IntentQA: Context-aware Video Intent Reasoning

Jiapeng Li¹,², Ping Wei¹, Wenjuan Han³, Lifeng Fan²
¹National Key Laboratory of Human-Machine Hybrid Augmented Intelligence, Xi’an Jiaotong University, Xi’an, China
²National Key Laboratory of General Artificial Intelligence, Beijing Institute for General Artificial Intelligence (BIGAI), Beijing, China
³School 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.

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 TGIFQB [24] and MSVRVT-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 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] 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]’, 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

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

2https://www.nltk.org/_modules/nltk/stem/wordnet.html
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 A as input, and outputting the correct answer \( a^* \) from the answer set A.

\[
a^* = \arg \max_{a \in A} f_w(a|q, v, A),
\]

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 \( G_{f,r} \):

\[
G_{f,r} = \operatorname{Concat}(V_f, \text{DGT}(V_r)),
\]

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:

\[
F_{q,A} = \operatorname{Bert}(\text{Concat}(q, A)).
\]

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

\[
G_{r|q,A} = \operatorname{VQL}(G_r, F_{q,A}).
\]

Next, we use the cross-modal graph \( G_{r|q,A} \) 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 \( F_{f,r|q,A} \) of the video:

\[
F_{f,r|q,A} = \operatorname{MHSA}(G_{f,r} + G_{r|q,A}).
\]

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 \( F_{q,A} \) extracted by BERT, and calculate the similarity matrix \( S_{r|q,A} \):

\[
S_{r|q,A} = G_r(F_{q,A})^\top.
\]

Multiplying the similarity matrix \( S_{r|q,A} \) and the language feature \( F_{q,A} \), transforming the language feature \( F_{q,A} \)
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.

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 $S_{r|q,A}$:

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

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

$$F^p_{r|q,A} \text{ align} = (F^a_{r|q,A} (F^p_{r|q,A})^T) F^p_{r|q,A},$$
$$F^n_{r|q,A} \text{ align} = (F^a_{r|q,A} (F^n_{r|q,A})^T) F^n_{r|q,A}.$$  

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

$$d(a,p) = (F^n_{r|q,A} - F^p_{r|q,A} \text{ align})^2,$$
$$d(a,n) = (F^n_{r|q,A} - F^n_{r|q,A} \text{ align})^2.$$  

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/answer’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:

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

The triplet loss is:

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

(12)

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

\[ L_{\text{ce}} = -\sum_{i=1}^{|A|} y_i \log S_i. \]  

(13)

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

\[ L = L_{\text{ce}}^a + L_{\text{ce}}^p + L_{\text{ce}}^n + L_{\text{triplet}}. \]  

(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 \( F_{q,A} \) and the composite feature \( F_{f,r|q,A} \) after the global MHSA transformer are calculated by dot product to obtain the matching scores \( S \) of the answer set \( A \):

\[ S = F_{f,r|q,A}(F_{q,A})^T. \]  

(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 \( S_{\text{gpt}} \) of the question’s answer set from GPT [41]. We combine the two distributions with a penalty coefficient \( \lambda \) as:

\[ S_{\text{joint}} = S + \lambda S_{\text{gpt}}, \]  

(16)

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

\[ a^* = \arg \max_{a \in A} S_{\text{joint}}. \]  

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

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

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

<table>
<thead>
<tr>
<th>Model ID</th>
<th>Model</th>
<th>CW Val.</th>
<th>Test</th>
<th>CH Val.</th>
<th>Test</th>
<th>TP&amp;TN Val.</th>
<th>Test</th>
<th>Total Val.</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>Ours full</td>
<td>-</td>
<td>58.40</td>
<td>-</td>
<td>65.46</td>
<td>-</td>
<td>50.48</td>
<td>-</td>
<td>57.64</td>
</tr>
<tr>
<td>4</td>
<td>Ours - GPT</td>
<td>52.57</td>
<td>55.28</td>
<td>57.47</td>
<td>60.72</td>
<td>46.10</td>
<td>47.81</td>
<td>51.71</td>
<td>54.50</td>
</tr>
<tr>
<td>6</td>
<td>Ours - Triplet Loader - Triplet Loss</td>
<td>-</td>
<td>57.20</td>
<td>-</td>
<td>63.51</td>
<td>46.10</td>
<td>49.86</td>
<td>-</td>
<td>55.72</td>
</tr>
<tr>
<td>7</td>
<td>Ours - VQL</td>
<td>-</td>
<td>58.56</td>
<td>-</td>
<td>62.67</td>
<td>-</td>
<td>49.71</td>
<td>-</td>
<td>57.08</td>
</tr>
</tbody>
</table>

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, $t_1 = 1$’) 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 ($G_{r|q,A}$, 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 experiment 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.
Table 5: Analysis of contrastive learning (best shown in bold) and context attention (best shown with underline). We use accuracy (%) as the metric.

<table>
<thead>
<tr>
<th>Model ID</th>
<th>Model</th>
<th>CW Val.</th>
<th>Test</th>
<th>CH Val.</th>
<th>Test</th>
<th>TP&amp;TN Val.</th>
<th>Test</th>
<th>Total Val.</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>Base Model + VQL</td>
<td>51.65</td>
<td>52.32</td>
<td>54.49</td>
<td>58.77</td>
<td>47.62</td>
<td>48.00</td>
<td>51.08</td>
<td>52.34</td>
</tr>
<tr>
<td>4-1</td>
<td>Action ID, $t_1 = 0.85$</td>
<td>51.48</td>
<td>52.32</td>
<td>50.60</td>
<td>58.77</td>
<td>48.38</td>
<td>47.62</td>
<td>51.08</td>
<td>52.34</td>
</tr>
<tr>
<td>4-2</td>
<td>Action ID, $t_1 = 1$</td>
<td>50.13</td>
<td>53.12</td>
<td>55.99</td>
<td>62.12</td>
<td>44.00</td>
<td>45.52</td>
<td>50.80</td>
<td>53.84</td>
</tr>
<tr>
<td>4-3</td>
<td>Lemmatized Verb, $t_1 = 0.85$</td>
<td>50.80</td>
<td>54.56</td>
<td>55.09</td>
<td>62.67</td>
<td>48.00</td>
<td>46.10</td>
<td>50.78</td>
<td>53.84</td>
</tr>
<tr>
<td>4-5</td>
<td>Action, $t_1 = 0.85$</td>
<td>50.21</td>
<td>52.00</td>
<td>56.29</td>
<td>59.61</td>
<td>48.95</td>
<td>49.52</td>
<td>50.88</td>
<td>52.67</td>
</tr>
<tr>
<td>4</td>
<td>Action, $t_1 = 1$</td>
<td>52.57</td>
<td>55.28</td>
<td>57.47</td>
<td>61.56</td>
<td>46.10</td>
<td>47.81</td>
<td>51.71</td>
<td>54.50</td>
</tr>
<tr>
<td>4-1 Mask Randomly</td>
<td>50.89</td>
<td>51.28</td>
<td>53.89</td>
<td>56.27</td>
<td>45.14</td>
<td>48.76</td>
<td>49.90</td>
<td>51.50</td>
<td></td>
</tr>
<tr>
<td>4-7 Mask Lowest Attention</td>
<td>52.83</td>
<td>54.72</td>
<td>57.49</td>
<td>59.89</td>
<td>46.86</td>
<td>48.38</td>
<td>52.05</td>
<td>54.05</td>
<td></td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Model ID</th>
<th>Model</th>
<th>CW Val.</th>
<th>Test</th>
<th>CH Val.</th>
<th>Test</th>
<th>TP&amp;TN Val.</th>
<th>Test</th>
<th>Total Val.</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>prompt 1</td>
<td>-</td>
<td>58.40</td>
<td>-</td>
<td>65.46</td>
<td>-</td>
<td>50.48</td>
<td>-</td>
<td>57.64</td>
</tr>
<tr>
<td>5-1</td>
<td>prompt 2</td>
<td>-</td>
<td>58.64</td>
<td>-</td>
<td>64.07</td>
<td>-</td>
<td>49.52</td>
<td>-</td>
<td>57.31</td>
</tr>
<tr>
<td>5-2</td>
<td>prompt 3</td>
<td>-</td>
<td>58.24</td>
<td>-</td>
<td>63.51</td>
<td>-</td>
<td>50.29</td>
<td>-</td>
<td>57.17</td>
</tr>
<tr>
<td>5-3</td>
<td>prompt 1 + ‘Let’s think step by step’</td>
<td>-</td>
<td>59.12</td>
<td>-</td>
<td>65.74</td>
<td>-</td>
<td>51.81</td>
<td>-</td>
<td>58.43</td>
</tr>
<tr>
<td>5-4</td>
<td>prompt 2 + ‘Let’s think step by step’</td>
<td>-</td>
<td>59.28</td>
<td>-</td>
<td>65.74</td>
<td>-</td>
<td>48.95</td>
<td>-</td>
<td>57.83</td>
</tr>
<tr>
<td>5-5</td>
<td>prompt 3 + ‘Let’s think step by step’</td>
<td>-</td>
<td>58.00</td>
<td>-</td>
<td>64.07</td>
<td>-</td>
<td>51.24</td>
<td>-</td>
<td>57.36</td>
</tr>
</tbody>
</table>

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-
A2 the girl hit herself
Predicted Answer
1 hitting the …
2
Predicted Answer
4 to feed himself
to feed himself
eat
to feed himself
2
2
4
scoop food
2
the girl hit herself
ready to dance
to feed himself
do not want to
scoop food
2
press the keys
3
2
Predicted Answer
1 ready to dance
hitting the …
2 press the keys
4 ...
(c)

Human - 77.76 57.47 - 79.05 78.49 80.22 -

'cry' is very subtle, being the instant moment when the girl
text exaggerate the girl's physical movement and choose the
but it's wrong. The two basic models with situational con-
taking off the first sock', Blind GPT
(Fig. 5 (c), to answer the question 'why the girl cry after
knowledge, and thus only our model with GPT and
Blind GPT get the answer right. Blind GPT can even answer cor-
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 con-
trastive 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. 2018AA0102501) and the Fundamental Research Funds for the Central Universities. Wenjuan Han is supported by the Talent Fund of Beijing Jiaotong University (2023XKRC006).


