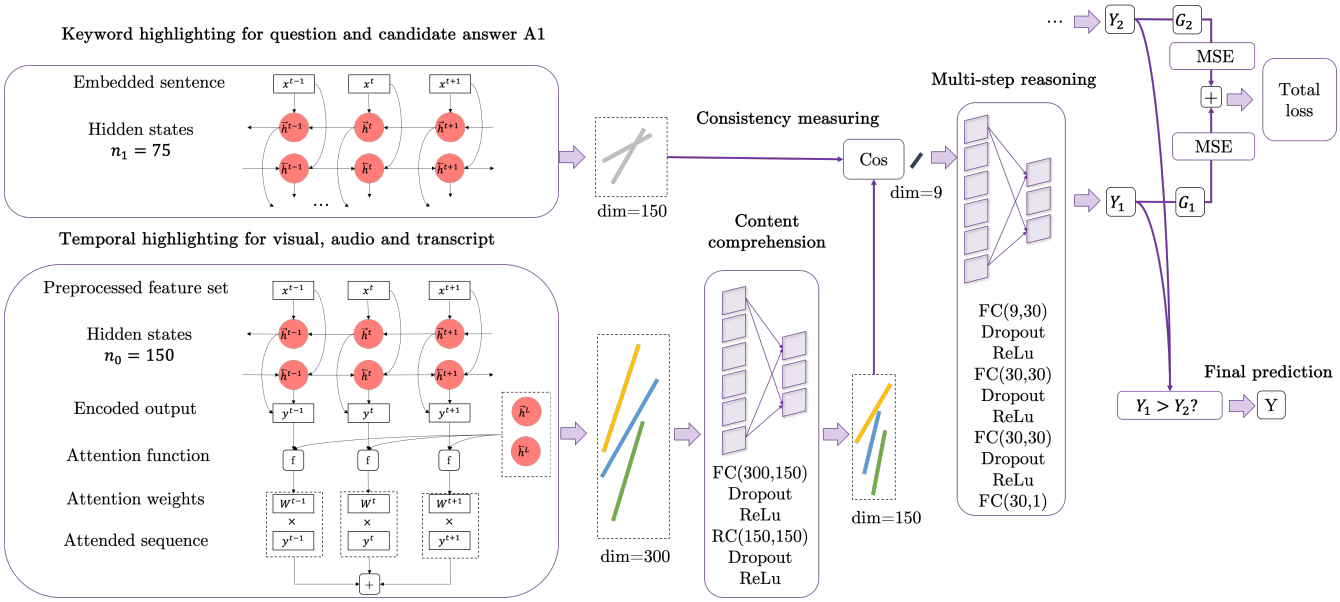


Figure 4: The network details of the proposed TACO-Net.

4.5 Training process

For each training sample, we are given the question Q, candidate answers A_1 and A_2, and the multimodal data X. Following the procedure in [31], we perform two-step training. Assuming A_1 is the ground-truth correct answer:

$$\begin{aligned} Y_1 &= \Theta(Q, A_1, X) \\ Y_2 &= \Theta(Q, A_2, X) \end{aligned} \tag{4}$$

where Θ represents the TACO-Net model, and Y_1, Y_2 denote the predictions. Since A_1 is correct, the ground truth values are $G_1 = 1, G_2 = 0$, and we use the mean-squared loss during the training process:

$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N}\left((\sigma(Y_1^i) - G_1^i)^2 + (\sigma(Y_2^i) - G_2^i)^2\right) \tag{5}$$

where N is the batch size, i represents the index of the training sample and σ represents the sigmoid function.

After training, for each testing sample $\{Q, A_1, A_2, X\}$, the final prediction value Y is

$$Y = \begin{cases} 1 & \text{if } Y_1 > Y_2 \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

where Y_1 and Y_2 are outputs from the model in (4).

5 EXPERIMENTS AND DISCUSSION

5.1 Implementation details

We used the Adam optimizer [17] for training the algorithm. During training, we set the initial learning rate to be 0.001, the batch size to be 32, and the maximum number of epochs to be 60. It takes about 1 hour for the network to converge when using an Nvidia Quadro M4000 to train. During the two-step training, after alternately feeding one correct candidate answer and one incorrect candidate answer, the network only performs one update using one joint loss function. During testing, the network has no prior knowledge about the position of the correct answer.

5.2 Qualitative analysis

To verify that the trained model can correctly highlight the important parts in the QA pair as well as in the multimodal data, we extracted the highlighting weights of the trained model and visualize one example in Figure 5. The lightness of the colors in the cell and the numbers adjacent to the text/images indicate the weight at each position (darker cells = higher weights).

In this example, the video content is a conversation about trying a makeup product, where the camera switches between closeup and wide views of several pairs of people. As shown in Figure 5, in the question, the key phrases related to the main subjects "woman wearing a denim shirt" and "blond woman" are automatically highlighted with largest weights. Additionally, the words related to the critical moment "starts feeling pain" are marked as important by the highlighting module. In the two candidate answers, the key distinctive words "shocked" and "unconcerned" are highlighted with the largest weights, making the model emphasize the different attitudes in the two answers. Additionally, in the transcript, we can see that the critical moments in which the two women are talking about pain-related feelings such as "starting to itch me", "cause my skin..." and "I'll be fine" are highlighted.

In the visual information, the temporal windows that contains the "shocked" and "worried" facial expressions of the woman in denim are correctly highlighted. We can see that there are more than two people shown in the video and TACO-Net correctly identifies the querying subject **woman in denim** among the multiple people, demonstrating the effectiveness of the model.

Figure 5: An example visualization of automatic temporal highlighting results for both a QA pair and a video.

We then consider the consistency measurements corresponding to the two different candidate answers generated by the trained model. According to Table 3, we can see that the consistency scores between candidate answer A_1 and the question and the multimodal data are higher than those for candidate answer A_2, demonstrating the effectiveness of the consistency measuring module.

Table 3: An example visualization of consistency measuring results.

Candidate answer	A-Q	A-Vis	A-Trs	A-Aud
A1	0.1644	-0.0185	-0.3852	-0.0210
A2	0.0054	-0.0461	-0.4555	-0.0355

5.2.1 Case Analysis. Figure 6 shows an example of a success case where TACO-Net makes the right decisions on all 12 correct/incorrect answer pairs for the given video and question. The video contains three people being interviewed about a movie. The question involves the attitude of the man in white towards the costumes. There is a large portion of the video that shows a closeup of the query subject (man in white), providing visual cues including the facial expression (smile) and the action (laughter). Additionally, phrases in the transcript such as "I'll go with it" convey the consistent meaning of the answers $A1, A2, A3, A4$ as well as the opposite meaning of the answers $A5, A6$, making it easier for the network to sense the correct signals from the multimodal input. In contrast, Figure 7

Did the man in white like the costumes?
A1: He thought that the costumes fit perfectly into the movie.
A2: He accepted them, as they made sense given the movie.
A3: He didn't mind them despite them being cheesy.
A4: He tolerated the costumes.
A5: He disliked the costumes.
A6: He didn't like the costumes, as they were too cheesy for his taste.
A7: He loved the costumes because of how well-done and professional they looked.
Performance:
(A1, A5) √ (A2, A5) √ (A3, A5) √ (A4, A5) √
(A1, A6) √ (A2, A6) √ (A3, A6) √ (A4, A6) √
(A1, A7) √ (A2, A7) √ (A3, A7) √ (A4, A7) √
Total: 12/12 √

Figure 6: A success case in which all 12 correct/incorrect pairs were decided correctly.

gives an example where TACO-Net makes 6 correct decisions out of the 12 possible answer pairs, which is no better than random guessing. The video contains a discussion between hosts of a TV show talking about the reconciliation of a couple, opining that communication and honesty are important. The goal of the question is to select the correct description of the conversation. Some important clues can be read from the multimodal data streams including facial expression (smile) and audio tone feature (quiet) that don't match with a debate, and actions (head nodding) and gaze activity (looking at each other) that show signs of agreement. In this case, the model selects the correct answer when it is paired with $A5$, which is understandable since $A5$ contains the words "arguing" and "debate" that are not consistent with the meaning conveyed by the multimodal data. On the other hand, $A7$ can be hard to judge since the conversation shows signs of an "open" and "friendly" atmosphere and the two people stay in one place without anyone else involved, which is prone to misinterpretation as a "private" discussion.

Describe the nature of the conversation between the man and woman.
A1: They seem to be agreeing with each other, and are promoting a unified opinion.
A2: The conversation is formal in nature.
A3: The conversation is educational in nature.
A4: The conversation is scripted and professional in nature.
A5: They seem to be arguing with each other, and are having a heated debate.
A6: The conversation between the man and woman is private and intimate in nature evidenced by the man and woman's body language.
A7: The man and woman are having a open and friendly conversation judging by their relaxed tone of voice.
Performance:
(A1, A5) √ (A2, A5) √ (A3, A5) √ (A4, A5) √
(A1, A6) (A2, A6) (A3, A6) √ (A4, A6)
(A1, A7) (A2, A7) (A3, A7) √ (A4, A7)
Total: 6/12 √

Figure 7: A failure case where the algorithm decides 6 of 12 answer pairs correctly.

5.3 Quantitative performance

We compare our model with the state-of-the-art VQA model on the Social-QA dataset proposed by Zadeh et al. [31]. We note that [31] already compares itself against 4 other high-performing algorithms on the same dataset as ours and achieves the best performance, so we do not duplicate those figures here.

As discussed in Section 3, we report experimental results on the 888 training videos and 127 testing videos designated as the split in [31]. Thus, the training set and the testing set are exactly the same for direct comparison. Table 4 reports the testing performance figures, which are the average of the 127×6×12 possible videos and QA pairs. Our approach surpasses Tensor-MFN by nearly 3 percent, demonstrating the effectiveness of our method.

Table 4: Accuracy on the testing set of Social-IQ.

Model	Testing accuracy (%)
Tensor-MFN [31]	65.73
TACO-Net (Ours)	**68.19**

We further ran our best-performing model on a multiple-choice task. In particular, the goal is to select the single correct answer from among 3 additional incorrect answers. Assuming A_1 is correct and A_2, A_3, A_4 are incorrect, the method discussed in Section 4.5 can be straightforwardly extended to the 4-way case, computing

$$Y_i = \Theta(Q, A_i, X) \tag{7}$$

and selecting the correct answer as the index

$$i^* = \arg\max_i Y_i \tag{8}$$

Our model achieves 49.08% accuracy on the multiple-choice task over the 127-video Social-IQ testing dataset, much better than random chance. Unfortunately, no published work reports comparison figures on exactly this task and dataset. However, Zadeh et al. reported performance of 34.14% on the same task over their sequestered 235-video portion of the Social-IQ dataset, which we assume has similar characteristics to the publicly available training and testing data. In this case, our model substantially outperforms the state of the art.

5.4 Ablation study

To investigate the effectiveness of our key network mechanisms, we compare our full model with several baseline models with different pieces removed or replaced. These models include:

TA-Net: The network without consistency measuring. We remove all the consistency measuring parts and directly concatenate the intermediate feature vectors into the final reasoning module. The number of nodes in the fully connected layers of the final reasoning module is slightly adjusted according to the different input size.

CO-Net: The network without temporal attention. We remove the temporal attention weighting on the multimodal data and the question/answer pair to check whether this mechanism is essential for successfully modeling the interactive behaviors. All the data sequences after the bi-directional LSTM temporal encoder are directly fed into the consistency measuring module for a final prediction.

TACO-M1/M3: TACO-Net with M2 and one of either M1 or M3. The goal is to determine whether both self-consistency measures are required or if only one is necessary.

TACO-P1: The network without the content comprehension part in the consistency measuring module (Section 4.2.2).

TACO-RS: TACO-Net with single-step reasoning for the final prediction. Only a single layer is present in the final reasoning module to check the effectiveness of the progressive reasoning from the multiple layers.

TACO-CLASS: Instead of two-step training using a regression-based framework with mean-squared loss, we investigate the performance of the classification-based VQA framework used in [1, 9]. We slightly modify the network structure to take two candidate answers at once. Specifically, we construct two branches for the two candidate answers A_1 and A_2. Denoting c_1 and c_2 as the intermediate vector before the last layer of the original multi-step reasoning module for A_1 and A_2, we merge c_1 and c_2 and add a fully connected layer $\tilde{p}$ for the final binary prediction.

$$\tilde{Y} = \tilde{p}\,(c_1, c_2) \tag{9}$$

Denoting $\tilde{G}$ as the ground truth, we use the cross-entropy loss during training:

$$\tilde{\mathcal{L}} = \frac{1}{N}\sum_{i=1}^{N}\left(-\tilde{G}^i \log\sigma(\tilde{Y}^i) + (1-\tilde{G}^i)\log(1-\sigma(\tilde{Y}^i))\right) \tag{10}$$

Table 5 reports the comparison results of the various baseline models on the Social-IQ testing set. We can see that the two key pieces (temporal attention and consistency measuring) are essential to the model's success, and that the other mechanisms improve performance to a lesser degree.

Table 5: Comparing the full model against the baseline models with key mechanisms removed.

Model	Accuracy on testing dataset
TA-Net	54.08%
CO-Net	53.54%
TACO-P1	65.10%
TACO-RS	67.20%
TACO-MS3	67.05%
TACO-MS1	66.61%
TACO-CLASS	66.89%
Full model	**68.19%**

6 CONCLUSIONS AND FUTURE WORK

We demonstrated the success of combining temporal attention and consistency measuring for the visual question answering task on the challenging Social-IQ dataset. Currently, the pre-processed visual feature set is directly extracted from the intermediate feature map in a pre-trained CNN without any semantic information about people or objects. One possible future direction is to process the image frame with models such as Mask-RCNN [11] to include semantic labels for the interacting subjects. In this way, the environmental context could be captured and used to distinguish different scenarios. For example, the mood, relationship and atmosphere of a group of people are very different depending on whether the scenario is a happy hour in a bar or a formal meeting in a conference room. This environmental context could aid in social signal processing algorithms by providing additional evidence for the final judgement.

We currently measure consistency between the question and the multimodal feature set based on the entire video. In the future, we plan to measure *dynamic* consistency scores across the timeline during the interaction and use the changes in the consistency scores to locate the key moments that are used to answer the question. In this way, the model can not only select the correct answer but also automatically demonstrate its reasoning process and the corresponding supporting materials for why it chooses the specific answer.

ACKNOWLEDGMENTS

This work was supported by the US National Science Foundation under award IIP-1631674 from the PFI:BIC program.

REFERENCES

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*. 2425–2433.

[2] Cigdem Beyan, Francesca Capozzi, Cristina Becchio, and Vittorio Murino. 2017. Prediction of the leadership style of an emergent leader using audio and visual nonverbal features. *IEEE Transactions on Multimedia* 20, 2 (2017), 441–456.

[3] Jose M Chaquet, Enrique J Carmona, and Antonio Fernández-Caballero. 2013. A survey of video datasets for human action and activity recognition. *Computer Vision and Image Understanding* 117, 6 (2013), 633–659.

[4] Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Doha, Qatar, 1724–1734. https://doi.org/10.3115/v1/D14-1179

[5] Pradipto Das, Chenliang Xu, Richard F Doell, and Jason J Corso. 2013. A thousand frames in just a few words: Lingual description of videos through latent topics and sparse object stitching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2634–2641.

[6] Gilles Degottex, John Kane, Thomas Drugman, Tuomo Raitio, and Stefan Scherer. 2014. COVAREP—A collaborative voice analysis repository for speech technologies. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 960–964.

[7] 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).

[8] Lucca Eloy, Angela EB Stewart, Mary Jean Amon, Caroline Reinhardt, Amanda Michaels, Chen Sun, Valerie Shute, Nicholas D Duran, and Sidney D'Mello. 2019. Modeling Team-level Multimodal Dynamics during Multiparty Collaboration. In *2019 International Conference on Multimodal Interaction*. 244–258.

[9] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6904–6913.

[10] Sergio Guadarrama, Niveda Krishnamoorthy, Girish Malkarnenkar, Subhashini Venugopalan, Raymond Mooney, Trevor Darrell, and Kate Saenko. 2013. Youtube2text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition. In *Proceedings of the IEEE International Conference on Computer Vision*. 2712–2719.

[11] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. Mask R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision*. 2961–2969.

[12] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation* 9, 8 (1997), 1735–1780.

[13] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4700–4708.

[14] Mohamed E Hussein, Marwan Torki, Mohammad A Gowayyed, and Motaz El-Saban. 2013. Human action recognition using a temporal hierarchy of covariance descriptors on 3d joint locations. In *Twenty-Third International Joint Conference on Artificial Intelligence*.

[15] Yu-Sian Jiang, Garrett Warnell, and Peter Stone. 2018. Inferring user intention using gaze in vehicles. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*. 298–306.

[16] Ahmet Alp Kindiroglu, Lale Akarun, and Oya Aran. 2017. Multi-domain and multi-task prediction of extraversion and leadership from meeting videos. *EURASIP Journal on Image and Video Processing* 2017, 1 (2017), 77.

[17] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[18] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L. Berg. 2018. TVQA: Localized, Compositional Video Question Answering. In *EMNLP*.

[19] Yuncheng Li, Yale Song, Liangliang Cao, Joel Tetreault, Larry Goldberg, Alejandro Jaimes, and Jiebo Luo. 2016. TGIF: A new dataset and benchmark on animated GIF description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4641–4650.

[20] Shogo Okada, Oya Aran, and Daniel Gatica-Perez. 2015. Personality trait classification via co-occurrent multiparty multimodal event discovery. In *Proceedings of the 2015 ACM International Conference on Multimodal Interaction*. ACM, 15–22.

[21] Anna Rohrbach, Marcus Rohrbach, Siyu Tang, Seong Joon Oh, and Bernt Schiele. 2017. Generating descriptions with grounded and co-referenced people. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4979–4989.

[22] Dairazalia Sanchez-Cortes, Oya Aran, and Daniel Gatica-Perez. 2011. An audio visual corpus for emergent leader analysis. In *Workshop on Multimodal Corpora for Machine Learning: Taking Stock and Road Mapping the Future, ICMI-MLMI*.

[23] Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing* 45, 11 (1997), 2673–2681.

[24] Angela EB Stewart, Zachary A Keirn, and Sidney K D'Mello. 2018. Multimodal modeling of coordination and coregulation patterns in speech rate during triadic collaborative problem solving. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*. 21–30.

[25] T Váradi, Gy Kovács, I Szekrényes, H Kiss, and K Takács. [n.d.]. Human-human, human-machine communication: on the HuComTech multimodal corpus. In *CLARIN Annual Conference 2018*. 56.

[26] Bo Wang, Youjiang Xu, Yahong Han, and Richang Hong. 2018. Movie question answering: Remembering the textual cues for layered visual contents. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

[27] Xin Wang, Wenhu Chen, Jiawei Wu, Yuan-Fang Wang, and William Yang Wang. 2018. Video captioning via hierarchical reinforcement learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4213–4222.

[28] Suowei Wu, Zhengyin Du, Weixin Li, Di Huang, and Yunhong Wang. 2019. Continuous Emotion Recognition in Videos by Fusing Facial Expression, Head Pose and Eye Gaze. In *2019 International Conference on Multimodal Interaction*. 40–48.

[29] Huijuan Xu, Boyang Li, Vasili Ramanishka, Leonid Sigal, and Kate Saenko. 2019. Joint event detection and description in continuous video streams. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 396–405.

[30] Yunan Ye, Zhou Zhao, Yimeng Li, Long Chen, Jun Xiao, and Yueting Zhuang. 2017. Video question answering via attribute-augmented attention network learning. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 829–832.

[31] Amir Zadeh, Michael Chan, Paul Pu Liang, Edmund Tong, and Louis-Philippe Morency. 2019. Social-IQ: A question answering benchmark for artificial social intelligence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 8807–8817.

[32] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor Fusion Network for Multimodal Sentiment Analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Copenhagen, Denmark, 1103–1114. https://doi.org/10.18653/v1/D17-1115

[33] Aurélie Zara, Valérie Maffiolo, Jean Claude Martin, and Laurence Devillers. 2007. Collection and annotation of a corpus of human-human multimodal interactions: Emotion and others anthropomorphic characteristics. In *International Conference on Affective Computing and Intelligent Interaction*. Springer, 464–475.

[34] Lingyu Zhang and Richard J. Radke. 2020. A Multi-Stream Recurrent Neural Network for Social Role Detection in Multiparty Interactions. *IEEE Journal of Selected Topics in Signal Processing* (2020), 1–14.

[35] Zhou Zhao, Qifan Yang, Deng Cai, Xiaofei He, Yueting Zhuang, Zhou Zhao, Qifan Yang, Deng Cai, Xiaofei He, and Yueting Zhuang. 2017. Video Question Answering via Hierarchical Spatio-Temporal Attention Networks.. In *IJCAI*. 3518–3524.