# IntentQA: Context-aware Video Intent Reasoning

Jiapeng Li[1,2], Ping Wei[1], Wenjuan Han[3], Lifeng Fan[2]
[1]National Key Laboratory of Human-Machine Hybrid Augmented Intelligence,
Xi'an Jiaotong University, Xi'an, China
[2]National Key Laboratory of General Artificial Intelligence,
Beijing Institute for General Artificial Intelligence (BIGAI), Beijing, China
[3]School of Computer and Information Technology, Beijing Jiaotong University, Beijing, China
lijiapeng@stu.xjtu.edu.cn, pingwei@xjtu.edu.cn, wjhan@bjtu.edu.cn, lifengfan@bigai.ai

## Abstract

*In this paper, we propose a novel task IntentQA, a special VideoQA task focusing on video intent reasoning, which has become increasingly important for AI with its advantages in equipping AI agents with the capability of reasoning beyond mere recognition in daily tasks. We also contribute a large-scale VideoQA dataset for this task. We propose a Context-aware Video Intent Reasoning model (CaVIR) consisting of i) Video Query Language (VQL) for better cross-modal representation of the situational context, ii) Contrastive Learning module for utilizing the contrastive context, and iii) Commonsense Reasoning module for incorporating the commonsense context. Comprehensive experiments on this challenging task demonstrate the effectiveness of each model component, the superiority of our full model over other baselines, and the generalizability of our model to a new VideoQA task. The dataset and codes are open-sourced at: https://github.com/JoseponLee/IntentQA.git.*

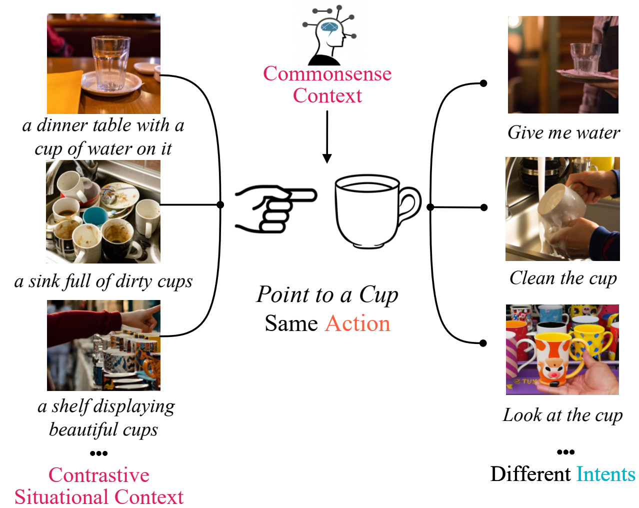


Figure 1: Illustration of challenges brought by varied contexts in video intention reasoning. The same action under different contexts could mean different underlying intents.

## 1. Introduction

Among the recent flourishing studies on cross-modal vision-language understanding, video question answering (VideoQA) is one of the most prominent to support interactive AI with the ability to understand and communicate dynamic visual scenarios via natural languages [75]. Despite its popularity, VideoQA is still quite challenging, because it demands the models to comprehensively understand the videos to correctly answer questions, which include not only factual but also inferential ones. The former (factoid VideoQA) directly asks about the visual facts (e.g., humans, objects, actions, etc.), while the latter (inference VideoQA) requires logical reasoning of latent variables (e.g., the spatial, temporal and causal relationships among entities, mental states, etc.) beyond observed visual facts [75]. The future trend for AI is to study inference VideoQA beyond factoid VideoQA [75], requiring more reasoning ability beyond mere recognition. In this paper, we propose a new task called **IntentQA**, i.e., a special kind of inference VideoQA that focuses on intent reasoning.

Intent understanding is a key building block of human intelligence. Humans have a strong inclination to interpret events as a series of goals driven by intentions [10, 58, 59]. In fact, humans do not encode the entirety of action details but rather interpret actions in terms of intentions and store these interpretations for later retrieval [3]. As a fundamental organizing principle that regulates how humans comprehend one another and act in the environment, the concept of intent has been awarded a central position within social intelligence and should thus be an essential component of future AI [76, 17]. However, as far as we know, there is no

VideoQA work focusing on intent understanding. Therefore, we believe our proposed new task is a great contribution to the development of intent reasoning in VideoQA.

The biggest challenge for video intent reasoning is **context** because intent understanding is quite context-sensitive. As illustrated in Fig. 1, humans can interpret different intents underlying the same action 'point to a cup' given different video contexts along with commonsense knowledge. The intent is more likely to be 'give me water' if the given context is 'a table at a restaurant', and 'clean the cup' if the context is 'a sink full of dirty cups', and 'look at the cup' if the context is 'a store selling beautiful cups'. The uncertainty does not come from the protruding finger, but from the context, which is the key to solving the overloaded signal and the 'dark matter' mystery. Here, the context includes the immediate communicative context, the shared experience, and the commonsense. Context-aware reasoning ability plays a significant role in human intelligence.

We contribute a new dataset for IntentQA, as detailed in Section 3. We also propose a model with three key modules that deals with three major contexts respectively: (I) **Situational Context**; (II) **Contrastive Context**; (III) **Commonsense Context**. Module I (Video Query Language (VQL)) integrates cross-modal contextual information from both videos and languages. Module II (Contrastive Learning) learns to reason from contrasting a triplet of anchor, positive and negative samples. Module III (Commonsense Reasoning) further incorporates the commonsense knowledge from the large language model.

Our main **contributions** can be summarized as follows. First, we propose a new task IntentQA, a special VideoQA task focusing on intent reasoning. Given a video and a question, the aim is to select the correct answer with the understanding of intent. Second, we collect and annotate a large-scale VideoQA dataset with natural social scene videos. Finally, we propose a Context-aware Video Intent Reasoning model (CaVIR) and provide benchmark results.

## 2. Related Work

### 2.1. Video Question Answering

As a typical cross-modal task, VideoQA answers the natural language question according to the given video, which is challenging because it requires a deep and comprehensive understanding of the semantic information of the video and question. Notably, recent studies in this domain have shifted away from the traditional reliance on 3D convolutions [6, 31] as the primary video backbone models. Instead, approaches harnessing fine-grained information, such as objects and relations, are increasingly gaining traction [64, 65]. A growing body of work recognizes the paramount importance of 'context' in addressing this problem. On the one hand, VideoQA datasets and techniques jointly evolve over time [75]. In addition to the early datasets, such as TGIFQA [24] and MSRVTT-QA [66], many more challenging datasets have emerged recently, such as NExT-QA [61], CLEVRER [68], CLEVR_HYP [49], AGQA 2.0 [20] and Causal-VidQA [30], which usually invoke complicated spatial, temporal and causal inference among multiple entities and relations [75]. On the other hand, various techniques have been developed for VideoQA [54, 75], such as Memory [13, 56], Attention [71, 72], Transformer [64, 67], Neural Modular Networks [28, 47], Neural-Symbolic methods [68, 7, 11], and Graph-structured methods [62, 64].

Such an inspiring and promising trend from recognition to reasoning in the field of VideoQA is great progress. Answering questions like 'what' is no longer the core of VideoQA, we further want to answer questions like 'why' and 'how'. However, although there are studies aiming to reason about various relationships between the visual facts (e.g., [32, 42]), few VideoQA work studies the unobserved human mental state underlying the apparent entities. To our best knowledge, our study is the first VideoQA work focusing on 'intent'. We believe intent-related VideoQA features human-level in-depth understanding of videos, demands higher-level reasoning abilities, and would promote VideoQA toward the core of human intelligence.

### 2.2. Intent Understanding

Upon seeing human actions, humans have an inherent tendency to infer other people's intentions from their actions [4]. Intent understanding plays a key role in human social intelligence [76, 17, 45, 46]. There have been some studies exploring intent inference in computer vision, robotics, etc. Jia et al. [25] collected an social media image dataset *Intentonomy* with an aim to analyze how visual information can facilitate recognition of human intent. Pei et al. [43] inferred the goals and intents of agents through an event parsing algorithm. Some studies [38, 44, 55, 52] manifest human intentions by predicting their trajectories. Holtzen et al. [21] proposed a method for robots to infer a human's hierarchical intent from partially observed RGBD videos. Yu-Ching et al. [8] used a QA approach in robotic systems to construct interactive dialogue systems, assisting robots in understanding user intentions. Sap et al. [50] measured the large language model's ability to understand intents and reactions of participants in social interactions. However, there has not yet been a cross-modal intent reasoning video dataset nor a benchmark model in VideoQA.

### 2.3. Context-aware Reasoning

Context, including not only the immediate context in videos and languages but also the commonsense knowledge, is very important for answering inference questions because knowledge underpins reasoning [33, 74, 69]. Research has demonstrated that when relevant knowledge is provided as additional context to commonsense question answering, it can substantially enhance performance [33]. Many methods that utilize objects in images to integrate context have been proposed [57, 35, 14, 16, 15]. Zheng et al. [73] proposed a novel approach for generating image captions with guiding objects. Li et al. [29] introduced a novel relation consistency loss to address the multi-instance confusion problem in video relation (context) grounding.

AI continues to be narrow and brittle due to its lack of reasoning ability of context, such as commonsense [9]. Recent years have brought about a renewed interest in commonsense representation and reasoning [23, 22, 51, 18, 70]. Current systems either rely on external knowledge bases (KBs) to incorporate additional relevant knowledge, or resort to pre-trained language models as the sole implicit source of world knowledge [53, 5]. Hwang et al. [23] built a new commonsense knowledge graph, ATOMIC2020. Lourie et al. [36] argued that QA-based commonsense datasets transfer well with each other, while commonsense knowledge graphs do not. Rainier [33] learns to generate contextually relevant knowledge in response to given commonsense questions. Arabshahi et al. [2] used a transformer-based generative commonsense knowledge base as its source of background knowledge for reasoning. In contrast to crowdsourcing, a pre-trained language model like GPT [41] [1] is a more flexible source of external knowledge and a better way to generate large-scale dialogue datasets with social commonsense knowledge, such as SODA [27]. West et al. [60] show how to selectively distill high-quality causal commonsense from GPT-3. Liu et al. [34] used external knowledge generated from a language model to improve model performance on four commonsense reasoning tasks.

## 3. Dataset

We contribute an IntentQA dataset with diverse intents in daily social activities. Examples are shown in Fig. 2.

**Dataset Construction and Annotation.** We utilize NExT-QA [61] as the source dataset to construct our dataset. NExT-QA dataset is a comprehensive VideoQA dataset with rich natural daily social activities and detailed QA annotations. Originally, the NExT-QA dataset categorizes itself into three types, i.e., *Causal*, *Temporal*, *Descriptive*. We select the inference QA types, i.e., *Causal* and *Temporal*, rather than the factoid *Descriptive*, to build our IntentQA dataset. Particularly, we select both the *Causal Why* and *Causal How* subtypes under *Causal*, and the *Temporal Previous* and *Temporal Next* subtypes under *Temporal* (see examples shown in Fig. 2). The *Causal Why (CW)* QA usually takes the form of 'Why [action]? For [intent]',

[1] In this paper, 'GPT' without a specified version refers to instructGPT (text-davinci-003) [41].

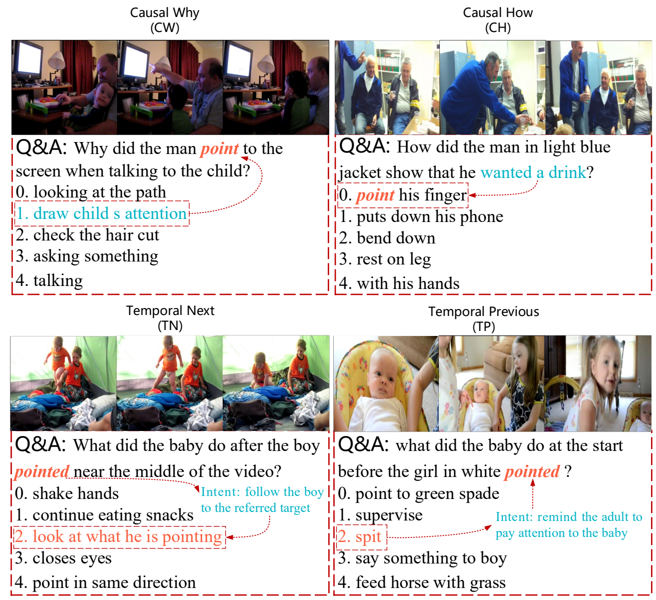


Figure 2: Illustration of four types of QA in our dataset. In the example of *CW*, a man points to the screen to draw the child's attention. In the example of *CH*, the man points his finger to express his intent for a drink. The *TN* example shows that the boy's 'pointing' leads to the baby's 'looking at what he is pointing'. The *TP* example shows that the girl's 'pointing' is motivated by the baby's 'spit' action. The red box frames the correct answer.

with the key action appearing in the question and the intent in the answer. On the contrary, the *Causal How (CH)* QA usually takes the form of 'How [intent]? By [action]', with the key action appearing in the answer and the intent in the question. The *Temporal Previous (TP)* QA usually takes the form of 'What [action A] before [action B]? ', while the *Temporal Next (TN)* QA takes the form of 'What [action B] after [action A]? '. In the *TP&TN* QA, the intent is not explicitly expressed in the question nor answer, but is the implicit causal factor linking the two sequential actions.

We use AllenNLP [12] for dependency parsing to extract the key action in QA, and obtain the Lemmatized Verb [2] of the action from the dictionary [39, 40]. We searched for synonyms based on each action's Lemmatized Verb, and merge the synonyms to assign an action ID for each cluster. After the preliminary filtering and processing, we further annotate the dataset on Amazon Mechanical Turk (AMT). We carefully design four questions to select QAs satisfying the following criteria: 1) The key action is physical, observable in the video, and conducted by a person; 2) The *same actions* refer to semantically the same and physically similar actions in the videos, rather than different actions under

[2] https://www.nltk.org/_modules/nltk/stem/wordnet.html

| *IntentQA* | Training | Validation | Testing | Total |
|---|---|---|---|---|
| # Video | 3212 | 524 | 567 | 4303 |
| # Action | 605 | 399 | 397 | 624 |
| # Lemmatized Verb | 193 | 188 | 167 | 193 |
| # Action ID | 162 | 162 | 144 | 162 |

Table 1: Statistics of IntentQA dataset.

| # VQA | *CW* | *CH* | *TP&TN* | Total |
|---|---|---|---|---|
| Training | 6989 | 1940 | 3190 | 12119 |
| Validation | 1185 | 334 | 525 | 2044 |
| Testing | 1250 | 359 | 525 | 2134 |

Table 2: Statistics of dataset splitting. # VQA refers to the number of video question answering samples.

the same or similar action words. We construct our dataset in a contrastive manner that the same actions under different contexts lead to different underlying intents, as illustrated in Fig. 1. To ensure the annotation quality, we apply the cross-validation principle and assign at least three annotators for each data sample; only when all three annotators agree will the sample be included in the final IntentQA dataset.

**Dataset Statistics.** After the filtering and annotation, our IntentQA dataset eventually contains $4,303$ videos and $16,297$ question-answer pairs. And there are $624$ actions, $193$ Lemmatized Verbs, and $162$ action IDs. See Table 1. The whole dataset is split into training, validation and testing sets in a ratio of approximately 6:1:1. After splitting, the training set contains $12,119$ QAs, the validation set contains $2,044$ QAs, and the testing set contains $2,134$ QAs (see Table 2). We guarantee that each action's Lemmatized Verb appearing in the validation/testing sets appears at least twice in the training set. For action's Lemmatized Verbs with sufficient video samples, we try to maintain a 6:1:1 ratio in the three sets. To avoid overfitting, we make sure that the same video only appears in one set.

# 4. Model

## 4.1. Overview

We define the task of IntentQA to be the same as VideoQA in terms of input and output forms, taking a video $v$, a question $q$ and a corresponding answer set $\mathbb{A}$ as input, and outputting the correct answer $a^*$ from the answer set $\mathbb{A}$.

$$a^* = \arg\max_{a\in\mathbb{A}} \; f_w(a|q, v, \mathbb{A}), \tag{1}$$

where $f_w$ represents a mapping function with learnable parameters $w$. Compared to traditional VideoQA tasks, the difference in our proposed intentQA task lies in that all the QA are related to intent understanding.

To solve this problem, we propose a Context-aware Video Intent Reasoning model (CaVIR), as shown in Fig. 3, which can sense context from three aspects. Firstly, we obtain the **Situational Context** from the video related to the question through VQL. Then, we select the positive and negative samples with the same action randomly, align the top-k highest attention nodes in the **Situational Context**, calculate the triplet loss, and obtain the **Contrastive Context**. Finally, we use GPT [41] to obtain the **Commonsense Context**, and combine the predicted distribution of our model based on the Situational Context and Contrastive Context to get the final result. To further explain the overall structure of the model, we will start with a single sample.

For a single sample, we use a simplified version of VGT [64] as our baseline model. As shown in Fig. 4, we use the frame features $V_f$ and the region features $V_r$ of the video as inputs. The region features $V_r$ are first modeled by $N$ DGTs [64] to form the region graph $G_r$, and then concatenated with $V_f$ to obtain the frame/region graph $G_{f,r}$:

$$G_{f,r} = \text{Concat}(V_f, DGT(V_r)), \tag{2}$$

where DGT is from VGT [64], and we use the same settings. For the language part, we concatenated the questions and answers together and extract language features with Bert:

$$F_{q,\mathbb{A}} = \text{Bert}(\text{Concat}(q, \mathbb{A})). \tag{3}$$

Then, we use the region graph $G_r$ and the language feature $F_{q,\mathbb{A}}$, and employ the VQL model to extract the cross-modal graph $G_{r|q,\mathbb{A}}$, i.e., the **Situational Context**:

$$G_{r|q,\mathbb{A}} = \text{VQL}(G_r, F_{q,\mathbb{A}}). \tag{4}$$

Next, we use the cross-modal graph $G_{r|q,\mathbb{A}}$ extracted by VQL to obtain **Contrastive Context** through contrastive learning. Finally, we use a multi-head self-attention (MHSA) transformer to fuse all the features and obtain a composite feature representation $F_{f,r|q,\mathbb{A}}$ of the video:

$$F_{f,r|q,\mathbb{A}} = \text{MHSA}(G_{f,r} + G_{r|q,\mathbb{A}}). \tag{5}$$

In Section 4.4, we detail how we predict the results through **Commonsense Context** for the test pipeline.

## 4.2. Video Query Language (VQL)

We use a video query language (VQL) approach to obtain visual contexts related to the question from the video. As shown in Fig. 4, we use the video region graph $G_r$ obtained by extracting features from $N$ DGTs to query the QA features $F_{q,\mathbb{A}}$ extracted by BERT, and calculate the similarity matrix $S_{r|q,\mathbb{A}}$:

$$S_{r|q,\mathbb{A}} = G_r(F_{q,\mathbb{A}})^\intercal. \tag{6}$$

Multiplying the similarity matrix $S_{v|q,\mathbb{A}}$ and the language feature $F_{q,\mathbb{A}}$, transforming the language feature $F_{q,\mathbb{A}}$

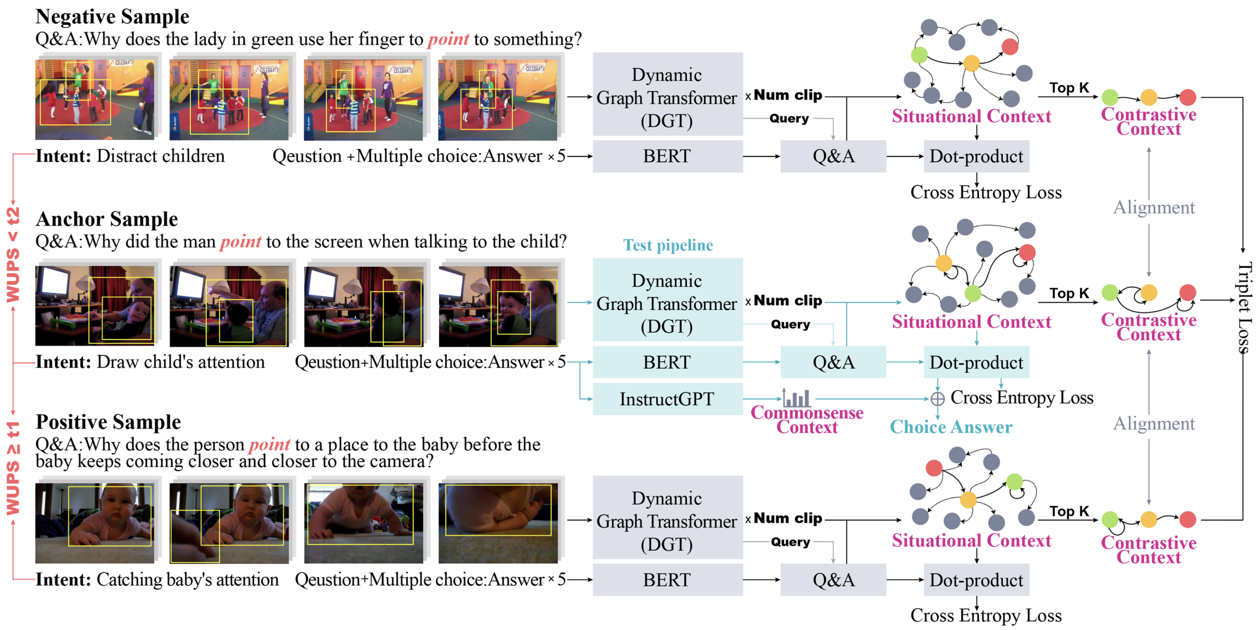


Figure 3: Overview of our Context-aware Video Intent Reasoning model (CaVIR). The figure contains a triplet of samples, i.e., the anchor sample, the positive sample and the negative sample. In the anchor sample, the agent guides the children to look at the screen. In the positive example, the agent guides the children to learn the skill of 'crawling'. In the negative example, the 'point' action is a trick used by the agent in green to distract the children's attention and win the game. Our model utilizes situational context, contrastive context and commonsense context to solve the IntentQA task. The blue color highlights the test pipeline. The yellow bounding boxes show the region features fed into our model.

into the visual feature space $G_v$, and then fusing them together to form the cross-modal graph $G_{r|q,\mathbb{A}}$, which is the question-relevant video contexts we need:

$$G_{r|q,\mathbb{A}} = G_r + S_{r|q,\mathbb{A}} F_{q,\mathbb{A}}. \tag{7}$$

## 4.3. Contrastive Learning

We select positive and negative examples based on two similarity conditions between the action and answer of two QAs. To control the action similarity, we divide it into three levels according to action consistency/action's Lemmatized Verb consistency/action ID consistency. In order to determine whether two samples with the same action are positive or negative to each other, we compare the similarity of their answers via WUPS score [37]. As formula Eq. (8) shows, when two QA samples, A and B, have a WUPS score between their correct answers ($a^*_A$,$a^*_B$) that is greater than or equal to a threshold $t_1$, we consider A and B to be positive samples for each other; otherwise, if the WUPS score is below a threshold $t_2$, we regard A and B as negative samples of each other:

$$\text{Relation}(A,B) = \begin{cases} \text{Pos.} & \text{WUPS } (a^*_A, a^*_B) \geq t_1, \\ \text{Neg.} & \text{WUPS } (a^*_A, a^*_B) < t_2. \end{cases} \tag{8}$$

We collect positive and negative examples for each QA to allow the anchor sample to randomly select one positive and one negative example to form a triplet as the input.

As shown in Fig. 3, we first extract the features $F_{r|q,A}$ of the top-k nodes from cross-modal graph $G_{r|q,A}$ responding most to the question and answer set according to the similarity matrix $S_{r|q,\mathbb{A}}$:

$$F_{r|q,\mathbb{A}} = \text{top-k}(G_{r|q,\mathbb{A}}). \tag{9}$$

We repeat this operation for the three samples in the triplet to obtain $F^a_{r|q,\mathbb{A}}$, $F^p_{r|q,\mathbb{A}}$, $F^n_{r|q,\mathbb{A}}$. Then we align the features of the negative example $F^n_{r|q,\mathbb{A}}$ and the positive example $F^p_{r|q,\mathbb{A}}$ to the anchor sample:

$$\begin{aligned} {F^p_{r|q,\mathbb{A}}}^{\text{align}} &= (F^a_{r|q,\mathbb{A}} (F^p_{r|q,\mathbb{A}})^\intercal) F^p_{r|q,\mathbb{A}}, \\ {F^n_{r|q,\mathbb{A}}}^{\text{align}} &= (F^a_{r|q,\mathbb{A}} (F^n_{r|q,\mathbb{A}})^\intercal) F^n_{r|q,\mathbb{A}}. \end{aligned} \tag{10}$$

The distance between the anchor sample and the positive/negative samples, $d(a,p)$ and $d(a,n)$ are computed as:

$$\begin{aligned} d(a,p) &= (F^a_{r|q,\mathbb{A}} - {F^p_{r|q,\mathbb{A}}}^{\text{align}})^2, \\ d(a,n) &= (F^a_{r|q,\mathbb{A}} - {F^n_{r|q,\mathbb{A}}}^{\text{align}})^2. \end{aligned} \tag{11}$$

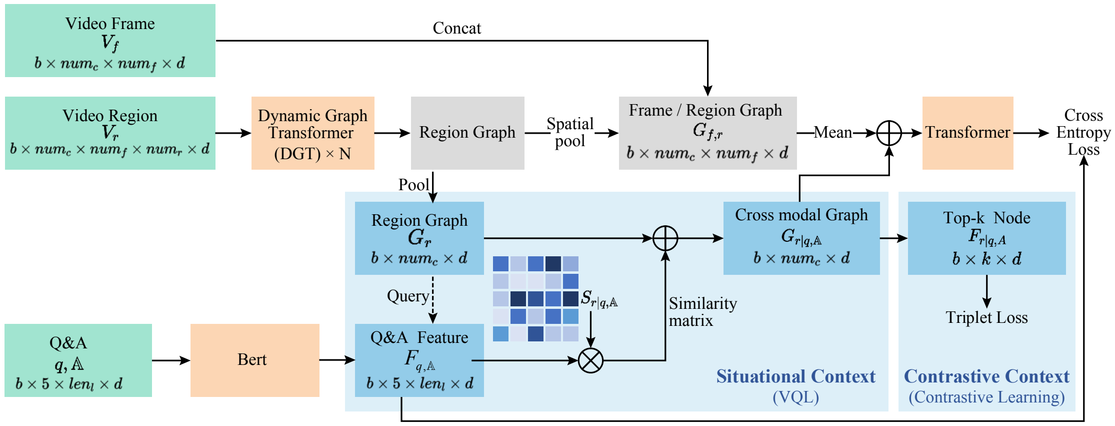


Figure 4: The model architecture for a single sample input. Green colors highlight the input. Orange highlights the modules borrowed from VGT [64]. Blue highlights our new modules for extracting different contexts. $b$ represents the batch size. $num_c$ indicates the number of clips. $num_f$ denotes the number of frames in each clip. $num_r$ denotes the number of regions per frame. $d$ is the dimension of the features. $len_l$ is the length of Q&A. $k$ refers to the top-k nodes selected.

The triplet loss is:

$$L_{\text{triplet}} = \max(d(a,p) - d(a,n) + \text{margin}, 0). \tag{12}$$

For each sample of the triplet, the cross-entropy loss is:

$$L_{\text{ce}} = -\sum_{i=1}^{|\mathbb{A}|} y_i \log S_i. \tag{13}$$

The matching score $S$ is calculated as Eq. (15). The complete loss $L$ calculated as:

$$L = L_{\text{ce}}^a + L_{\text{ce}}^p + L_{\text{ce}}^n + L_{\text{triplet}}. \tag{14}$$

### 4.4. Commonsense Reasoning

We propose a simple method that allows the model to combine the prior commonsense information provided by GPT [41] in the test stage.

As shown in the Test Pipeline of Fig. 3, the language feature $F_{q,\mathbb{A}}$ and the composite feature $F_{f,r|q,\mathbb{A}}$ after the global MHSA transformer are calculated by dot product to obtain the matching scores $S$ of the answer set $\mathbb{A}$:

$$S = F_{f,r|q,\mathbb{A}}(F_{q,\mathbb{A}})^{\intercal}. \tag{15}$$

Then, we prompt GPT [41] with the following template: '[question]. Please choose the most likely answer from the following options according to the given question and commonsense. [answer set]' to get the confidence distribution $S_{gpt}$ of the question's answer set from GPT [41]. We combine the two distributions with a penalty coefficient $\lambda$ as:

$$S_{\text{joint}} = S + \lambda S_{\text{gpt}}, \tag{16}$$

where $S_{\text{joint}}$ is the joint distribution of the matching score for the answer set $\mathbb{A}$. Consequently, the candidate answer with highest confidence is returned as the final prediction:

$$a^* = \arg\max_{a \in \mathbb{A}} S_{\text{joint}}. \tag{17}$$

## 5. Experiments

### 5.1. Ablation Experiments

#### 5.1.1 Model Component Diagnosis

To assess the effectiveness of our essential components, we design the following comprehensive ablation experiments, as shown in Table 3. **'Blind GPT'** only use GPT [41] for the IntentQA task, and thus with no video input. **'Base Model'** is a simplified VGT model. **'+ VQL'** adds Video Query Language onto the base model to get a cross-modal graph for better situational context representation. **'+ Triplet Loader'** loads the anchor sample together with the positive and negative samples during training. **'+ Triplet Loss'** continue to add triplet margin loss. **'+ GPT'** adds commonsense prior of GPT during the test. All the components are **subsequently** and **cumulatively** added to the previous model.

As shown in Table 3, our entire model achieves the best performance during all tests, and each component of our model contributes remarkably to the performance improvements. In particular, **'+ GPT'** performs the best on all three types of QAs, and brings the biggest performance increase in the total test (+3.14%). In addition, **'+ Triplet Loader'** also achieves great total accuracy improvement (the second largest). Among all the experiments, *TP&TN*-type QAs seem to be the hardest compared to *CW* and *CH*, which might be due to the fact that intents are not explicitly expressed in *TP&TN*-type QAs. We further examined the

| Model ID | Model | *CW* | | *CH* | | *TP&TN* | | Total | |
|---|---|---|---|---|---|---|---|---|---|
| | | Val. | Test | Val. | Test | Val. | Test | Val. | Test |
| 0 | Blind GPT | - | 52.16 | - | 61.28 | - | 43.43 | - | 51.55 |
| 1 | Base Model | 50.89 | 51.76 | 54.79 | 56.27 | 48.00 | 47.05 | 50.78 | 51.36 |
| 2 | + VQL | 51.65 | 52.32 | 54.49 | 58.77 | 47.62 | 48.00 | 51.08 | 52.34 (+0.98) |
| 3 | + Triplet Loader | 51.56 | 53.60 | 56.89 | 60.72 | 48.00 | 49.52 | 51.52 | 53.80 (+1.46) |
| 4 | + Triplet Loss | 52.57 | 55.28 | 57.47 | 61.56 | 46.10 | 47.81 | 51.71 | 54.50 (+0.70) |
| 5 | + GPT | - | **58.40** | - | **65.46** | - | **50.48** | - | **57.64** (+3.14) |

Table 3: Ablation diagnosis of our model components. We use accuracy (%) as the metric.

| Model ID | Model | *CW* | | *CH* | | *TP&TN* | | Total | |
|---|---|---|---|---|---|---|---|---|---|
| | | Val. | Test | Val. | Test | Val. | Test | Val. | Test |
| 5 | Ours full | - | 58.40 | - | **65.46** | - | **50.48** | - | **57.64** |
| 4 | Ours - GPT | 52.57 | 55.28 | 57.47 | 61.56 | 46.10 | 47.81 | 51.71 | 54.50 |
| 6 | Ours - Triplet Loader - Triplet Loss | - | 57.20 | - | 63.51 | - | 46.86 | - | 55.72 |
| 7 | Ours - VQL | - | **58.56** | - | 62.67 | - | 49.71 | - | 57.08 |

Table 4: Ablation diagnosis of individual components. We use accuracy (%) as the metric.

performance contributions of each component within the model by selectively removing individual components. As shown in Table 4, the omission of any single component led to a discernible degradation in the performance of the complete model.

#### 5.1.2 Contrastive Learning Analysis

To further verify our contrastive learning approach, we analyze how the selection criterion for contrastive samples would influence the performance. We control two factors for selecting positive and negative samples: (1) What is used to calculate the action similarity, which could be 'Action', 'lemmatized verb' or 'Action ID'; (2) The WUPS score threshold for answer similarity. We set the threshold $t_2$ in Eq. (8) to 0.5, and discuss on the value of threshold $t_1$, i.e., $t_1 = 0.85$ or $t_1 = 1$. As shown in Table 5, model 4 ('Action, $t1 = 1$') achieves the best performance. The result indicates the stricter criterion of action/answer similarity, the better performance.

#### 5.1.3 Context Attention Analysis

In order to verify whether our model learns to extract the most significant context information to solve the IntentQA task, we add two more analysis experiments: (i) **Mask Randomly**. We randomly mask $k$ nodes of the cross-modal graph ($G_{r|q,\mathbb{A}}$, see Section 4.1). (ii) **Mask Lowest Attention**. We mask the bottom $k$ nodes of the cross-modal graph with the lowest attention. As shown in Table 5, randomly masking the nodes severely hurts the model performance (decrease from $54.5$ to $51.5$), while masking the nodes with the lowest attention only influences the model performance very slightly (decrease from $54.5$ to $54.05$). The results verify that our model's capability in paying attention to the most valuable parts of the context.

#### 5.1.4 Prompt Engineering Analysis

To mitigate the potential impact of prompt engineering on performance, we experimented with several other prompts. We took into account the influence of the 'chain of thought' and further incorporated 'let's think step by step' to achieve additional improvements. As can be observed in Table 6, without considering the 'chain of thought', the performance across different prompts is comparable. However, after adding 'let's think step by step', the model's performance showed a notable enhancement.

### 5.2. Comparison with VideoQA Baselines

We compare our full model with several established VideoQA baseline models, as shown in Table 7. We select several established VideoQA models from 2015 to 2022 as the baselines, including *EVQA* [1] proposed for the earliest VQA task, *CoMem* [19] and *HME* [13] using memory modules to model visual appearance, motion and language, as well as *HGA* [26], *VGT* [64] and *HQGA* [63] using the graph to model videos. These selected baseline models respectively represent several typical methods for VideoQA.

As shown in Table 7, our full model performs the best, and our model without GPT performs the second best. The early VideoQA models may focus on QA about video content description, i.e., the factoid VideoQA, they perform poorly on our IntentQA task, which requires better reasoning abilities of the unobservable intent. However, even the most recent SOTA models *VGT* and *HQGA* still have a large performance gap with our model. Contrastive situational context effectively improves model performance

| Model ID | Model | *CW* | | *CH* | | *TP&TN* | | Total | |
|---|---|---|---|---|---|---|---|---|---|
| | | Val. | Test | Val. | Test | Val. | Test | Val. | Test |
| 2 | Base Model + VQL | 51.65 | 52.32 | 54.49 | 58.77 | 47.62 | 48.00 | 51.08 | 52.34 |
| 4-1 | Action ID, $t1 = 0.85$ | 51.48 | 52.32 | 50.60 | 58.77 | 48.38 | 47.62 | 50.54 | 52.25 |
| 4-2 | Action ID, $t1 = 1$ | 50.13 | 53.12 | 54.49 | 57.66 | 47.24 | 48.19 | 50.10 | 52.67 |
| 4-3 | Lemmatized Verb, $t1 = 0.85$ | 51.90 | 53.36 | 55.99 | 62.12 | 44.00 | 45.52 | 50.54 | 52.91 |
| 4-4 | Lemmatized Verb, $t1 = 1$ | 50.80 | 54.56 | 55.09 | **62.67** | 48.00 | 46.10 | 50.78 | 53.84 |
| 4-5 | Action, $t1 = 0.85$ | 50.21 | 52.00 | 56.29 | 59.61 | **48.95** | **49.52** | 50.88 | 52.67 |
| 4 | Action, $t1 = 1$ | **52.57** | **55.28** | **57.47** | 61.56 | 46.10 | 47.81 | **51.71** | **54.50** |
| 4-6 | Mask Randomly | 50.89 | 51.28 | 53.89 | 56.27 | 45.14 | <u>48.76</u> | 49.90 | 51.50 |
| 4-7 | Mask Lowest Attention | <u>52.83</u> | <u>54.72</u> | <u>57.49</u> | <u>59.89</u> | <u>46.86</u> | 48.38 | <u>52.05</u> | <u>54.05</u> |

Table 5: Analysis of contrastive learning (best shown in bold) and context attention (best shown with underline). We use accuracy (%) as the metric.

| Model ID | Model | *CW* | | *CH* | | *TP&TN* | | Total | |
|---|---|---|---|---|---|---|---|---|---|
| | | Val. | Test | Val. | Test | Val. | Test | Val. | Test |
| 5 | prompt 1 | - | 58.40 | - | 65.46 | - | 50.48 | - | 57.64 |
| 5-1 | prompt 2 | - | 58.64 | - | 64.07 | - | 49.52 | - | 57.31 |
| 5-2 | prompt 3 | - | 58.24 | - | 63.51 | - | 50.29 | - | 57.17 |
| 5-3 | prompt 1 + 'Let's think step by step' | - | 59.12 | - | **65.74** | - | **51.81** | - | **58.43** |
| 5-4 | prompt 2 + 'Let's think step by step' | - | **59.28** | - | **65.74** | - | 48.95 | - | 57.83 |
| 5-5 | prompt 3 + 'Let's think step by step' | - | 58.00 | - | 64.07 | - | 51.24 | - | 57.36 |

Table 6: Prompt ablations. Prompt 1 is the original one. Prompt 2 is 'According to the given question and common sense, please choose the most likely intention of the protagonist in the question from the following options.' Prompt 3 is 'From the perspective of understanding the intention of the protagonist in the question, select the most likely answer from the following options.' We use accuracy (%) as the metric.

on *CW* and *CH* QA, but only slightly improves the performance on *TP&TN* QA. Commonsense context further significantly improve the model performance in all types of QA tasks.

In addition, we reported human results in Table 7, which are far superior to our model and other established models. This indicates that compared to existing models, humans still have a great advantage in understanding the intentions of humans in social contexts. At the same time, this also highlights the importance of the task we proposed, and the exploration of model understanding of social intentions and human cognition is still in its early stages. This problem is distinct from the traditional video understanding problem. It comprehends the video from the perspective of human cognition, exploring the hidden human intentions beneath the surface visual context, providing a novel perspective for video understanding.

## 5.3. Generalization Test

We test our IntentQA model's generalization ability to other VideoQA tasks. We choose a large-scale open-ended VideoQA dataset *MSRVTT-QA*, which contains 244k descriptive QA pairs and is a challenging traditional factoid VideoQA dataset, different from our inference VideoQA dataset. All the models, i.e., *VGT*, *Ours (w/o triplet loss)* and *Ours (w/ triplet loss)*, are pre-trained on our *IntentQA* dataset, and then finetuned on *MSRVTT-QA*. Table 8 shows the results. Both of our two models achieve better accuracy than the baseline, and the model with triplet loss generalizes better. The test verifies our conjecture that intent reasoning and understanding based on contrastive situational context would help the model to better understand the video contexts, and generalize well to a new factoid VideoQA task.

## 5.4. Qualitative Results and Analysis

**How Does VQL Work?** In the example illustrated in Fig. 5 (a), there are three men playing different instruments. To answer the question correctly, the model needs to understand that the question is asking about 'the man in brown checkered', and pay attention to the correct context in the video while ignoring other contexts. Our base model gets the wrong answer, but our model with VQL successfully predicts the correct answer. *Blind GPT* could not answer correctly without any video context input.

**How Does Commonsense Context Work?** In the example shown in Fig. 5 (b), the question asks why the child put the spoon into his mouth, and the candidates 'scoop food', 'eat', 'feed', and 'drop food' all appear in the video, which might confuse the model a lot. The basic two models simply choose the most obvious action 'scoop food' in the video. The two models with contrastive learning correctly understand that the subject is the baby, but still get the wrong answer. Note that it's the mother that is feeding the baby, thus the most appropriate answer is 'eat food'. The slight differ-

| Model ID | Model | Text Rep. | *CW* | | *CH* | | *TP&TN* | | Total | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Val. | Test | Val. | Test | Val. | Test | Val. | Test |
| - | EVQA [1] | GloVe | 25.99 | 25.92 | 37.43 | 34.54 | 28.00 | 25.52 | 28.38 | 27.27 |
| - | CoMem [19] | GloVe | 31.56 | 30.00 | 35.63 | 28.69 | 28.57 | 28.95 | 31.46 | 29.52 |
| - | HGA [26] | GloVe | 29.45 | 32.00 | 35.03 | 30.64 | 29.71 | 31.05 | 30.43 | 31.54 |
| - | HME [13] | GloVe | 30.97 | 34.40 | 35.33 | 34.26 | 34.29 | 29.14 | 32.53 | 33.08 |
| - | HQGA [63] | GloVe | 32.49 | 33.20 | 38.32 | 34.26 | 34.48 | 36.57 | 33.95 | 34.21 |
| - | CoMem [19] | BERT | 46.75 | 47.68 | 57.49 | 54.87 | 41.71 | 39.05 | 47.21 | 46.77 |
| - | HGA [26] | BERT | 43.54 | 44.88 | 56.89 | 50.97 | 42.48 | 39.62 | 45.45 | 44.61 |
| - | HME [13] | BERT | 46.50 | 46.08 | 51.20 | 54.32 | 44.76 | 40.76 | 46.82 | 46.16 |
| - | HQGA [63] | BERT | 45.91 | 48.24 | 57.78 | 54.32 | 44.76 | 41.71 | 47.55 | 47.66 |
| - | VGT [64] | BERT | 50.46 | 51.44 | 55.99 | 55.99 | **48.19** | 47.62 | 50.78 | 51.27 |
| - | Blind GPT [41] | - | - | 52.16 | - | 61.28 | - | 43.43 | - | 51.55 |
| 4 | Ours w/o GPT | BERT | **52.57** | 55.28 | **57.47** | 61.56 | 46.10 | 47.81 | **51.71** | 54.50 |
| 5 | **Ours** | BERT | - | **58.40** | - | **65.46** | - | **50.48** | - | **57.64** |
| - | Human | - | - | 77.76 | - | 80.22 | - | 79.05 | - | 78.49 |

Table 7: Comparison results with the established VideoQA baseline models. We use accuracy (%) as the metric.

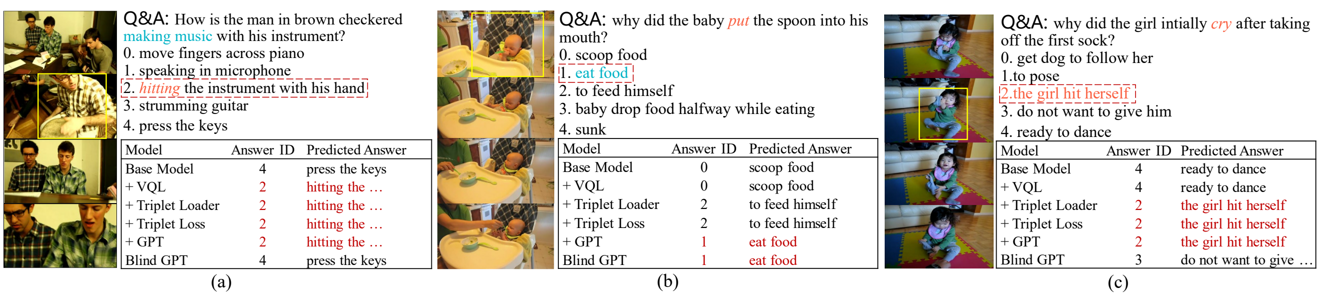


Figure 5: Qualitative Results and Analysis. The yellow boxes highlight the evidence contexts used to determine the correct answer. The red box frames the correct answer. Actions are colored in red while intents are colored in blue.

| Model ID | Model | MSRVTT-QA (OE) | |
|---|---|---|---|
| | | Val. | Test |
| - | VGT [64] | 38.26 | 39.00 |
| 3 | Ours (w/o triplet loss) | 38.11 | 39.21 |
| 4 | Ours (w/ triplet loss) | **38.98** | **39.39** |

Table 8: Generalization test on dataset *Open-ended MSRVTT-QA*. We use accuracy (%) as the metric.

ence between 'feed' and 'eat' requires deep commonsense knowledge, and thus only our model with GPT and *Blind GPT* get the answer right. *Blind GPT* can even answer correctly based solely on text and commonsense without the video context, just as humans do.

**How Does Contrastive Learning Work?** As shown in Fig. 5 (c), to answer the question 'why the girl cry after taking off the first sock', *Blind GPT* guessed the answer to be 'do not want to give him' based on commonsense, but it's wrong. The two basic models with situational context exaggerate the girl's physical movement and choose the wrong answer 'ready to dance'. The real context causing 'cry' is very subtle, being the instant moment when the girl hit foot on ground after taking off the sock. Through contrastive learning with other positive and negative samples, the model learns that usually cry is caused by injury; thus the three models with contrastive context are correct.

## 6. Conclusion

We address a new problem of IntentQA, and build a new large-scale VideoQA dataset. We propose a new model called Context-aware Video Intent Reasoning model (CaVIR), which utilizes three different contexts including situational, contrastive, and commonsense contexts. Comprehensive experiments verify the effectiveness, superiority and generalizability of our model. We hope our work will draw the field's attention and serve as important resources.

## Acknowledgement

This work was supported by the National Key Research and Development Program of China (No. 2018AAA0102501) and the Fundamental Research Funds for the Central Universities. Wenjuan Han is supported by the Talent Fund of Beijing Jiaotong University (2023XKRC006).

## References


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

[2] Forough Arabshahi, Jennifer Lee, Antoine Bosselut, Yejin Choi, and Tom Mitchell. Conversational multi-hop reasoning with neural commonsense knowledge and symbolic logic rules. *arXiv preprint arXiv:2109.08544*, 2021.

[3] Dare A Baldwin and Jodie A Baird. Discerning intentions in dynamic human action. *Trends in cognitive sciences*, 5(4):171–178, 2001.

[4] Sarah-Jayne Blakemore and Jean Decety. From the perception of action to the understanding of intention. *Nature reviews neuroscience*, 2(8):561–567, 2001.

[5] Antoine Bosselut, Ronan Le Bras, and Yejin Choi. Dynamic neuro-symbolic knowledge graph construction for zero-shot commonsense question answering. In *Proceedings of the AAAI conference on Artificial Intelligence*, volume 35, pages 4923–4931, 2021.

[6] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.

[7] Zhenfang Chen, Jiayuan Mao, Jiajun Wu, Kwan-Yee Kenneth Wong, Joshua B Tenenbaum, and Chuang Gan. Grounding physical concepts of objects and events through dynamic visual reasoning. *arXiv preprint arXiv:2103.16564*, 2021.

[8] Yu-Ching Chiu, Nanyi Bi, Richard Tsai, et al. Enhancing multi-modal intent classification in assembly scenarios through multi-task learning. *Available at SSRN 4013382*, 2022.

[9] Yejin Choi. The curious case of commonsense intelligence. *Daedalus*, 151(2):139–155, 2022.

[10] Gergely Csibra and György Gergely. 'obsessed with goals': Functions and mechanisms of teleological interpretation of actions in humans. *Acta psychologica*, 124(1):60–78, 2007.

[11] Mingyu Ding, Zhenfang Chen, Tao Du, Ping Luo, Josh Tenenbaum, and Chuang Gan. Dynamic visual reasoning by learning differentiable physics models from video and language. *Advances In Neural Information Processing Systems*, 34:887–899, 2021.

[12] Timothy Dozat and Christopher D. Manning. Deep biaffine attention for neural dependency parsing. *ArXiv*, abs/1611.01734, 2017.

[13] Chenyou Fan, Xiaofan Zhang, Shu Zhang, Wensheng Wang, Chi Zhang, and Heng Huang. Heterogeneous memory enhanced multimodal attention model for video question answering. In *IEEE/CVF conference on computer vision and pattern recognition*, pages 1999–2007, 2019.

[14] Lifeng Fan, Yixin Chen, Ping Wei, Wenguan Wang, and Song-Chun Zhu. Inferring shared attention in social scene videos. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6460–6468, 2018.

[15] Lifeng Fan, Shuwen Qiu, Zilong Zheng, Tao Gao, Song-Chun Zhu, and Yixin Zhu. Learning triadic belief dynamics in nonverbal communication from videos. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7312–7321, 2021.

[16] Lifeng Fan, Wenguan Wang, Siyuan Huang, Xinyu Tang, and Song-Chun Zhu. Understanding human gaze communication by spatio-temporal graph reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5724–5733, 2019.

[17] Lifeng Fan, Manjie Xu, Zhihao Cao, Yixin Zhu, and Song-Chun Zhu. Artificial social intelligence: A comparative and holistic view. *CAAI Artificial Intelligence Research*, 1(2):144–160, 2022.

[18] Zhiyuan Fang, Tejas Gokhale, Pratyay Banerjee, Chitta Baral, and Yezhou Yang. Video2commonsense: Generating commonsense descriptions to enrich video captioning. *arXiv preprint arXiv:2003.05162*, 2020.

[19] Jiyang Gao, Runzhou Ge, Kan Chen, and Ram Nevatia. Motion-appearance co-memory networks for video question answering. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 6576–6585, 2018.

[20] Madeleine Grunde-McLaughlin, Ranjay Krishna, and Maneesh Agrawala. Agqa 2.0: An updated benchmark for compositional spatio-temporal reasoning. *arXiv preprint arXiv:2204.06105*, 2022.

[21] Steven Holtzen, Yibiao Zhao, Tao Gao, Joshua B Tenenbaum, and Song-Chun Zhu. Inferring human intent from video by sampling hierarchical plans. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1489–1496, 2016.

[22] Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Cosmos qa: Machine reading comprehension with contextual commonsense reasoning. *arXiv preprint arXiv:1909.00277*, 2019.

[23] Jena D Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. (comet-) atomic 2020: On symbolic and neural commonsense knowledge graphs. In *the AAAI Conference on Artificial Intelligence*, volume 35, pages 6384–6392, 2021.

[24] Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *IEEE conference on computer vision and pattern recognition*, pages 2758–2766, 2017.

[25] Menglin Jia, Zuxuan Wu, Austin Reiter, Claire Cardie, Serge Belongie, and Ser-Nam Lim. Intentonomy: a dataset and study towards human intent understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12986–12996, 2021.

[26] Pin Jiang and Yahong Han. Reasoning with heterogeneous graph alignment for video question answering. In *AAAI Conference on Artificial Intelligence*, volume 34, pages 11109–11116, 2020.

[27] Hyunwoo Kim, Jack Hessel, Liwei Jiang, Ximing Lu, Youngjae Yu, Pei Zhou, Ronan Le Bras, Malihe Alikhani, Gunhee Kim, Maarten Sap, et al. Soda: Million-scale dialogue distillation with social commonsense contextualization. *arXiv preprint arXiv:2212.10465*, 2022.

[28] Thao Minh Le, Vuong Le, Svetha Venkatesh, and Truyen Tran. Hierarchical conditional relation networks for video

question answering. In *IEEE/CVF conference on computer vision and pattern recognition*, pages 9972–9981, 2020.

[29] Huan Li, Ping Wei, Jiapeng Li, Zeyu Ma, Jiahui Shang, and Nanning Zheng. Asymmetric relation consistency reasoning for video relation grounding. In *European Conference on Computer Vision*, pages 125–141. Springer, 2022.

[30] Jiangtong Li, Li Niu, and Liqing Zhang. From representation to reasoning: Towards both evidence and commonsense reasoning for video question-answering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21273–21282, 2022.

[31] Jiapeng Li, Ping Wei, Yongchi Zhang, and Nanning Zheng. A slow-i-fast-p architecture for compressed video action recognition. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2039–2047, 2020.

[32] Chen Liang, Wenguan Wang, Tianfei Zhou, and Yi Yang. Visual abductive reasoning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15565–15575, 2022.

[33] Jiacheng Liu, Skyler Hallinan, Ximing Lu, Pengfei He, Sean Welleck, Hannaneh Hajishirzi, and Yejin Choi. Rainier: Reinforced knowledge introspector for commonsense question answering. *arXiv preprint arXiv:2210.03078*, 2022.

[34] Jiacheng Liu, Alisa Liu, Ximing Lu, Sean Welleck, Peter West, Ronan Le Bras, Yejin Choi, and Hannaneh Hajishirzi. Generated knowledge prompting for commonsense reasoning. *arXiv preprint arXiv:2110.08387*, 2021.

[35] Chao Lou, Wenjuan Han, Yuhuan Lin, and Zilong Zheng. Unsupervised vision-language parsing: Seamlessly bridging visual scene graphs with language structures via dependency relationships. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15607–15616, 2022.

[36] Nicholas Lourie, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Unicorn on rainbow: A universal commonsense reasoning model on a new multitask benchmark. In *AAAI Conference on Artificial Intelligence*, volume 35, pages 13480–13488, 2021.

[37] Mateusz Malinowski and Mario Fritz. A multi-world approach to question answering about real-world scenes based on uncertain input. *Advances in neural information processing systems*, 27, 2014.

[38] R. Mehran, A. Oyama, and M. Shah. Abnormal crowd behavior detection using social force model. In *CVPR*, 2009.

[39] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[40] George A Miller. *WordNet: An electronic lexical database*. MIT press, 1998.

[41] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*, 2022.

[42] Jae Sung Park, Sheng Shen, Ali Farhadi, Trevor Darrell, Yejin Choi, and Anna Rohrbach. Exposing the limits of video-text models through contrast sets. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3574–3586, 2022.

[43] Mingtao Pei, Yunde Jia, and Song-Chun Zhu. Parsing video events with goal inference and intent prediction. In *International Conference on Computer Vision*, pages 487–494, 2011.

[44] S. Pellegrini, A. Ess, K. Schindler, and L. van Gool. You'll never walk alone: Modeling social behavior for multi-target tracking. In *IEEE/CVF conference on computer vision and pattern recognition*, 2009.

[45] Yujia Peng, Jiaheng Han, Zhenliang Zhang, Lifeng Fan, Tengyu Liu, Siyuan Qi, Xue Feng, Yuxi Ma, Yizhou Wang, and Song-Chun Zhu. The tong test: Evaluating artificial general intelligence through dynamic embodied physical and social interactions. *Engineering*, 2023.

[46] Siyuan Qi, Baoxiong Jia, Siyuan Huang, Ping Wei, and Song-Chun Zhu. A generalized earley parser for human activity parsing and prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(8):2538–2554, 2021.

[47] Zi Qian, Xin Wang, Xuguang Duan, Hong Chen, and Wenwu Zhu. Dynamic spatio-temporal modular network for video question answering. In *ACM International Conference on Multimedia*, pages 4466–4477, 2022.

[48] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

[49] Shailaja Keyur Sampat, Akshay Kumar, Yezhou Yang, and Chitta Baral. Clevr_hyp: A challenge dataset and baselines for visual question answering with hypothetical actions over images. *arXiv preprint arXiv:2104.05981*, 2021.

[50] Maarten Sap, Ronan LeBras, Daniel Fried, and Yejin Choi. Neural theory-of-mind? on the limits of social intelligence in large lms. *arXiv preprint arXiv:2210.13312*, 2022.

[51] Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. Socialiqa: Commonsense reasoning about social interactions. *arXiv preprint arXiv:1904.09728*, 2019.

[52] Tianmin Shu, Yujia Peng, Lifeng Fan, Hongjing Lu, and Song-Chun Zhu. Perception of human interaction based on motion trajectories: From aerial videos to decontextualized animations. *Topics in cognitive science*, 10(1):225–241, 2018.

[53] Vered Shwartz, Peter West, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Unsupervised commonsense question answering with self-talk. *arXiv preprint arXiv:2004.05483*, 2020.

[54] Guanglu Sun, Lili Liang, Tianlin Li, Bo Yu, Meng Wu, and Bolun Zhang. Video question answering: a survey of models and datasets. *Mobile Networks and Applications*, pages 1–34, 2021.

[55] Haowen Tang, Ping Wei, Jiapeng Li, and Nanning Zheng. Evostgat: Evolving spatiotemporal graph attention networks for pedestrian trajectory prediction. *Neurocomputing*, 491:333–342, 2022.

[56] Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. Movieqa: Understanding stories in movies through question-answering. In *IEEE conference on computer vision and pattern recognition*, pages 4631–4640, 2016.

[57] Bo Wan, Wenjuan Han, Zilong Zheng, and Tinne Tuytelaars. Unsupervised vision-language grammar induction with shared structure modeling. In *International Conference on Learning Representations*, 2021.

[58] Ping Wei, Yang Liu, Tianmin Shu, Nanning Zheng, and Song-Chun Zhu. Where and why are they looking? jointly inferring human attention and intentions in complex tasks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6801–6809, 2018.

[59] Ping Wei, Dan Xie, Nanning Zheng, and Song-Chun Zhu. Inferring human attention by learning latent intentions. In *International Joint Conference on Artificial Intelligence*, page 1297–1303, 2017.

[60] Peter West, Chandra Bhagavatula, Jack Hessel, Jena D Hwang, Liwei Jiang, Ronan Le Bras, Ximing Lu, Sean Welleck, and Yejin Choi. Symbolic knowledge distillation: from general language models to commonsense models. *arXiv preprint arXiv:2110.07178*, 2021.

[61] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9777–9786, 2021.

[62] Junbin Xiao, Angela Yao, Zhiyuan Liu, Yicong Li, Wei Ji, and Tat-Seng Chua. Video as conditional graph hierarchy for multi-granular question answering. In *AAAI Conference on Artificial Intelligence*, volume 36, pages 2804–2812, 2022.

[63] Junbin Xiao, Angela Yao, Zhiyuan Liu, Yicong Li, Wei Ji, and Tat-Seng Chua. Video as conditional graph hierarchy for multi-granular question answering. In *AAAI Conference on Artificial Intelligence*, volume 36, pages 2804–2812, 2022.

[64] Junbin Xiao, Pan Zhou, Tat-Seng Chua, and Shuicheng Yan. Video graph transformer for video question answering. In *European Conference on Computer Vision*, pages 39–58, 2022.

[65] Junbin Xiao, Pan Zhou, Angela Yao, Yicong Li, Richang Hong, Shuicheng Yan, and Tat-Seng Chua. Contrastive video question answering via video graph transformer. *arXiv preprint arXiv:2302.13668*, 2023.

[66] Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. Video question answering via gradually refined attention over appearance and motion. In *ACM international conference on Multimedia*, pages 1645–1653, 2017.

[67] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Just ask: Learning to answer questions from millions of narrated videos. In *IEEE/CVF International Conference on Computer Vision*, pages 1686–1697, 2021.

[68] Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B Tenenbaum. Clevrer: Collision events for video representation and reasoning. In *International Conference on Learning Representations*, 2020.

[69] Tao Yuan, Hangxin Liu, Lifeng Fan, Zilong Zheng, Tao Gao, Yixin Zhu, and Song-Chun Zhu. Joint inference of states, robot knowledge, and human (false-) beliefs. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5972–5978. IEEE, 2020.

[70] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *IEEE/CVF conference on computer vision and pattern recognition*, pages 6720–6731, 2019.

[71] Kuo-Hao Zeng, Tseng-Hung Chen, Ching-Yao Chuang, Yuan-Hong Liao, Juan Carlos Niebles, and Min Sun. Leveraging video descriptions to learn video question answering. In *AAAI Conference on Artificial Intelligence*, volume 31, 2017.

[72] Zhou Zhao, Jinghao Lin, Xinghua Jiang, Deng Cai, Xiaofei He, and Yueting Zhuang. Video question answering via hierarchical dual-level attention network learning. In *ACM international conference on Multimedia*, pages 1050–1058, 2017.

[73] Yue Zheng, Yali Li, and Shengjin Wang. Intention oriented image captions with guiding objects. In *IEEE/CVF conference on computer vision and pattern recognition*, pages 8395–8404, 2019.

[74] Zilong Zheng, Shuwen Qiu, Lifeng Fan, Yixin Zhu, and Song-Chun Zhu. Grice: A grammar-based dataset for recovering implicature and conversational reasoning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2074–2085, 2021.

[75] Yaoyao Zhong, Wei Ji, Junbin Xiao, Yicong Li, Weihong Deng, and Tat-Seng Chua. Video question answering: datasets, algorithms and challenges. *arXiv preprint arXiv:2203.01225*, 2022.

[76] Yixin Zhu, Tao Gao, Lifeng Fan, Siyuan Huang, Mark Edmonds, Hangxin Liu, Feng Gao, Chi Zhang, Siyuan Qi, Ying Nian Wu, et al. Dark, beyond deep: A paradigm shift to cognitive ai with humanlike common sense. *Engineering*, 6:310–345, 2020.