BDIQA: A New Dataset for Video Question Answering to Explore Cognitive Reasoning through Theory of Mind

Yuanyuan Mao[1,2], **Xin Lin**[1,2], **Qin Ni**[3], **Liang He**[1,2]

[1] Shanghai Key Laboratory of Multidimensional Information Processing, ECNU, Shanghai, China
[2]Department of Computer Science and Technology, East China Normal University
[3]Key Laboratory of Multilingual Education with AI, Shanghai International Studies University
51215901051@stu.ecnu.edu.cn, xlin@cs.ecnu.edu.cn,niqin@shisu.edu.cn,lhe@cs.ecnu.edu.cn

Abstract

As a foundational component of cognitive intelligence, theory of mind (ToM) can make AI more closely resemble human thought processes, thereby enhancing their interaction and collaboration with human. In particular, it can significantly improve a model's comprehension of videos in complex scenes. However, current video question answer (VideoQA) datasets focus on studying causal reasoning within events, few of them genuinely incorporating human ToM. Consequently, there is a lack of development in ToM reasoning tasks within the area of VideoQA. This paper presents BDIQA, the first benchmark to explore the cognitive reasoning capabilities of VideoQA models in the context of ToM. BDIQA is inspired by the cognitive development of children's ToM and addresses the current deficiencies in machine ToM within datasets and tasks. Specifically, it offers tasks at two difficulty levels, assessing **B**elief, **D**esire and **I**ntention (BDI) reasoning in both simple and complex scenarios. We conduct evaluations on several mainstream methods of VideoQA and diagnose their capabilities with zero-shot, few-shot and supervised learning. We find that the performance of pre-trained models on cognitive reasoning tasks remains unsatisfactory. To counter this challenge, we undertake thorough analysis and experimentation, ultimately presenting two guidelines to enhance cognitive reasoning derived from ablation analysis.

Introduction

In the attempt to understand the mechanisms of human for advanced intelligence, cognitive intelligence of AI has gained much attention in recent years, such as affecting computing (Poria et al. 2017) and human value alignment (Carroll 2018). Theory of mind (ToM) is the basis of human cognition. It represents a set of cognitive abilities which attribute mental states (beliefs, intentions, desires, knowledge, perspectives, etc.) to others and recognizes that these mental states may differ from one's own (Premack and Woodruff 1978). The development of ToM also has fundamental significance for AI cognition development. It can make AI more closely resemble human thought processes, thereby enhancing their interaction and collaboration (Smart 2018; Agarwal and Bansal 2021). In particular, integrating ToM reasoning into video question answering (VideoQA), can significantly improve a model's comprehension of videos in complex scenes (Zhong et al. 2022; Zellers et al. 2019). Since ToM reasoning on VideoQA requires to infer hidden information and relationships related to human understanding with the simultaneous verification of multiple skills as well as integrating visual and auditory information.

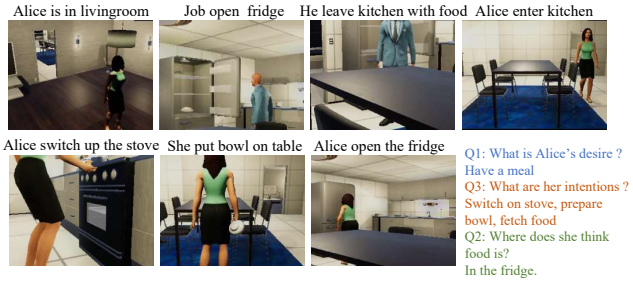

Figure 1: An example of how ToM is involved in human action explanation. Job takes the only food in fridge away and leaves the kitchen while Alice is in the living room. Alice's desire is to have a meal (desire). In the last picture, she is planning to fetch food (intention). She is walking to the empty fridge because she mistakenly thinks that the food is in the fridge (belief) and hold a false belief about the food.

However, there is a lack of development in cognitive reasoning tasks within the VideoQA domain. Some of current work on VideoQA study causal reasoning of events but few of them are involved in human mental states. For example, when asked "why does Alice walk to the **empty** fridge?" in Figure 1, only by establishing human cognitive processes can models answer this question correctly; most of existing research tend to provide straightforward action descriptions such as wanting to open fridge or fetch food. In contrast to descriptions, a comprehensive understanding necessitates an exploration of intrinsic motivation and mechanisms of a complex cognitive process for action generation. As is shown above, belief, desire, and intention (BDI) of ToM play fundamental roles and can be used to better explain human actions. Figure 2 and Figure 1 show the definition of BDI and how humans engage in cognitive reasoning with BDI reasoning. These three elements work together in a dynamic and complex manner in the human's mind. Even though there exists some work (Xiao et al. 2021; Ko et al. 2021;

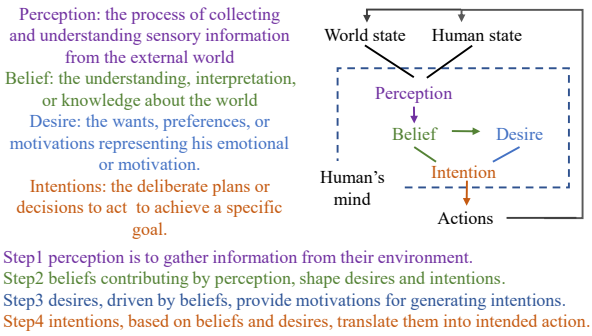

Figure 2: The definition of perception, desire, belief and intention and the relationship of them during human cognitive process.

Fang and López 2019) for intention understanding in computer vision, we argue that these work don't reveal truly understanding of intentions because they do not explore more behind these actions from a human ToM perspective.

Indeed, there have been prior studies on machine ToM. Nevertheless, persistent challenges remain within this research field, especially concerning tasks and datasets. In most current work, desire or intention is usually represented as a target or object (Baker et al. 2017; Gandhi et al. 2021). Mao et al. argue that this setting fails to accurately reflect the real-life experiences of human beings (Mao et al. 2023). And it also lacks joint reasoning on the three concepts of BDI. Another limitation is that most datasets are primarily formatted in 2D grids or simple videos (Shu et al. 2021; Gandhi et al. 2021). Worthy of reference is that most current work evaluates machine ToM with the children's ToM (Baker et al. 2017; Gandhi et al. 2021). This integration enables AI to mimic and harness the core mechanisms of human development, leading to more sophisticated and adaptable intelligent systems.

Thus, inspired by the development process of children's ToM, we contribute BDIQA, a benchmark to explore the cognitive reasoning of VideoQA models in **B**elief, **D**esire and **I**ntention of ToM. BDIQA fills the gap of cognitive reasoning tasks in VideoQA and also addresses the limitations of current machine ToM, which suffers from lacks of diversity in both dataset format and joint reasoning tasks. The questions of BDI are involved in human actions, revealing mental states and cognitive processes. BDIQA contains a two-level structure at different difficulty for ToM subabilities. Children's ToM development exhibits unique cognitive abilities, ordered developmental stages, and essential learning processes. BDIQA is leveraged by these characteristics to provide enriching and instructive experiences.

As a pioneering cognitive intelligence task, BDIQA is evaluated on several VideoQA methods such as memory network (Gao et al. 2018), graph neural network (GNN) (Duan et al. 2022a), modular networks (Le et al. 2021) and pre-trained models (Yang et al. 2021) as well as models of video classification (Li et al. 2021) with with zero-shot, few-shot and supervised learning. Notably, the current pre-trained models do not exhibit satisfactory performance on BDIQA. Thus, we delve deeper into pre-trained models' exploration on BDIQA and build our cognition reasoning models. To enable pre-trained models to adapt to cognitive reasoning, we employ leading visual techniques and expand the pre-trained model through the integration of a memory module to proceed multi-step reasoning in complex scences. Finally, our method outperforms with an overall margin of 1.17% over the baseline and we derive two cognitive reasoning guidelines through ablation study.

Our contributions can be summarized as follows:

- We contribute BDIQA, a benchmark to explore the cognitive reasoning of VideoQA in BDI of ToM with a two-level structure to facilitate phased evaluation.
- We extensively analyze the baselines and establish our VideoQA models on BDIQA, providing detailed results for various question types, and heuristic observations that can guide future research in this area.

Related Work

Machine ToM has been developed with a lot of work to incorporate ToM capabilities into machines (Rabinowitz et al. 2018; Ko et al. 2021) but the existing research has some limitations on machine ToM. On one hand, the experimental setup and data format are too simple for intention and desire reasoning within a 2D grid where the agents view their desire or intention as a goal with discrete actions (Gandhi et al. 2021; Shu et al. 2021), which is challenging to apply to real-world situations with humans. Baker et al. provided fewer than 100 samples in a 2D grid to facilitate joint inference of desires and beliefs. In overcook (Strouse et al. 2021), they collected about 400 trajectories from a two-player cooperative game in 2D grid for intention reasoning and action prediction. On the other hand, most datasets of belief reasoning are in isolation from intentions and desires in spite of some of joint inference datasets (Baker et al. 2017). Some have developed NLP question-answer (QA) datasets formalized "Sally and Anne" test with a story (Grant, Nematzadeh, and Griffiths 2017; Nematzadeh et al. 2018; Le, Boureau, and Nickel 2019) as well as image-based datasets (Eysenbach, Vondrick, and Torralba 2016; Duan et al. 2022b).

VideoQA tasks can be categorized into two types based on questions: factoid VideoQA and inference VideoQA (Zhong et al. 2022). Neither of these tasks currently involves human cognitive reasoning within ToM. Factoid questions directly inquire about visual facts, such as locations or colors (Wang et al. 2018; Lei et al. 2018; Maharaj et al. 2017). On the other hand, inference VideoQA explores the logic within dynamic scenarios (Xiao et al. 2021; Yi et al. 2019; Mao et al. 2022). As proposed by recent works, VideoQA currently emphasizes temporal and causal relationships that feature temporal dynamics. CLEVRER (Yi et al. 2019) and CLEVRER-Human (Mao et al. 2022) specially studied temporal and causal relationships of physical motions. NExT-QA (Xiao et al. 2021) is the first VideoQA dataset for casual and temporal action reasoning towards deeper explanation. With causal reasoning, VideoQA has been developing in a

more intelligent direction while few of these datasets has incorporated human cognitive reasoning within ToM.

Various methods enhance VideoQA via four components: video encoder, question encoder, cross-modal interaction, and answer decoder. Video encoders evolve for effective representation, using 2D CNNs (e.g., ResNet (He et al. 2016)) for appearance and 3D CNNs (e.g., C3D, (Tran et al. 2015), I3D (Carreira and Zisserman 2017)) for motion. Question encoders use pre-trained models (e.g., GloVe (Pennington, Socher, and Manning 2014), BERT (Devlin et al. 2018)) for semantic understanding. Sequential models (RNNs (Lev et al. 2015), CNNs (Kato and Harada 2015), Transformers (Vaswani et al. 2017)) process vision and language. Cross-modal modules (spatial/temporal attention(Dang et al. 2021), co-attention (Fan et al. 2019), multi-cycle memory (Fan et al. 2019; Gao et al. 2018), GNN (Jiang and Han 2020; Wang 2021), and conditional relation networks (Dang et al. 2021; Le et al. 2021)) fuse information for reasoning. The answer decoder emerges as the linchpin in the VideoQA pipeline to synthesize coherent and contextually appropriate responses. Leveraging the enriched features, the decoder is adept at inferring answers that reflect a profound understanding of the input data. Recent strides in visual-language pre-training profoundly affect VideoQA. Abundant pre-trained models leverage advanced techniques and transfer learning (Yang et al. 2021; Radford et al. 2021; Yang et al. 2022), even adapting from other visual-language tasks (Wang et al. 2023a; Ni et al. 2022).

Dataset

With a two-level structure inspired by children's ToM development, we introduce BQIQA, a VideoQA dataset to assess machine ToM of BDI. BDIQA assesses the sub-abilities of machine ToM at each level and for every concept. A video in BDIQA includes two characters performing household activities. And our dataset asks questions about characters' belief, desire and intention as well as perception. We generate videos in the way of animation. Synthetic datasets can control the generation of annotation, and generate large-scale datasets. Further, we conduct human evaluation of BDIQA to quantify human reasoning ability on BDIQA. Finally, by human validation and manual filtering, we obtain 3,527 videos and 19,932 QA pairs.

Task Setup

There are two characters in the video, a male character called *Job* and a female character called *Alice*. For each video, *Alice*'s desire is to complete a household activity and she makes intended plans. For the three mental states of BDI, psychology currently subdivides them into multiple ToM sub-abilities from simple to complex tasks. We build a two-level dataset to examine different aspects of ToM according to these difficulty divisions. There are five types of questions in our BDIQA - belief question, desire question, intention question, "where" question and "yes/no" question. Each mental state question is divided into two levels: *level 0* and *level 1*, where *level 0* is easier than *level 1* in infants. At the first level, BDIQA dataset involves simple reasoning tasks with satisfied desires, simple intentions and true beliefs; while at the second level, we set for harder tasks with unsatisfied desires, complex intentions and false beliefs. And we list sub-abilities of all levels of the three mental states in Table. 1 and give examples in our dataset with explanation in detail shown in Figure 3 and Figure 1.

level 0: Desire	*level 0:*Intention	*level 0:*Belief
know that human take action to meet one's desires	know the pursuit of goals	know true beliefs
*level 1:*Desire	*level 1:*Intention	*level 1:*Belief
know that desires are not always satisfied	know the choice of plans	know false beliefs

Table 1: Two levels ToM sub-abilities of BDIQA.

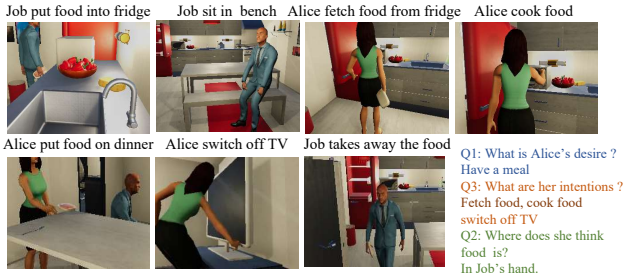

Figure 3: A example for true belief and unsatisfied desire for Alice. Alice fails to have a meal because of Job. And during that time they never leave kitchen and they have a true belief about the food which is consistent with real world. "Fetch food" is a required sub-task."Switch off TV" is an optional sub-task because it is a necessary step for "have a meal".

Desire is represented by the household activity which the character in the video wants to complete. We design 10 major household activities. For each household activity, the character called *Alice* will make plans to complete the household activity. In human desire reasoning, "desire-outcome matching strategy" (Schult 2002) refers to a cognitive process in which humans match their desires with the outcomes they expect to achieve. This strategy demonstrates a simple understanding of relationship between actions and satisfied desired outcomes while it can not be expanded to unsatisfied desire. For example in Figure 3 we can only infer Alice's desires from her actions, not the final outcomes. It is reported that 3-5 years old children show worse performance on unsatisfied desire (Schult 2002). Therefore, in *level 1* BDIQA, we add a harder situation where Job would make Alice's desire unsatisfied.

Intention is represented by the sub-task which can help to accomplish one's desire. Although there is still controversy on time, Vaish and Woodward has shown that beginning around 9 months, infants understand others' actions as driven by goals, and by 12 to 14 months infants understand others' choice of plans. It indicates that infants understand that humans take perception into account and attend to only a subset of all things before choosing an action plan. Therefore, in *level 0*, Alice will carry out only a set of steps for planning; in *level 1*, a more varied set of steps and random

sequences actions will be executed. To be specific, we divide sub-tasks into two categories: *required sub-task* and *optional sub-task*. The *required tasks* are directly related to the corresponding household activities. *Optional tasks* are not usually necessary steps but also accord with human commonsense. These optional sub-tasks will increase the difficulty in judging the true desire of the character. Because there are one more intentions for one video, intention questions are in the way of that "what does *char* do after *intention*? ".

Belief is represented by the location of an object which a character thinks. We adapt the "Sally and Anne" test (Wimmer and Perner 1983) to our household activity and ask true belief (*level 0*) and false belief questions (*level 1*). Mastering false belief means understanding that beliefs belong to people's minds and may not correspond to the external world which has been taken as the standard measure that children have a "mentalistic" understanding of beliefs (Broekhof et al. 2015). The two examples in Figure 1, 3 can help understand true and false beliefs. In a video, the character may leave the room and result in false belief. Following the classic "Sally and Anny" test , supposing that when a man leaves a room, he mistakenly thinks an object which has moved to another place is still in its original place and causes a false belief for the object. Belief questions can be formatted "where does *char* think *object* is? ".

"Where" questions are with "where is the *object*" which is different with the belief questions because the special word "think". "Where" question can be classified as perception task. For each object, our dataset asks the initial location and the last location of objects following (Grant, Nematzadeh, and Griffiths 2017). Actually, reasoning for belief questions often requires first to answer "where" questions in our setting, and then choosing one of the candidate's locations based on the character's trajectory. Therefore, belief reasoning is a multi-step task.

Yes/no questions are added to identify whether the models can judge true belief and false belief. They are templated as "does *char* have a false/true belief about *object*?". We ask the belief question in the way of "think" instead of "belief", because we believe that the current models do not yet understand the concept of "belief" which leads to the wrong answer. "Yes/no" questions explicitly complement the belief questions.

Dataset Construction

We choose *VirtualHome* (Puig et al. 2021) for us to generate videos. Based on objects and actions in *Virtualhome*, we design 10 major household activities and 28 categories of sub-tasks. *VirtualHome* is composed of 50 custom-designed departments and 4 kinds of room (bathroom, living room, kitchen and bedroom) in each department. We first identified specific sub-tasks for each household activity and design 12 templates for each household task for story generation in our videos.

Variation of data Fifty different scenarios are provided with various layouts, objects with different sizes and colors that we can create a large set of physical scene. In addition to that, we categorize each object so that the character can interact with a class of objects rather than a single object. For example, when Alice cooks food, she can cook chicken with the stove or heat a cake with the microwave. In addition, Alice may randomly take actions within common sense.

Question Generation Since these characters follow scripts to take actions and housework activities that we have designed, the housework activities and sub-tasks naturally become the labels of the desire and intention. For a video, there will be a desire question of Alice. The intention questions are based on consecutive sub-tasks in each video. In addition, we track the locations of each object and character, and following the rule Grant, Nematzadeh, and Griffiths, we can get the beliefs of each character about each object at different times. There are two belief questions, two "yes/no" question about the two characters, two "where" questions for the initial location and the last location of each object. We generate belief questions of objects which one character hold a true belief and the other hold the false belief about.

Human Evaluation and Quality Control

We conduct a crowdsourced evaluation to quantify human cognitive reasoning ability over BDIQA. We randomly sample round 2,000 QA pairs with 90% test set and 10% train set and design a web interface for data collection online. Each person is randomly assigned 6 videos and their questions. Each QA pair is assigned to over 3 random annotators. All human data is filtered based on time spent answering questions and accuracy of certain participant. Subsequently, we conduct quality control with expert re-labeling on questions with poor human performance. More detail can be found in supplementary material[1]. Finally, the human performance reaches at 84.23% on filtered test set and we compare the results of human and models in Section .

Data Statistics

BDIQA[1] contains 3,527 videos with 320*240 RGB frames and 19,932 QA pairs, including 90% for training, 10% for testing. Detailed statistics are given in Figure 4. From Figure 4(a) we can see that the number videos of each level accounts for about half of the whole dataset. And the average video length is about 192 frames and the number of video length in the dataset ranges from 30 to 1000 frames with a large time span which also proposes a challenge of long sequence videos for VideoQA. As is shown in Figure 4(c), BDI questions are the majority, constituting 54%. "Yes/no" questions of understanding as auxiliary reasoning for belief questions compose 22% of the whole dataset. Apart from these, there are 22% of "where" questions which focus on describing the locations. There are 58 answers and 6 templates to generate questions. Therefore, in Figure 4(d) the distribution of the question length is only from 5 to 11. On the whole, the questions and answers in ours are simpler than that of the counterparts. We provide various information about human behaviour, including action scripts, action localization, and scene details. We hope that researchers can use this information to expand our dataset and offer more intricate tasks, thereby exploring a wider range of human mental states. Detailed statistics can be found in supplementary material.

[1]Supplementary https://github.com/mao-yy/BDIQA.git

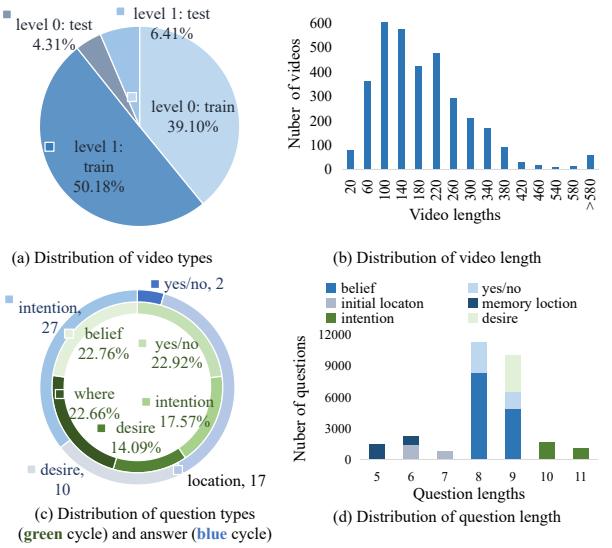

Figure 4: Data statistics.

Experiments

This section contains comprehensive experiments and intensive analyses of BDIQA. We conduct evaluations on several mainstream methods of VideoQA models and diagnose their capabilities to deal with different tasks respectively. In our primary experiments, we conduct zero-shot, few-shot, and supervised learning experiments on the video models. Subsequently, we run all models on both *level 0* and *level 1* datasets to examine the dataset hierarchy. Unexpectedly, Our discoveries unveil a distinct disparity between existing pre-trained models and the end-to-end models when applied to BDIQA. To address the lack of cognitive reasoning ability exhibited by existing pre-trained models trained solely on existing datasets, we endeavor to improve them with the best visual backbone and memory module. And we propose guidances on cognitive reasoning tasks through ablation study to enhance their performance on BDIQA. More detailed analysis can be found in supplementary material.

Experimental Setup

Baselines

- **Memory based models.** HME (Fan et al. 2019) introduces dynamic memory network (DMN) for multi-step reasoning. Comem (Gao et al. 2018) follows HME and design co-attention for multi-modal integration.
- **Hierarchical Module.** HCRN (Le et al. 2021) introduces a reusable conditional relation network (CRN) module to obtain the relationship between the various parts of the video at different levels.
- **GNN.** HGA (Jiang and Han 2020) utilizes a heterogeneous graph reasoning module and co-attention unit to capture local and global correlations between video clips and linguistic concepts, while Dual (Wang 2021) employs a stacked, iterative graph-based reasoning unit for multi-step reasoning.
- **Pre-trained Model.** Frozen (Yang et al. 2022) and JustAsk (Yang et al. 2021) are transformer-based models with VideoQA dataset. ClipBERT (Lei et al. 2021) is an efficient framework for end-to-end learning for image-text and video-text tasks with sparse sampling.

Setup Frames are resized to 224×224 as input. We divide a video into 12 clips, and randomly sample 16 frames from each clip. We employ two visual backbone models. Follow original version of JustAsk, one is S3D (Xie et al. 2018) pre-trained on HowTo100M (Miech et al. 2019). The other one is a composed of ResNet-101 (He et al. 2016) trained on ImageNet (He et al. 2016) to extract the per-frame appearance feature, and 3D ResNeXt-101 (Hara et al. 2017; Xie et al. 2017) pre-trained on Kinetics (Kay et al. 2017) to extract clip-level motion information. On the language side, all models use Bert (Devlin et al. 2018) to tokenize text inputs and get the token embedding. The initial learning rate is from 1e-4 to 1e-5 with cosine decayed in subsequent epochs. Experiments are performed on an NVIDIA 3090 GPU. More setup can be found in supplementary material.

Besides overall accuracy (overall Acc), we also report the question per-type accuracy and the accuracy of different tasks, i.e., cognitive reasoning including all BDI questions (Acc_{bdi}) and perception reasoning including initial location and memory location question (Acc_p).

Main Result

Zero-shot on BDIQA Table 2 (left) illustrates the zero-shot performance of video-language models on BDIQA. JustAsk which is trained on large-scale automatically generated VideoQA data, outperforms the other methods significantly in both BDI reasoning and perception reasoning. However, we also observe that, Frozen only achieves approximately 0.40% and ClipBERT (Lei et al. 2021) only achieves 1.09% for cognitive reasoning tasks.

Few-shot on BDIQA Table 2 (right) displays the results of few-shot performance on BDIQA with 10% data in the train set. ClipBERT presents great improvement with few-shot VideoQA and shows the best performance among the pre-trained models. Yet, there is only a slight improvement in Frozen with few-shot VideoQA. JustAsk correctly answers only 16.72% of BDI questions, in contrast to its higher accuracy on perception reasoning.

Furthermore, we conducted a video classification task on existing video models on desired reasoning. In comparison to kinetics400 (Kay et al. 2017) with 400 labels, the desired

Models	zero-shot		few-shot	
	Acc_{bdi}	Acc_p	Acc_{bdi}	Acc_p
ClipBERT	1.09%	2.05%	32.14%	40.87%
JustAsk	13.55%	29.98%	16.72%	30.80%
Frozen	0.40%	2.72%	5.04%	10.89%

Table 2: Comparison with baselines for zero-shot and few-shot VideoQA on BDIQA.

Models	desire	level 0	level 1
Uniformer (Li et al.)	26.13%	28.04%	26.65%
VideoMAE (Tong et al.)	9.14%	9.76%	7.48%
VideoMAE (Wang et al.)	10.66%	11.85%	9.35%
XClip (Ni et al.)	32.81%	37.38%	20.56%

Table 3: Comparison with pre-trained models of video classification for few-shot on BDIQA desire reasoning.

Models	overall Acc	Acc_{bdi}	Acc_p
Visual backbone: ResNet+ResNeXt-101			
Dual (Wang)	75.39%	74.78%	70.30%
HCRN (Le et al.)	76.39%	73.49%	**73.02%**
HME (Fan et al.)	71.51%	69.05%	59.58%
Comem (Gao et al.)	**79.84%**	**78.73%**	71.93%
Visual backbone: S3D (Xie et al. 2018)			
HGA (Jiang and Han)	58.63%	52.82%	63.76%
JustAsk (Yang et al.)	64.59%	64.29%	64.30%
Frozen (Yang et al.)	71.49%	70.03%	68.66%
Human	-	84.23%	80.54%

Table 4: A comparison with results of end-to-end models methods and pre-trained models of VideoQA.

reasoning task in BDIQA is considerably easier with only 10 labels, and there is some overlap actions between the two datasets. Table 3 exhibits the results of desired reasoning at two levels. Xclip outperforms other methods, achieving an accuracy of 32.81%. However, the overall accuracy of VideoMAE and VideoMAEV2 are around 10%, which are even worse than random asking. Moreover, it is evident that each model performs poorly at *level 1* compared to *level 0*.

Overall, relies on training from existing datasets and current pre-trained models, they cannot solve our BDI reasoning task effectively. In particular, at *level 1* of BDI reasoning, all models can not answer well.

Supervised Learning As is shown in Table 4, Comem (Gao et al. 2018) demonstrates the best performance among all methods. While HGA (Jiang and Han 2020) exhibits the poorest performance. The inadequate performance of HGA on the desire question task contributes to the overall result. Dual, a GNN based model and HCRN, a modular networks perform also well in overall accuracy. The performance of these better models may thank to the structures of multi-hop inference with a fusion of visual and language features. With the same epoches, we also compare the fine-tuning results of the pre-trained models. Frozen and JustAsk don't present good results as expected on the two tasks, especially JustAsk, despite its training on numerous VideoQA datasets and prior experience in solving similar tasks. It is observed that models with S3D as the video backbone such as Frozen, JustAsk, and HGA, do not perform well on BDIQA. Thus, the poor performance of the two pre-trained models may be due to the video backbone. It is also hypothesized that this outcome can be attributed to the significant disparity between our dataset and the previous VideoQA datasets.

Dataset Validation In order to obtain the validity of the dataset hierarchy of BDIQA, we conduct experiments on *level 0* and *level 1* datasets respectively. We provide an intensive comparison of baselines on BDIQA with human performance. As is shown in Table 5 (the best results for *level 0* are **bolded** and these of *level 1* are underlined), human always perform better than models on every question type except the initial location question and belief question of *level 0*. The gap performance of *level 0* and *level 1* on human and models shows that the task of *level 1* is harder than that of *level 0*. Despite demonstrating that belief questions are a multi-step task built upon "where" questions, the major models do not prioritize "where" questions over belief questions. For example, for Comem (Gao et al. 2018) the accuracy of belief question is 81.36% over that of "where" question (73.93%) at *level 1*. One possible explanation for this is the imbalance in answers across different question types. Another factor could be that the models' inference process does not align with our expectations. Additionally, we discover that the small gap between the performance of the best VideoQA model and human is 5.50% in cognitive reasoning potentially attributed to the less diversity of questions and answers of BDIQA.

Improving Pre-trained Models Performance

During analyzing visual reasoning techniques on BDIQA, we are surprised to find that the pre-trained models performed poorly. In order to transfer the prior knowledge of the pre-trained models to ours tasks, we improve the pre-trained models [1] and propose two suggestions to solve the cognitive reasoning task. Case study for our models and baseline can be found in supplementary material.

Perception is the basis of reasoning. As the above said, perception provides the foundation upon higher-order cognitive processes. Therefore, in visual cognitive tasks, the processing of visual inputs is indeed essential for successfully solving tasks. Our first idea is to replace the visual backbone. Yang (Yang et al. 2022) proposed a VideoQA pre-trained method based on freezing the visual model and bidirectional language model using light trainable modules. And we replace the visual backbone of Frozen to conduct ablation experiments. We choose S3D (Xie et al. 2018), ResNext (Hara et al. 2017), ResNet(He et al. 2016), TimeSformer (Bertasius et al. 2021) and CLIP ViT-L/14 (Kolesnikov et al. 2021) as the visual backbones of Frozen. We also follow Comem to use a simple GRN to extract appearance and motion features with ResNext and ResNet (RR).

As is shown in Table 6, compared with S3D, video representations with appearance and motion features (RR) effectively improve the overall accuracy as well as Acc_{bdi} which also surpasses the single features with only one of appearance and motion features. Other video representations also improve on BDIQA to varying degrees. This conclusion can also been validated on JustAsk, with a 14.24% improvement shown in Table 7. However, the improvements of the visual backbone are not enough for Frozen to surpass the state-of-the-art model (Comem).

[1] Code https://github.com/mao-yy/BDIQA.git

Models	Level	overall Acc	Acc_{bdi}	Acc_p	desire	intention	belief	initial	memory	yes/no
Dual	l0	82.13%	77.43%	**76.30%**	77.54%	70.65%	79.76%	88.79%	**63.46%**	96.22%
	l1	76.76%	74.78%	66.67%	77.21%	75.03%	75.32%	62.34%	70.89%	91.97%
HGA	l0	69.37%	55.09%	67.77%	12.32%	75.01%	76.95%	79.44%	55.77%	97.90%
	l1	56.54%	43.65%	66.03%	15.81%	52.78%	70.35%	63.64%	68.35%	88.33%
Comem	l0	**83.25 %**	**80.23%**	73.93%	**82.68%**	**78.26%**	81.36%	**89.72%**	57.69%	97.06%
	l1	78.19%	77.24%	66.67%	79.44%	75.56%	76.22%	63.64%	69.62%	91.12%
JustAsk	l0	77.80%	69.03%	73.46%	53.62%	73.91%	**82.24%**	84.11%	62.50%	**98.32%**
	l1	62.35%	54.38%	64.74%	43.72%	51.67%	64.66%	62.34%	67.09%	85.00%
Frozen	l0	81.13%	75.22%	74.68%	63.77%	77.17%	81.09%	88.79%	60.58%	97.90%
	l1	70.61%	66.55%	62.18%	58.60%	71.67%	74.33%	62.34%	62.03%	90.56%
Human	l0	-	84.38%	83.22%	87.12%	88.83%	80.13%	79.21%	86.10%	-
	l1	-	84.18%	79.77%	84.55%	87.68%	79.75%	79.82%	79.70%	-

Table 5: A comprehensive comparison of VideoQA methods and human evaluation on BDIQA.

Models	overall Acc	Acc_{bdi}	Acc_p
TimeSformer	76.89%	75.67%	71.65%
ResNet	73.94%	73.89%	66.75%
ResNeXt-101	71.99%	70.09%	68.38%
CLIP ViT-L-14	72.83%	72.90%	70.84%
RR	**77.84%**	**79.13%**	**71.93%**
S3D	71.49%	70.03%	68.66%

Table 6: Results with different video backbones for Frozen.

Models	overall Acc	Acc_{bdi}	Acc_p
Frozen	71.49%	70.03%	68.66%
+M	72.22%	71.51%	68.11%
+RR	77.81%	**79.13%**	71.93%
+M+RR (ours)	78.62%	76.36%	74.36%
JustAsk	64.59%	64.29%	64.30%
+RR	78.83%	76.76%	70.83%
+M+RR (ours)	**81.01%**	78.54%	**75.46%**
Comem	79.84%	78.73%	71.92%

Table 7: A comprehensive comparison of ours methods in Frozen and JustAsk. +RR means models with ResNet and ResNeXt-101. +M means models with memory module.

Reason like human. Children's ToM development is characterized by their ability to follow certain steps or patterns when solving problems. The process involves employing multiple steps of reasoning, starting from simpler concepts and gradually moving towards more complex ones. Inspired by human reasoning, some of cross-modal modules are designed to reason about complex tasks step by step such as memory network, hierarchical structure, etc. In order to help the performance of pre-trained models on BDIQA, it is available to incorporate mechanisms of human reasoning into pre-trained models. Our main idea is to incorporate the memory network into pre-trained models which enables to extract videos and questions features as well as their context. As is shown in Table 7, although the memory network module doesn't contribute much improvement to the original version of Frozen, when combined with state-of-the-art visual feature techniques, it achieves a trivial enhancement of 0.81%. This module contributes a 2.18% improvement for JustAsk and it exceeds the baseline by achieving an overall narrow margin of 1.17%.

In conclusion, our approach is simple but our improvements result in the enhancement of both the JustAsk and Frozen. Specifically, JustAsk has emerged as the superior model, surpassing Comem as the second-best model.

Conclusion and Future Work

The paper introduces a new benchmark called BDIQA, which aims to explore the cognitive reasoning capabilities of VideoQA models. It offers tasks at different difficulty levels to assess the model's understanding of BDI, and fills a gap in existing machine ToM by including BDI joint inference and video-language data. Using BDIQA, we evaluate several different VideoQA methods. After analysis, we have come to two guidelines for enhancing cognitive reasoning in VideoQA models. Our approach is simple but does improve, and these guidelines likely provide recommendations or strategies for augmenting the models to improve their performance on BDIQA tasks.

Although BDIQA is not large and the complexity of the questions to be challenging for VideoQA models , in order to provide cognitive intelligence to AI, we incorporate psychological theory into the design process, which is worth considering because it offers theoretical guidance for the development of AI learning. The additional information provided enables the expansion of our dataset for more detailed behavioral studies with mental states. The analysis of our experiment demonstrates that existing models cannot solve this task. Thus, BDIQA has also presented existing researchers with insights into how to develop novel architectures specifically tailored for cognitive reasoning in VideoQA. This is the issue we are going to face next, and we advocate these architectures can incorporate techniques from cognitive science, neuroscience, or other relevant fields to enhance AI's cognition development.

Acknowledgments

This work is supported by National Key Research and Development Program of China (2021ZD0111000/ 2021ZD0111004), the Science and Technology Commission of Shanghai Municipality Grant (No. 21511100101, 22511105901, 22DZ2229004) and Fund of the International Conference of Graduate Students of East China Normal University. Xin Lin is the corresponding author.

References

Agarwal, M.; and Bansal, S. 2021. Making AI'Smart': Bridging AI and Cognitive Science. *arXiv preprint arXiv:2112.15360*.

Baker, C. L.; Jara-Ettinger, J.; Saxe, R.; and Tenenbaum, J. B. 2017. Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour*.

Bertasius, G.; et al. 2021. Is Space-Time Attention All You Need for Video Understanding? In *Proc. of ICML*.

Broekhof, E.; Ketelaar, L.; Stockmann, L.; van Zijp, A.; Bos, M. G.; and Rieffe, C. 2015. The understanding of intentions, desires and beliefs in young children with autism spectrum disorder. *Journal of autism and developmental disorders*.

Carreira, J.; and Zisserman, A. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *Proc. of CVPR*.

Carroll, M. 2018. Overview of current AI Alignment Approaches.

Dang, L. H.; Le, T. M.; Le, V.; and Tran, T. 2021. Hierarchical object-oriented spatio-temporal reasoning for video question answering. *arXiv preprint arXiv:2106.13432*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Duan, J.; Yu, S.; Tan, H. L.; Zhu, H.; and Tan, C. 2022a. A survey of embodied ai: From simulators to research tasks. *IEEE Transactions on Emerging Topics in Computational Intelligence*.

Duan, J.; Yu, S.; Tan, N.; Yi, L.; and Tan, C. 2022b. BOSS: A Benchmark for Human Belief Prediction in Object-context Scenarios. *arXiv preprint arXiv:2206.10665*.

Eysenbach, B.; Vondrick, C.; and Torralba, A. 2016. Who is Mistaken?

Fan, C.; Zhang, X.; Zhang, S.; Wang, W.; Zhang, C.; and Huang, H. 2019. Heterogeneous memory enhanced multimodal attention model for video question answering. In *Proc. of CVPR*.

Fang, Z.; and López, A. M. 2019. Intention recognition of pedestrians and cyclists by 2d pose estimation. *IEEE Transactions on Intelligent Transportation Systems*.

Gandhi, K.; Stojnic, G.; Lake, B. M.; and Dillon, M. R. 2021. Baby Intuitions Benchmark (BIB): Discerning the goals, preferences, and actions of others. *Proc. of NeurIPS*.

Gao, J.; Ge, R.; Chen, K.; and Nevatia, R. 2018. Motion-appearance co-memory networks for video question answering. In *Proc. of CVPR*.

Grant, E.; Nematzadeh, A.; and Griffiths, T. L. 2017. How Can Memory-Augmented Neural Networks Pass a False-Belief Task? *Cognitive Science*.

Hara, K.; Kataoka, H.; Satoh, Y.; and on, S. 2017. Can Spatiotemporal 3D CNNs Retrace the History of 2D CNNs and ImageNet? *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6546–6555.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *Proc. of CVPR*.

Jiang, P.; and Han, Y. 2020. Reasoning with Heterogeneous Graph Alignment for Video Question Answering.

Kato, H.; and Harada, T. 2015. Visual Language Modeling on CNN Image Representations. *ArXiv*.

Kay, W.; Carreira, J.; Simonyan, K.; Zhang, B.; Hillier, C.; Vijayanarasimhan, S.; Viola, F.; Green, T.; Back, T.; Natsev, P.; et al. 2017. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*.

Ko, W.-R.; Jang, M.; Lee, J.; and Kim, J. 2021. AIR-Act2Act: Human–human interaction dataset for teaching non-verbal social behaviors to robots. *The International Journal of Robotics Research*.

Kolesnikov, A.; Dosovitskiy, A.; Weissenborn, D.; Heigold, G.; Uszkoreit, J.; Beyer, L.; Minderer, M.; Dehghani, M.; Houlsby, N.; Gelly, S.; Unterthiner, T.; and Zhai, X. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale.

Le, M.; Boureau, Y.; and Nickel, M. 2019. Revisiting the Evaluation of Theory of Mind through Question Answering. In *Proc. of EMNLP*.

Le, T. M.; Le, V.; Venkatesh, S.; and Tran, T. 2021. Hierarchical conditional relation networks for multimodal video question answering. *International Journal of Computer Vision*.

Lei, J.; Li, L.; Zhou, L.; Gan, Z.; Berg, T. L.; Bansal, M.; and Liu, J. 2021. Less is More: ClipBERT for Video-and-Language Learningvia Sparse Sampling. In *CVPR*.

Lei, J.; Yu, L.; Bansal, M.; and Berg, T. L. 2018. TVQA: Localized, Compositional Video Question Answering. In *Proc. of EMNLP*.

Lev, G.; Sadeh, G.; Klein, B.; and Wolf, L. 2015. RNN Fisher Vectors for Action Recognition and Image Annotation. *ArXiv*.

Li, K.; Wang, Y.; Peng, G.; Song, G.; Liu, Y.; Li, H.; and Qiao, Y. 2021. UniFormer: Unified Transformer for Efficient Spatial-Temporal Representation Learning. In *Proc. of ICLR*.

Maharaj, T.; Ballas, N.; Rohrbach, A.; Courville, A.; and Pal, C. 2017. A dataset and exploration of models for understanding video data through fill-in-the-blank question-answering. In *Proc. of CVPR*.

Mao, J.; Yang, X.; Zhang, X.; Goodman, N.; and Wu, J. 2022. CLEVRER-Humans: Describing Physical and Causal Events the Human Way. *Proc. of NeurIPS*.

Mao, Y.; Liu, S.; Zhao, P.; Ni, Q.; Lin, X.; and He, L. 2023. A Review on Machine Theory of Mind. *arXiv preprint arXiv:2303.11594*.

Miech, A.; Zhukov, D.; Alayrac, J.-B.; Tapaswi, M.; Laptev, I.; and Sivic, J. 2019. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proc. of ICCV*.

Nematzadeh, A.; Burns, K.; Grant, E.; Gopnik, A.; and Griffiths, T. L. 2018. Evaluating theory of mind in question answering.

Ni, B.; Peng, H.; Chen, M.; Zhang, S.; Meng, G.; Fu, J.; Xiang, S.; and Ling, H. 2022. Expanding Language-Image Pretrained Models for General Video Recognition.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. Glove: Global vectors for word representation. In *Proc. of EMNLP*.

Poria, S.; Cambria, E.; Bajpai, R.; and Hussain, A. 2017. A review of affective computing: From unimodal analysis to multimodal fusion. *Information fusion*.

Premack, D.; and Woodruff, G. 1978. Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*.

Puig, X.; Shu, T.; Li, S.; Wang, Z.; Liao, Y.-H.; Tenenbaum, J. B.; Fidler, S.; and Torralba, A. 2021. Watch-And-Help: A Challenge for Social Perception and Human-{AI} Collaboration. In *Proc. of ICLR*.

Rabinowitz, N.; Perbet, F.; Song, F.; Zhang, C.; Eslami, S. M. A.; and Botvinick, M. 2018. Machine Theory of Mind. In *Proc. of ICML*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *Proc. of ICML*.

Schult, C. A. 2002. Children's understanding of the distinction between intentions and desires. *Child Development*.

Shu, T.; Bhandwaldar, A.; Gan, C.; Smith, K.; Liu, S.; Gutfreund, D.; Spelke, E.; Tenenbaum, J.; and Ullman, T. 2021. Agent: A benchmark for core psychological reasoning. In *Proc. of ICML*.

Smart, P. R. 2018. Human-extended machine cognition. *Cognitive Systems Research*.

Strouse, D.; McKee, K.; Botvinick, M.; Hughes, E.; and Everett, R. 2021. Collaborating with Humans without Human Data. In *Proc. of NeurIPS*.

Tong, Z.; Song, Y.; Wang, J.; and Wang, L. 2022. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Proc. of NeurIPS*.

Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; and Paluri, M. 2015. Learning spatiotemporal features with 3d convolutional networks. In *Proc. of ICCV*.

Vaish, A.; and Woodward, A. 2005. Baby steps on the path to understanding intentions. *Behavioral and Brain Sciences*.

Vaswani, A.; Shazeer, N. M.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is All you Need. In *NIPS*.

Wang, B.; Xu, Y.; Han, Y.; and Hong, R. 2018. Movie question answering: Remembering the textual cues for layered visual contents. In *Proc. of AAAI*.

Wang, J.; Ge, Y.; Yan, R.; Ge, Y.; Lin, K. Q.; Tsutsui, S.; Lin, X.; Cai, G.; Wu, J.; Shan, Y.; et al. 2023a. All in one: Exploring unified video-language pre-training. In *Proc. of CVPR*.

Wang, J. e. a. 2021. Dualvgr: A dual-visual graph reasoning unit for video question answering. *IEEE Transactions on Multimedia*.

Wang, L.; Huang, B.; Zhao, Z.; Tong, Z.; He, Y.; Wang, Y.; Wang, Y.; and Qiao, Y. 2023b. VideoMAE V2: Scaling Video Masked Autoencoders With Dual Masking. In *Proc. of CVPR*.

Wimmer, H.; and Perner, J. 1983. Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*.

Xiao, J.; Shang, X.; Yao, A.; and Chua, T.-S. 2021. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proc. of CVPR*.

Xie, S.; Girshick, R.; Dollár, P.; Tu, Z.; and He, K. 2017. Aggregated Residual Transformations for Deep Neural Networks. In *Proc. of CVPR*.

Xie, S.; Sun, C.; Huang, J.; Tu, Z.; and Murphy, K. 2018. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *Proc. of ECCV*.

Yang, A.; Miech, A.; Sivic, J.; Laptev, I.; and Schmid, C. 2021. Just ask: Learning to answer questions from millions of narrated videos. In *Proc. of ICCV*.

Yang, A.; Miech, A.; Sivic, J.; Laptev, I.; and Schmid, C. 2022. Zero-Shot Video Question Answering via Frozen Bidirectional Language Models. In *Proc. of NeurIPS*.

Yi, K.; Gan, C.; Li, Y.; Kohli, P.; Wu, J.; Torralba, A.; and Tenenbaum, J. B. 2019. Clevrer: Collision events for video representation and reasoning. *arXiv preprint arXiv:1910.01442*.

Zellers, R.; Bisk, Y.; Farhadi, A.; and Choi, Y. 2019. From Recognition to Cognition: Visual Commonsense Reasoning. In *Proc. of CVPR*.

Zhong, Y.; Xiao, J.; Ji, W.; Li, Y.; Deng, W.; and Chua, T.-S. 2022. Video question answering: Datasets, algorithms and challenges. *arXiv preprint arXiv:2203.01225*.