VEGAS: Towards Visually Explainable and Grounded Artificial Social Intelligence

Hao Li[1], **Hao Fei**[2], **Zechao Hu**[1], **Zhengwei Yang**[1], **Zheng Wang**[1]*

[1] National Engineering Research Center for Multimedia Software, School of Computer Science, Wuhan University
[2] School of Computing, National University of Singapore
{sc.lihao, hzc_aim, yzw_aim, wangzwhu}@whu.edu.cn, haofei37@nus.edu.sg

Abstract

Social Intelligence Queries (Social-IQ) serve as the primary multimodal benchmark for evaluating a model's social intelligence level. While impressive multiple-choice question (MCQ) accuracy is achieved by current solutions, increasing evidence shows that they are largely, and in some cases entirely, dependent on language modality, overlooking visual context. Additionally, the closed-set nature further prevents the exploration of whether and to what extent the reasoning path behind selection is correct. To address these limitations, we propose the Visually Explainable and Grounded Artificial Social Intelligence (**VEGAS**) model. As a generative multimodal model, VEGAS leverages open-ended answering to provide explainable responses, which enhances the clarity and evaluation of reasoning paths. To enable visually grounded answering, we propose a novel sampling strategy to provide the model with more relevant visual frames. We then enhance the model's interpretation of these frames through Generalist Instruction Fine-Tuning (GIFT), which aims to: i) learn multimodal-language transformations for fundamental emotional social traits, and ii) establish multimodal joint reasoning capabilities. Extensive experiments, comprising modality ablation, open-ended assessments, and supervised MCQ evaluations, consistently show that VEGAS effectively utilizes visual information in reasoning to produce correct and also credible answers. We expect this work to offer a new perspective on Social-IQ and advance the development of human-like social AI.

Code — https://github.com/lihao921/VEGAS

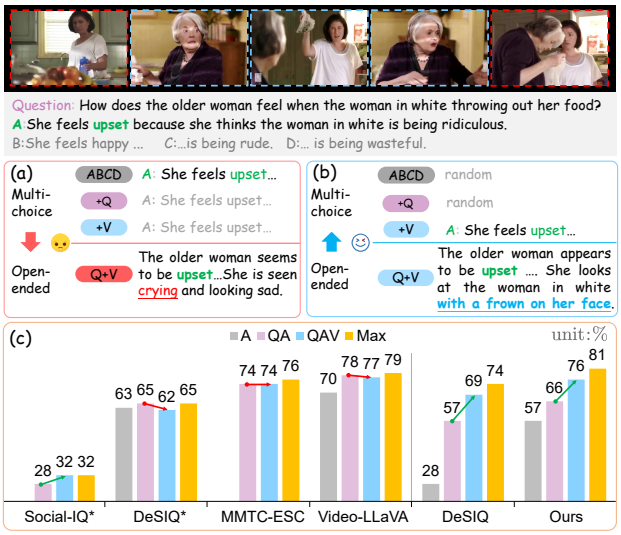


Figure 1: (a) An existing approach selects the correct option without knowing the question or even the video context, revealing incorrect rationale in the open-ended answers. (b) Our study begins with a correct reasoning path grounded in the video, ensuring reliable selection. (c) Our model enhances visual engagement, reduces the language shortcut, and achieves comparable but more reliable MCQ accuracy. * denotes the baseline of the corresponding method.

Introduction

Human-level AI necessitates not only formidable reasoning abilities at the individual level, but also must exhibit social intelligence akin to that of humans in social contexts. This motivates the study of the topic of Artificial Social Intelligence (ASI) (Bainbridge et al. 1994; Dautenhahn 2007), which aims to advance machine comprehension and interaction within complex social contexts. The Social-IQ benchmark (Zadeh et al. 2019) exemplifies this challenge through a multiple-choice question (MCQ) task, which allows the integration of video, audio, and subtitles to test models' reasoning ability regarding social dynamics. This benchmark is a valuable resource of ASI, as it mirrors the complexity of real-world social scenarios.

Social-IQ posed significant challenges for advanced VideoQA methods when it was first introduced (Zadeh et al. 2017; Lei et al. 2018; Liang et al. 2018): relying solely on the question and answer (BlindQA) resulted in near-random accuracy. With the ongoing advancements of Large Language Models (LLMs), recent multimodal approaches incorporating them have more than doubled this accuracy. However, as illustrated in Figure 1 (c), these improvements are largely due to the LLMs' shortcut effect, i.e., exploiting inherent spurious correlations between questions and options, while visual context is almost entirely disregarded in

*Corresponding author

the answering process. DeSIQ (Guo, Li, and Haffari 2023) even found that a T5-small language model (Raffel et al. 2020) could achieve 100% accuracy on Social-IQ-1.0 under BlindQA, highlighting a significant language bias in Social-IQ. Moreover, current solutions, whether based on Transformer (Yang et al. 2021; Pirhadi, Mirzaei, and Eetemadi 2023), LSTM (Kumar, Mittal, and Manocha 2020), contrastive learning (Wilf et al. 2023; Xie and Park 2023), or compositional models (Natu, Sural, and Sarkar 2023), typically rely on uniformly sampled frames that fail to provide question-relevant visual context, let alone capture the nuanced social traits embedded in the videos. These limitations and data biases lead models to exploit language shortcuts for selection (Niu et al. 2021; Cho et al. 2023; Lao et al. 2023).

The consequences are profound. **i)** Such models face doubts about their ability to meet Social-IQ's goal: discerning the correct option through comprehensive reasoning about multimodal social traits. **ii)** The closed-set nature of MCQ amplifies these concerns, as it restricts the exploration and assessment of whether the selected answers truly reflect the underlying reasoning. In contrast, open-ended answers provide deeper insights, as shown in Figure 1 (a), where a model reliant on shortcuts reveals its ungrounded rationale. These challenges and opportunities encourage us to develop a more transparent and reliable approach to Social-IQ.

To combat these issues, in this paper, we propose a novel **V**isually **E**xplainable and **G**rounded **A**rtificial **S**ocial Intelligence (**VEGAS**) framework for Social-IQ. Specifically, we opt for a generative approach, as seen in recent Multimodal Large Language Models (MLLMs), to enable *Explainable* open-ended responses, facilitating the probing and measuring of reasoning paths. To deliver visually *Grounded* answering, we employ a dual-pronged strategy consisting of **L**anguage **G**uided **S**ampling (**LGS**) and **G**eneralist **I**nstruction **F**ine-**T**uning (**GIFT**). **Firstly**, the LGS equips the model with the ability to sample question-relevant video frames in social interactions, guided by language cues in the form of explicit descriptions, causal questioning, and nuanced differentiation. We craft the dataset and learning strategy to enable effective LGS supervision in the absence of timestamp annotations. However, the sampled frame features often exhibit disordered temporal relationships due to the temporal embedding in the pre-trained video encoder, which is optimized for uniformly spaced frames. This leads to misinterpretations of social activities, such as *touching vs hitting*, which are crucial for understanding latent social attitudes. To address this, we propose a **T**emporal **A**ttention **M**odule (**TAM**) to restore the order in these frames, ensuring coherent sequencing without a secondary encoding. **Secondly**, the goal of GIFT is to learn an effective understanding of sampled visual features, which requires advanced abilities from the subsequent reasoning modules. To achieve that, we first integrate the **S**ocial **T**raits **P**rojector (**STP**) to learn transformation for fundamental emotional traits (video, image, and audio) into the language space. Following this, we perform joint fine-tuning of STP and LLM using an expansive multimodal social interaction dataset. This results in VEGAS-*generalist*, an enhanced version excelling in joint reasoning, enriched social commonsense, and advanced expertise in sociology and psychology.

In this study, we prioritize the correctness of the reasoning path over mere pursuit of maximum accuracy. We uncover and assess the reasoning of open-ended answers using ChatGPT. The evident accuracy improvement of VEGAS highlights the consistency between its reasoning and the correct selection, cf. Figure 1 (b). Furthermore, modality ablation in MCQ reveals that VEGAS significantly suppresses language shortcuts, improving visual context utilization from -0.43% to 9.28%. Finally, VEGAS achieves state-of-the-art performance with a more credible and scalable implementation. Contributions of this paper are summarized as follows:

- We for the first time introduce VEGAS, a visually explainable and grounded social intelligence model that mitigates language shortcut effect in Social-IQ and efficiently considers visual context answering.
- We propose a dual approach to enhance the relevance of video frames and improve their interpretation, ensuring accurate and visually grounded answers.
- We introduce VEGAS-*generalist*, a sophisticated human-like social AI that demonstrates profound understanding and analytical expertise in social dynamics.

Related Work

Social-IQ

The Social-IQ-1.0 benchmark (Zadeh et al. 2019) was introduced in 2019 to evaluate the social intelligence level of AI models with MCQ task. Social-IQ-2.0 soon updates the benchmark with newly annotated questions and answers. As solutions, (Natu, Sural, and Sarkar 2023) incorporate external knowledge retrieved from VisualCOMET (Park et al. 2020) to augment the multimodal features with social commonsense. MMTC-ESC (Xie and Park 2023) leverages contrastive learning with emotional cues to build cross-modal correlations of features. Just Ask Plus (Pirhadi, Mirzaei, and Eetemadi 2023) uses multi-headed attention and transformer encoders to compute representations for the questions and answers, then calculates their similarity for selection. F2F-CL (Wilf et al. 2023) conducts fine-grained graph contrastive learning by decomposing the social interaction according to speaking turns. Moreover, Social-IQ is also a popular benchmark in many generic video understanding models (Li et al. 2024; Xu et al. 2023; Fei et al. 2024b).

Despite ongoing efforts, few studies have addressed the language shortcut issue in Social-IQ. A model-side solution (Gat et al. 2020) once proposed to use loss regularization for generic classifier debiasing, achieving a 2% improvement on Social-IQ but lacked further analysis. DeSIQ (Guo, Li, and Haffari 2023) is the only approach so far addressing language biases by empirically substituting incorrect answers with correct ones from other samples. In contrast, our model-side approach directly enhances visual information usage as effectively as DeSIQ but with higher accuracy.

Multimodal Large Language Models

Recent advancements in MLLMs have significantly enhanced video question answering (VideoQA) (Chen and

Dolan 2011; Yu et al. 2019; Fei et al. 2024a,d). These models effectively integrate various modalities (Yu, Yoon, and Bansal 2024; Wu et al. 2024a; Zhu et al. 2023; Yu, Yoon, and Bansal 2024; Wu et al. 2024b; Fei et al. 2024c), such as audio, video, and depth, by projecting features from frozen encoders into the language space, leveraging them to produce natural language responses. Video-LLaMA (Zhang, Li, and Bing 2023) integrates visual and audio features from frozen encoders using Q-Former (Li et al. 2023). Video-ChatGPT (Maaz et al. 2023) uses liner layers to project temporal and spatial features extracted from videos to the LLM, and generate conversations accordingly. PG-Video-LLaVA (Munasinghe et al. 2023) strengthens the MLLM with pixel level grounding ability by introducing grounding modules like scene detector and object tracker. Video-LLaVA (Lin et al. 2023) incorporates visual encoders pre-aligned with language for unified understanding.

Although appealing, using them for social intelligence is non-trivial, as they default to uniformly sampled frames, missing critical visual details for Social-IQ. Recent generic models that retrieve relevant video segments (Xu et al. 2023) or frames (Li et al. 2024) have shown some improvements on Social-IQ, but the absence of targeted social designs limits their abilities. Despite this, the generative pipeline holds potential for bridging the gap between current research and human-like social AI (Chandra, Shirish, and Srivastava 2022; Duéñez-Guzmán et al. 2023; Liu et al. 2025).

Methodology

VEGAS Framework

As shown in Figure 2, the VEGAS framework integrates video, image, and audio encoders from LanguageBind (Zhu et al. 2023) to process inputs of various modalities, along with a word embedding layer for text encoding. While the video encoder is primarily used for the Social-IQ task, we include image and audio encoders for better scalability and general applicability.

First, all n candidate frames are encoded by the video encoder. In the LGS, the sampler selects k frames that align with the language hint—either a question (for inference) or its fusion with an answer (for training). The TAM then restores the temporal relationships among the sampled frames. To connect the encoders with the LLM, we use linear layers (Lin et al. 2023) to build the Social Traits Projector (STP), which learns transformations of modality social traits. Based on the multimodal features and word embeddings, the LLM generates either free-form text or selected options as per user instructions. During the GIFT stage, we fine-tune the STP along with the LLM, resulting in VEGAS-*generalist*.

Language Guided Sampling (LGS)

Sampler Structure. Let V represent the sequence of n uniformly sampled video frames, and Q denote the language hint consisting of m words. We encode these inputs using the video and text encoders, producing f_{v0} for the visual frames and f_q for the text. Here, f_{v0} is computed with pre-trained temporal embeddings, where each frame is associated with a `CLS` token that encapsulates its global visual feature. To

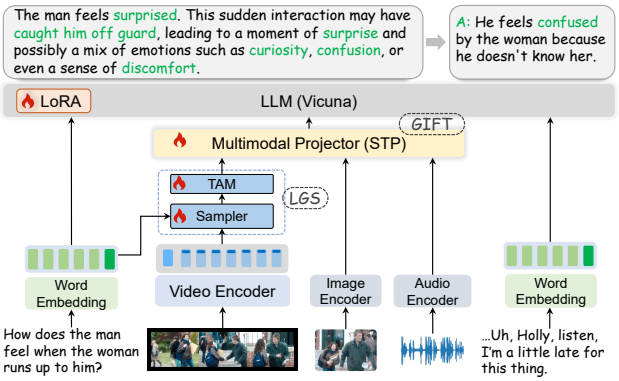


Figure 2: Architecture of VEGAS. The system encodes multimodal inputs with frozen encoders. These inputs are projected into LLM space using a trainable Multimodal Projector, enabling nuanced answer generation that captures social attitudes in interactions like emotions.

explore the causal relevance between V and Q, we compute a contextualized representation for the visual features using the ATP transformer block (Buch et al. 2022) as

$$f_{qv} = \text{ATP}(\text{Linear\_v}(f_{v0}^{\text{CLS}}), \text{Linear\_q}(f_q)). \quad (1)$$

Here, f_{v0}^{CLS} denotes the `CLS` token for each frame. Based f_{qv}, we calculate the relevance logits as:

$$logits = \text{Linear\_qv}(f_{qv}) \in \mathbb{R}^{n \times d}, \quad (2)$$

where d is the dimension of visual feature. Then, we select the top-k frames as

$$f_v = \text{Top-K}(logits, f_{v0}) \in \mathbb{R}^{k \times d}. \quad (3)$$

The Top-K function is implemented to be differentiable by introducing stochastic perturbations during training (Cordonnier et al. 2021) for optimization. At inference, we directly select the indices of $logits$ with the top-k values as key frame indicators.

Temporal Attention Module (TAM). The selected k frame features from Eq. (3) are encoded with temporal embedding T_v^k designed for k frames (typically $k = 8$). In such a process, encoding n frames ($n > k$) requires repeating T_v^k for n/k times, which disrupts temporal coherence. We introduce the TAM to restore the order of sampler output f_v. First, the f_v from Eq. (3), as illustrated in Figure 3, can be broken down as

$$f_v = f_v^k + T_v^k. \quad (4)$$

We additionally learn a new temporal embedding T_s^k for sampled frames, which is defined and initialized as

$$T_s^k \in \mathbb{R}^{k \times d} \sim \mathcal{N}(0, \frac{1}{\sqrt{d}}). \quad (5)$$

Then, we construct new relationships using a CLIP attention module (Zhu et al. 2023) as

$$f_{v1} = f_v + T_s^k, \quad (6)$$
$$f_{v2} = \text{CLIPAtt}(\text{LayerNorm}(f_{v1})), \quad (7)$$
$$f_{v3} = f_{v1} + f_{v2}. \quad (8)$$

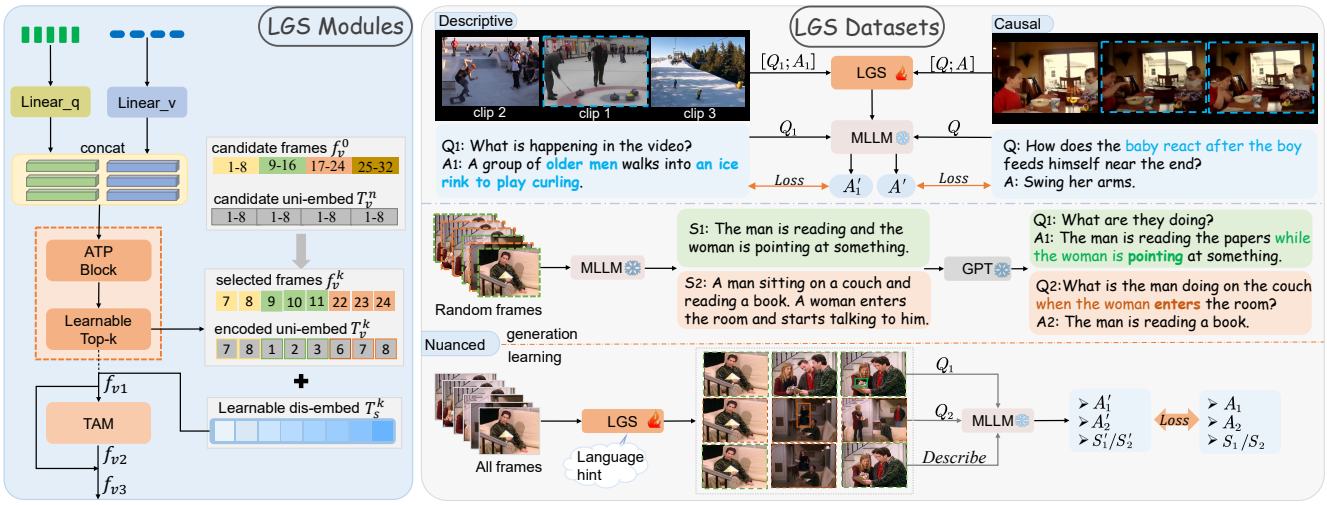


Figure 3: Left: The proposed Language Guided Sampling modules. Right: Three datasets are crafted for LGS training, incorporating descriptive, causal, and nuanced language cues.

Here, f_{v3} is the final output of LGS and then used as input of STP. We use residual connections considering uniform frames may still be needed in tasks like video captioning.

Sampling Data Construction. We craft targeted data for LGS learning due to the lack of timestamp annotations. We aim for three kinds of sampling ability as illustrated in Figure 3. **Firstly**, the basic localization ability based on *explicit description*. We create a composite video V by randomly selecting, shuffling, and merging three video clips from the Video-ChatGPT dataset (Maaz et al. 2023). The training sample is then defined as (V, Q_1, A_1), where the QA pair from clip1 is used for video captioning training. This setup compels the sampler to accurately locate frames in clip1 when provided with the language hints Q_1 and A_1. **Secondly**, the ability to capture frames that have *causal relations* with language hints. For this, we employ the QA samples from the temporal subset of Next-QA (Xiao et al. 2021).

Thirdly, the ability to distinguish key frames based on *nuanced details* in less varying social interactions.We use the TVQA dataset (Lei et al. 2018) for its rich social dynamics. We use a self-refinement strategy to enforce the model to recall and locate these dynamics according to its own memory. For each video, we begin by creating pseudo samples through two rounds of random sampling ($i \in \{1, 2\}$), each selecting k frames from the video. We then use the baseline model to generate captions S_1 and S_2 for the two clips, which serve as the memory. Next, we prompt ChatGPT to craft distinctive QA pairs based on the captions. Finally, we construct two samples of the same video as

$$\text{Sample}_i = \langle V, Q_i, A_i, S_i \rangle, i \in \{1, 2\}. \quad (9)$$

This design ensures that Q_i can only be correctly answered with A_i, with temporal moment in V captured by S_i.

Training Pipeline

LGS Training. To effectively train the LGS modules, we employ QA pairs as language hints of sampler for the descriptive and causal samples generated above. The predicted free-form answers A'_1 and A' are supervised with their ground truths. For the generated pseudo data, each sample is utilized in two ways, as depicted in Figure 3. **i)** The QA task is formulated as

$$A'_i = \text{LLM}(Q_i, \mathcal{S}(V, Q_i, A_i)), i \in \{1, 2\}, \quad (10)$$

where the sampler $\mathcal{S}(\cdot)$ generates key frame features for V conditioned on Q_i and A_i. **ii)** The video captioning task is conducted as

$$S'_i = \text{LLM}(Q, \mathcal{S}(V, S_i, A_i)), i \in \{1, 2\}, \quad (11)$$

where Q = *"Describe the video."*. The model predictions are supervised with the CrossEntropyLoss function.

GIFT 1: STP Emotion Transformation Learning. We initialize the STP module using the linear visual projector from Video-LLaVA (Lin et al. 2023), which ensures optimal alignment between visual and language modalities through captioning tasks. For social intelligence purposes, we customize STP by multimodal-language transformation learning using emotion recognition as the proxy task. The process is formulated as optimizing $p(E \mid \text{LLM}(\text{STP}(Z)))$, where Z represents the encoded multimodal features and E denotes the predicted emotional category. In this process, Mel-spectrogram features are extracted from audio tracks and encoded with a Vision Transformer (ViT). We also incorporate audio captioning data to augment the model's understanding of the ongoing audio events.

GIFT 2: STP&LLM Joint Representation Learning. We fine-tune the STP and the LLM together on an extensive multimodal social interaction dataset, which is helpful for joint reasoning and human-aligned answering. By "joint", we refer to both the model and data aspects, allowing a sample to contain multiple modalities instead of only one. Specifically, we process multimodal input <mm> concurrently by concatenating visual features with audio features (if any). For subtitles, we treat

them as an individual modality in training to differentiate the primary user instruction (question) from the dialog content. This also helps in understanding ongoing social interactions through conversations. To enhance robustness, we organize modalities of each training sample using two sequences: `<mm><subtitle><question>` and `<question><mm><subtitle>`.

Experiments

Settings

Data Details. In this study, we report results on Social-IQ-2.0 and leverage various datasets and their transformations in training as Table 1 shows. For the LGS, we craft data based on TVQA (Lei et al. 2018), Next-QA (Xiao et al. 2021), and Video-ChatGPT (Maaz et al. 2023). For the STP, we use RAVDESS (Livingstone and Russo 2018), AudioCaps (Kim et al. 2019), CMU-MOSEI (Zadeh et al. 2016), and Expression in-the-Wild (ExpW) (Zhang et al. 2018). For VEGAS-*generalist*, we integrate TVQA and CMU-MOSEI for multimodal joint training. We incorporate expert insights distilled by ChatGPT from Social-IQ data (Zadeh et al. 2019) to provide in-depth analysis. We also use a portion of Video-ChatGPT data to mitigate catastrophic forgetting. We use the original Social-IQ in MCQ experiments. The combined dataset totals approximately 240,000 samples.

Dataset	Sampling	STP	Joint
TVQA	✓		✓
Next-QA	✓		
Video-ChatGPT	✓		✓
RAVDESS		✓	
AudioCaps		✓	
CMU-MOSEI		✓	✓
ExpW		✓	
Social-IQ-*expert*&MCQ			✓
Data size	33k	165k	43k

Table 1: Datasets used at different training stages.

Training Details. For the sampler, we set $n = 32$ and $k = 8$, and encode language hints with the text encoder from CLIP ViT-B/32 (Radford et al. 2021). We initiate the LGS from scratch and train the sampling process with a learning rate of 2e-4. The STP module is pre-trained with a learning rate of 1e-6. For the joint tuning of the STP and LLM, we set their learning rates to 2e-5 and 2e-4, respectively and train for three epochs. The Vicuna-7b LLM (Chiang et al. 2023) is fine-tuned using Low-Rank Adaptation (LoRA) with $r = 128$ and $\alpha = 256$. Note that the joint tuning is performed for both VEGAS-*generalist* in open-ended QA and VEGAS in supervised MCQ, but on different datasets. All training is conducted on 4 A100 40G GPUs with a batch size of 64. All training proceeds for one epoch except for supervised MCQ, which is trained for three epochs.

Metrics. Following Video-ChatGPT (Maaz et al. 2023), we use GPT-3.5-turbo to assess Accuracy (%) and Score (1-5) for open-ended answers. Accuracy determines if the answer matches the correct option, while the Score measures how closely they are aligned. All four candidate options are included in the evaluation for rigor. For MCQ, Accuracy is calculated directly by literal matching. We use the first letter of **Q**uestion, **A**nswers, **V**ideo, and **S**ubtitles to denote each modality, respectively.

Open-Ended QA

Zero-shot Results. We start by probing the correctness of the reasoning path in the answers. Table 2 presents results comparison with strong MLLM baselines, and module ablations of our proposed designs. By zero-shot, we refer to the VEGAS model with LGS rather than the fine-tuned VEGAS-*generalist*.

The ablations in the **QV** setting demonstrate the effectiveness of the proposed sampling strategy and the temporal attention. We find that using fewer candidate frames ($n = 16$) improves accuracy but harms the consistency score. This might indicate that the generated answers only roughly align with the correct option, but fail to cover details or rationales. The TAM addresses this problem by reconstructing temporal relationships when working with more candidate frames. Interestingly, even though the VEGAS model primarily affects vision branch, the improved **QVS** metric (49.5% *vs* 51.2%) indicates that better visual features can enhance the model's understanding of the conversation in subtitles.

Model	QV		QVS	
	Score	Accuracy	Score	Accuracy
Video-LLAMA	1.6	36.1	1.7	37.9
Video-ChatGPT	3.4	43.5	3.4	49.2
PG-Video-LLaVA	3.4	42.8	3.5	48.5
Video-LLaVA	3.4	42.2	3.4	49.5
VEGAS $_{n=16}$ w/o TAM	2.7	43.6	-	-
VEGAS $_{n=16}$	2.8	46.1	-	-
VEGAS $_{n=32}$ w/o TAM	3.2	46.1	3.3	49.0
VEGAS $_{n=32}$	3.4	46.1	3.5	51.2
VEGAS-*generalist* w/o STP	3.1	48.4	3.6	51.5
VEGAS-*generalist*	3.4	48.5	3.9	54.9

Table 2: Modality and module ablation results in the open-ended setting.

VEGAS-*generalist*. The results in Table 2 highlight the modular effectiveness of GIFT and its learning strategy. Similar to TAM in VEGAS, the STP module significantly improves VEGAS-*generalist*, mitigating score drops from newly introduced designs. The STP shows notable gains in both Score and Accuracy across QV and QVS settings. The improvement in QVS further supports the beneficial interactions between modalities. Note that, despite leveraging Social-IQ expertise in GIFT, we avoid using answers as labels directly for better generalization ability.

Figure 4 shows examples comparing answers from frozen VEGAS and tuned VEGAS-*generalist*. VEGAS accurately identifies relevant frames and provides correct visual evidence (e.g., "*ironing the clothes*"), but struggles with in-depth analysis. As expected, VEGAS-*generalist* provides responses well-aligned with ground truths and enriched with

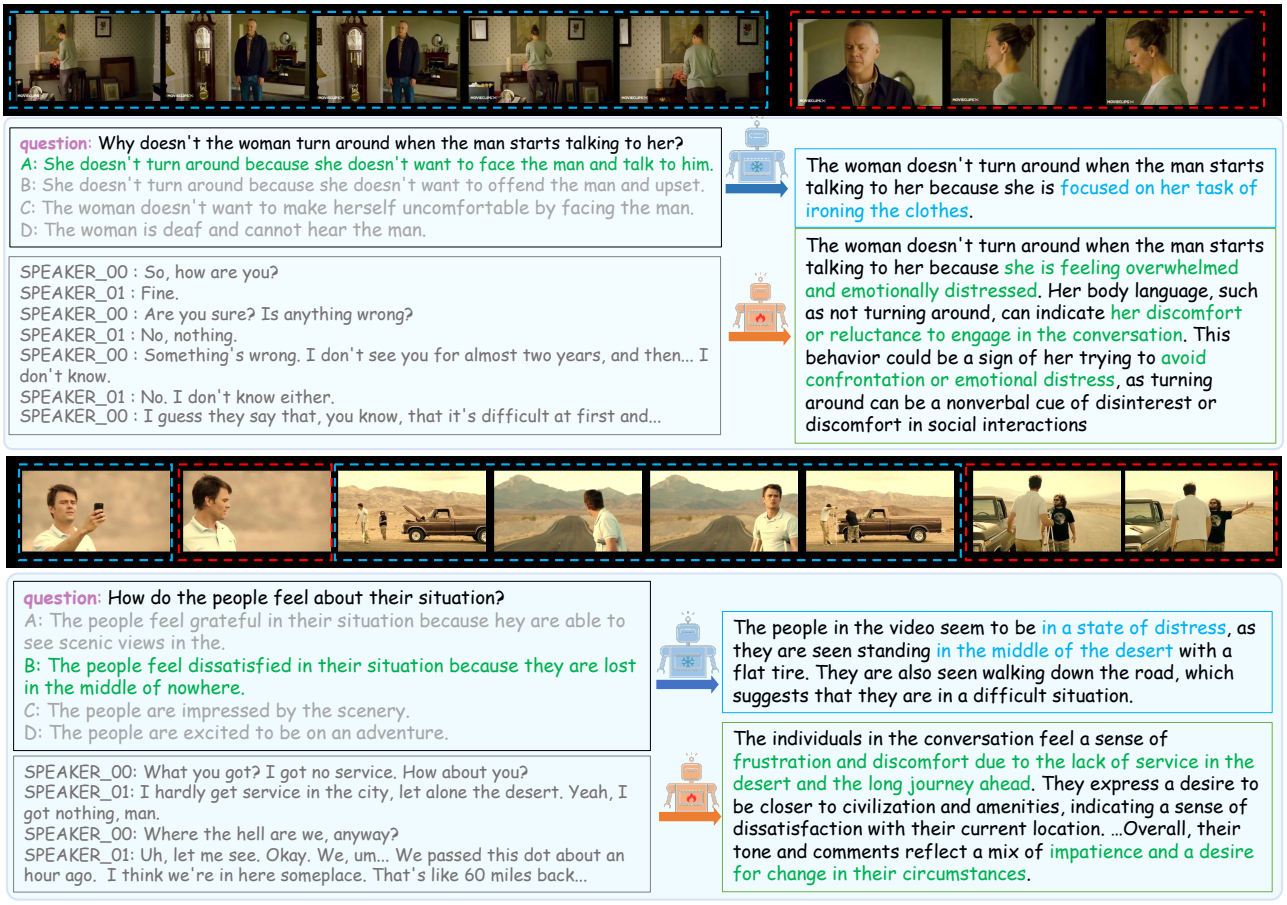


Figure 4: Open-ended QA examples of VEGAS (blue arrows) and VEGAS-*generalist* (orange arrows) using video alone and video with subtitles, respectively.

expert analysis, demonstrating a deeper understanding.

Multi-Choice QA

Modality Ablation. We first perform modality ablation studies in supervised MCQ. Table 3 shows that while DeSIQ improves performance in setting A, its advantage wanes in setting QA due to unresolved data bias. Additionally, it has difficulty understanding conversations in subtitles. In contrast, VEGAS mitigates shortcut effect in both A and QA settings while improving subtitle comprehension. More importantly, it significantly enhances the use of visual information, with a notable accuracy increase of 9.28%.

We also conduct full-parameter fine-tuning of LLM to explore potential improvements over LoRA tuning. Unexpectedly, the baseline method performs worse under this setting, likely due to the disruption of strong prior knowledge. Contrastively, despite a decrease in maximum accuracy for VEGAS-*full*, it surprisingly demonstrates significant improvements in reducing language shortcuts and enhancing the visual modality contribution by 30.63%. This evident advantage proves that, despite the overfitting risk associated with full-parameter fine-tuning, our approach avoids relying on spurious correlations in the language input. In subsequent experiments, we continue with the LoRA version of VEGAS due to its comprehensive and balanced performance.

Model	A↓	QA ↓	QAV	QAVS
DeSIQ*	63.35	64.63	62.28 $_{(-2.35)}$	-
DeSIQ	28.07	57.23	68.93$_{(+11.7)}$	37.72 🔊
Video-LLaVA	69.57	77.77	77.34 $_{(-0.43)}$	79.17
VEGAS	57.00	66.23	75.51 $_{(+9.28)}$	80.90
Video-LLaVA-*full*	66.01	73.46	74.75$_{(+1.29)}$	75.51
VEGAS-*full*	30.53	47.46	73.67$_{(+30.63)}$	76.37

Table 3: Modality ablation results in supervised MCQ. 🔊 denotes that the audio modality is used along with the subtitles. * denotes baseline of the method.

Maximum Accuracy Comparison. The upper part of Table 4 compares maximum accuracy across models. DeSIQ shows better performance when using audio rather than subtitles as an auxiliary modality. Overall, the proposed VEGAS model demonstrates state-of-the-art performance. There are some generative vision-language models (Xu et al. 2023; Li et al. 2024) also reported zero-shot **binary** Accu-

Mode	Model	Setting	Accuracy
Supervised	Just-Ask	MC	52.12
	Just-Ask-Plus	MC	53.35
	DeSIQ*	MC	64.63
	DeSIQ	MC	74.13 🔊
	MMTC-ESC*	MC	74.91
	MMTC-ESC	MC	75.94
	Video-LLaVA	MC	79.17
	VEGAS	MC	80.90
Zero-shot	R-VLM	Unknown	63.7
	IVA	Unknown	68.0
	Video-LLaVA	MC	60.6
	Video-LLaVA	OE	52.5
	VEGAS	MC	60.0
	VEGAS	OE	66.0

Table 4: Maximum accuracy comparison under supervised (upper) and zero-shot (lower) settings. 🔊 denotes that the audio is used as an auxiliary modality.

racy evaluated with ChatGPT on Social-IQ, which means they only compare the prediction with the correct answer. It is also unclear whether they handle open-ended or closed-set QA, the latter being defined by provided answer options. Therefore, we compare under both scenarios in the lower part of Table 4. Note that IVA was jointly trained with 34k NEXT-QA samples in addition to 136k instruction-tuning data. In contrast, our VEGAS model, trained on only 33k samples designed for the sampler, achieves 66.0% accuracy, closely matching IVA's 68.0%.

Exploring Emotion Understanding Ability

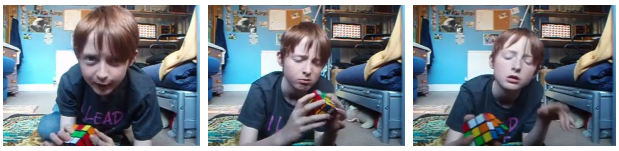
Emotion Recognition. Emotions are crucial indicators of people's attitudes during social interactions. We use the IEMOCAP (Busso et al. 2008) dataset for multimodal emotion recognition validation, which includes emotional conversations from actors in both scripted and spontaneous spoken scenarios. As shown in Table 5, baseline models display varying performance, likely due to characteristics of their instruction tuning datasets. For instance, Video-ChatGPT is trained on ActivityNet-200 (Caba Heilbron et al. 2015) with a stronger focus on human activities. VEGAS-*generalist* achieves superior performance: 25.5% with video alone, 35.8% with video and subtitles, and 37.4% when incorporating audio additionally.

Emotional Video Captioning. When observing social events, we humans naturally perceive the underlying sentiments and the nuanced emotions at play. In this part, we investigate such ability on a random subset of EmVidCap dataset (Wang et al. 2022). The dataset bridges visual content and linguistic sentences by combining factual and emotional elements in captions. We use uniformly sampled frames as input since a captioning task often requires full-length video understanding.

The lower part of Table 5 presents scores reported by GPT-3.5-turbo, which was prompted to focus on emotional consistency. Despite the zero-shot setting, VEGAS-*generalist* gains evident advantages over generic MLLM baselines. Examples in Figure 5 further demonstrate human-like reasoning and analyzing processes, underscoring the intelligence level and broader implications of our approach.

Model	V	V+Sub	V+Sub+Aud
Video-LLAMA	15.3	17.8	17.1
Video-ChatGPT	22.2	25.0	-
PG-Video-LLaVA	7.6	18.9	-
Video-LLaVA	16.4	16.6	-
VEGAS-*generalist*	25.5	35.8	37.4
Video-LLAMA	0.5	0.5	0.4
Video-ChatGPT	1.1	1.3	-
PG-Video-LLaVA	0.9	1.2	-
Video-LLaVA	1.5	2.0	-
VEGAS-*generalist*	1.7	2.3	1.7

Table 5: Accuracy of emotion recognition on IEMOCAP (upper), and ChatGPT score on EmVidCap (lower).



Predicted caption: The video captures a young boy's intense concentration and determination as he tries to solve a Rubik's cube. His facial expressions reveal his frustration and exasperation as he struggles to solve the puzzle. The boy's body language and gestures convey his emotional state, from his tense fingers gripping the cube to his exasperated sighs and throws of the puzzle.

Predicted caption: The video captures a young boy sitting at a desk, working on a math problem. He appears to be struggling with the problem, as he slams his head down on the desk in frustration. The camera captures his emotional reaction, showing the viewer the intensity of his feelings. The boy's facial expressions and body language convey a sense of overwhelm and despair, as he grapples with the difficulty of the problem. The scene is a poignant reminder of the challenges that children can face in their academic pursuits, and the emotional toll that these challenges can take on young individuals.

Figure 5: Video captioning examples from EmVidCap.

Conclusion

In this study, VEGAS is introduced to address the trust crisis in Social-IQ, where high selection accuracy often stems from significant language shortcuts. VEGAS utilizes a modality scalable, generative, multimodal large language model (MLLM) to deliver open-ended answers that reveal the reasoning behind selections. We first incorporate a novel Language Guided Sampling (LGS) technique to extract question-relevant visual features. Then we employ a specialized Generalist Instruction Fine-Tuning (GIFT) to produce VEGAS-*generalist* to interpret social traits from them, which excels in broader social reasoning with expert-level analytical capabilities. Extensive evaluations demonstrate that VEGAS significantly enhances the integration of visual context, ensuring that it plays a pivotal role in reasoning, while effectively mitigating reliance on language shortcuts. Notably, VEGAS-*generalist* excels in social understanding with expertise in psychology and sociology, positioning it as an advancing human-like social AI.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (62171325). The numerical calculations in this paper have been done on the supercomputing system in the Supercomputing Center of Wuhan University.

References

Bainbridge, W. S.; Brent, E. E.; Carley, K. M.; Heise, D. R.; Macy, M. W.; Markovsky, B.; and Skvoretz, J. 1994. Artificial social intelligence. *Annual review of sociology*, 20(1): 407–436.

Buch, S.; Eyzaguirre, C.; Gaidon, A.; Wu, J.; Fei-Fei, L.; and Niebles, J. C. 2022. Revisiting the" video" in video-language understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2917–2927.

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

Caba Heilbron, F.; Escorcia, V.; Ghanem, B.; and Carlos Niebles, J. 2015. ActivityNet: A Large-Scale Video Benchmark for Human Activity Understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Chandra, S.; Shirish, A.; and Srivastava, S. C. 2022. To be or not to be... human? Theorizing the role of human-like competencies in conversational artificial intelligence agents. *Journal of Management Information Systems*, 39(4): 969–1005.

Chen, D.; and Dolan, W. 2011. Collecting Highly Parallel Data for Paraphrase Evaluation. In Lin, D.; Matsumoto, Y.; and Mihalcea, R., eds., *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 190–200. Portland, Oregon, USA: Association for Computational Linguistics.

Chiang, W.-L.; Li, Z.; Lin, Z.; Sheng, Y.; Wu, Z.; Zhang, H.; Zheng, L.; Zhuang, S.; Zhuang, Y.; Gonzalez, J. E.; et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3): 6.

Cho, J. W.; Kim, D.-J.; Ryu, H.; and Kweon, I. S. 2023. Generative bias for robust visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11681–11690.

Cordonnier, J.-B.; Mahendran, A.; Dosovitskiy, A.; Weissenborn, D.; Uszkoreit, J.; and Unterthiner, T. 2021. Differentiable Patch Selection for Image Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2351–2360.

Dautenhahn, K. 2007. *A paradigm shift in artificial intelligence: why social intelligence matters in the design and development of robots with human-like intelligence*. Springer.

Duéñez-Guzmán, E. A.; Sadedin, S.; Wang, J. X.; McKee, K. R.; and Leibo, J. Z. 2023. A social path to human-like artificial intelligence. *Nature Machine Intelligence*, 5(11): 1181–1188.

Fei, H.; Wu, S.; Ji, W.; Zhang, H.; and Chua, T.-S. 2024a. Dysen-VDM: Empowering Dynamics-aware Text-to-Video Diffusion with LLMs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7641–7653.

Fei, H.; Wu, S.; Ji, W.; Zhang, H.; Zhang, M.; Lee, M.-L.; and Hsu, W. 2024b. Video-of-thought: Step-by-step video reasoning from perception to cognition. In *Forty-first International Conference on Machine Learning*.

Fei, H.; Wu, S.; Zhang, H.; Chua, T.-S.; and Yan, S. 2024c. VITRON: A Unified Pixel-level Vision LLM for Understanding, Generating, Segmenting, Editing.

Fei, H.; Wu, S.; Zhang, M.; Zhang, M.; Chua, T.-S.; and Yan, S. 2024d. Enhancing video-language representations with structural spatio-temporal alignment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Gat, I.; Schwartz, I.; Schwing, A.; and Hazan, T. 2020. Removing bias in multi-modal classifiers: Regularization by maximizing functional entropies. *Advances in Neural Information Processing Systems*, 33: 3197–3208.

Guo, X.-Y.; Li, Y.-F.; and Haffari, G. 2023. DeSIQ: Towards an Unbiased, Challenging Benchmark for Social Intelligence Understanding. *arXiv preprint arXiv:2310.18359*.

Kim, C. D.; Kim, B.; Lee, H.; and Kim, G. 2019. Audiocaps: Generating captions for audios in the wild. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 119–132.

Kumar, A.; Mittal, T.; and Manocha, D. 2020. Mcqa: Multimodal co-attention based network for question answering. *arXiv preprint arXiv:2004.12238*.

Lao, M.; Pu, N.; Liu, Y.; He, K.; Bakker, E. M.; and Lew, M. S. 2023. COCA: COllaborative CAusal Regularization for Audio-Visual Question Answering. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(11): 12995–13003.

Lei, J.; Yu, L.; Bansal, M.; and Berg, T. L. 2018. Tvqa: Localized, compositional video question answering. *arXiv preprint arXiv:1809.01696*.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In Krause, A.; Brunskill, E.; Cho, K.; Engelhardt, B.; Sabato, S.; and Scarlett, J., eds., *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, 19730–19742. PMLR.

Li, Y.; Chen, X.; Hu, B.; and Zhang, M. 2024. Llms meet long video: Advancing long video comprehension with an interactive visual adapter in llms. *arXiv preprint arXiv:2402.13546*.

Liang, J.; Jiang, L.; Cao, L.; Li, L.-J.; and Hauptmann, A. G. 2018. Focal visual-text attention for visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6135–6143.

Lin, B.; Zhu, B.; Ye, Y.; Ning, M.; Jin, P.; and Yuan, L. 2023. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*.

Liu, W.; Jia, X.; Zhong, X.; Jiang, K.; Yu, X.; and Ye, M. 2025. Dynamic and static mutual fitting for action recognition. *Pattern Recognition*, 157: 110948.

Livingstone, S. R.; and Russo, F. A. 2018. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *PloS one*, 13(5): e0196391.

Maaz, M.; Rasheed, H.; Khan, S.; and Khan, F. S. 2023. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*.

Munasinghe, S.; Thushara, R.; Maaz, M.; Rasheed, H. A.; Khan, S.; Shah, M.; and Khan, F. 2023. Pg-video-llava: Pixel grounding large video-language models. *arXiv preprint arXiv:2311.13435*.

Natu, S.; Sural, S.; and Sarkar, S. 2023. External Commonsense Knowledge as a Modality for Social Intelligence Question-Answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3044–3050.

Niu, Y.; Tang, K.; Zhang, H.; Lu, Z.; Hua, X.-S.; and Wen, J.-R. 2021. Counterfactual vqa: A cause-effect look at language bias. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12700–12710.

Park, J. S.; Bhagavatula, C.; Mottaghi, R.; Farhadi, A.; and Choi, Y. 2020. VisualCOMET: Reasoning About the Dynamic Context of a Still Image. In Vedaldi, A.; Bischof, H.; Brox, T.; and Frahm, J.-M., eds., *Computer Vision – ECCV 2020*, 508–524. Cham: Springer International Publishing. ISBN 978-3-030-58558-7.

Pirhadi, M. J.; Mirzaei, M.; and Eetemadi, S. 2023. Just ask plus: Using transcripts for videoqa. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3082–3085.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *International Conference on Machine Learning*.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, 21(140): 1–67.

Wang, H.; Tang, P.; Li, Q.; and Cheng, M. 2022. Emotion Expression With Fact Transfer for Video Description. *IEEE Transactions on Multimedia*, 24: 715–727.

Wilf, A.; Ma, M. Q.; Liang, P. P.; Zadeh, A.; and Morency, L.-P. 2023. Face-to-face contrastive learning for social intelligence question-answering. In *2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG)*, 1–7. IEEE.

Wu, S.; Fei, H.; Li, X.; Ji, J.; Zhang, H.; Chua, T.-S.; and Yan, S. 2024a. Towards Semantic Equivalence of Tokenization in Multimodal LLM. *arXiv preprint arXiv:2406.05127*.

Wu, S.; Fei, H.; Qu, L.; Ji, W.; and Chua, T.-S. 2024b. NExT-GPT: Any-to-Any Multimodal LLM. In *Proceedings of the International Conference on Machine Learning*, 53366–53397.

Xiao, J.; Shang, X.; Yao, A.; and Chua, T.-S. 2021. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9777–9786.

Xie, B.; and Park, C. H. 2023. Multi-Modal Correlated Network with Emotional Reasoning Knowledge for Social Intelligence Question-Answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3075–3081.

Xu, J.; Lan, C.; Xie, W.; Chen, X.; and Lu, Y. 2023. Retrieval-based Video Language Model for Efficient Long Video Question Answering. *arXiv preprint arXiv:2312.04931*.

Yang, A.; Miech, A.; Sivic, J.; Laptev, I.; and Schmid, C. 2021. Just ask: Learning to answer questions from millions of narrated videos. In *Proceedings of the IEEE/CVF international conference on computer vision*, 1686–1697.

Yu, S.; Yoon, J.; and Bansal, M. 2024. CREMA: Multimodal Compositional Video Reasoning via Efficient Modular Adaptation and Fusion. *arXiv preprint arXiv:2402.05889*.

Yu, Z.; Xu, D.; Yu, J.; Yu, T.; Zhao, Z.; Zhuang, Y.; and Tao, D. 2019. ActivityNet-QA: A Dataset for Understanding Complex Web Videos via Question Answering. *ArXiv*, abs/1906.02467.

Zadeh, A.; Chan, M.; Liang, P. P.; Tong, E.; and Morency, L.-P. 2019. Social-iq: A question answering benchmark for artificial social intelligence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8807–8817.

Zadeh, A.; Chen, M.; Poria, S.; Cambria, E.; and Morency, L.-P. 2017. Tensor fusion network for multimodal sentiment analysis. *arXiv preprint arXiv:1707.07250*.

Zadeh, A.; Zellers, R.; Pincus, E.; and Morency, L.-P. 2016. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligent Systems*, 31(6): 82–88.

Zhang, H.; Li, X.; and Bing, L. 2023. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*.

Zhang, Z.; Luo, P.; Loy, C. C.; and Tang, X. 2018. From facial expression recognition to interpersonal relation prediction. *International Journal of Computer Vision*, 126: 550–569.

Zhu, B.; Lin, B.; Ning, M.; Yan, Y.; Cui, J.; Wang, H.; Pang, Y.; Jiang, W.; Zhang, J.; Li, Z.; et al. 2023. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment. *arXiv preprint arXiv:2310.01852*.