Inferential Knowledge-Enhanced Integrated Reasoning for Video Question Answering

Jianguo Mao[1,2‡], Wenbin Jiang[3*], Hong Liu[1], Xiangdong Wang[1], Yajuan Lyu[3]
[1] Beijing Key Laboratory of Mobile Computing and Pervasive Device,
Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China
[2] University of Chinese Academy of Sciences, Beijing, China
[3] Baidu Inc., Beijing, China
{maojianguo20s, xdwang, hliu}@ict.ac.cn, {jiangwenbin, lvyajuan}@baidu.com

Abstract

Recently, video question answering has attracted growing attention. It involves answering a question based on a fine-grained understanding of video multi-modal information. Most existing methods have successfully explored the deep understanding of visual modality. We argue that a deep understanding of linguistic modality is also essential for answer reasoning, especially for videos that contain character dialogues. To this end, we propose an Inferential Knowledge-Enhanced Integrated Reasoning method. Our method consists of two main components: 1) an Inferential Knowledge Reasoner to generate inferential knowledge for linguistic modality inputs that reveals deeper semantics, including the implicit causes, effects, mental states, etc. 2) an Integrated Reasoning Mechanism to enhance video content understanding and answer reasoning by leveraging the generated inferential knowledge. Experimental results show that our method achieves significant improvement on two mainstream datasets. The ablation study further demonstrates the effectiveness of each component of our approach.

Introduction

Video question answering raises high demands on multi-modal representation learning and answer reasoning (Tapaswi et al. 2016; Jang et al. 2017; Xu et al. 2017; Lei et al. 2018, 2020; Garcia et al. 2020a). It requires a comprehensive understanding of visual and linguistic modality inputs (subtitles-based dialogues, questions, and candidate answers) to determine the answer (Lei et al. 2018; Garcia et al. 2020a). Recently, we have witnessed significant improvements in visual modality understanding (Gao et al. 2018; Fan et al. 2019; Kim, Tang, and Bansal 2020; Chadha, Arora, and Kaloty 2020; Garcia and Nakashima 2020a; Huang et al. 2020; Le et al. 2020; Li et al. 2020; Zellers et al. 2021). However, a deeper understanding of linguistic modality needs to be explored (Kim et al. 2021; Engin et al. 2021).

The general paradigm of video question answering is to encode the visual and linguistic inputs individually, then fuse them and feed them into a classifier to determine the answer. Many prior works have leveraged vision domain technologies to obtain a fine-grained understanding of visual modality. For instance, Kim et al. (Kim, Tang, and Bansal 2020) explored dense image captioning (Johnson, Karpathy, and Fei-Fei 2016) to extract local-to-global visual semantics to obtain a deep understanding of video content. Chadha et al. (Chadha, Arora, and Kaloty 2020) explored an unsupervised visual causal inference method to correct commonsense errors while generating dense image captioning. Garcia et al. (Garcia and Nakashima 2020a) adopted symbolized scene graphs to model the visual semantics. Due to the consistency and complementarity of multi-modal information, the linguistic modality also plays an essential role in video understanding and answer reasoning, especially for videos that contain character dialogues. To this end, Kim et al. Engin et al. (Engin et al. 2021) retrieved the plot summary of the video as external knowledge to complementary the background information related to the question. Kim et al. (Kim et al. 2021) applied contrastive learning to promote the representation learning of linguistic modality. Despite the success, it is still in its infancy. We believe that the understanding of linguistic modality can be further deepened by imitating the inference behavior of humans when watching videos.

When watching a video, humans can understand a complex story by leveraging their powerful inference ability. Based on the video dialogues, they usually infer implicit causes and effects of the event and the characters' mental states. This inferred implicit information is essential for video content understanding, called Inferential Knowledge in this paper. Intuitively, it is also beneficial for the intelligent system to understand the video content and reason the answer.

To this end, we propose the Inferential Knowledge-Enhanced Integrated Reasoning method to achieve this goal in two steps. **First, generating inferential knowledge for linguistic inputs (subtitles-based dialogues, questions, and candidate answers).** We design an Inferential Knowledge Reasoner, which receives a sentence and an inferential knowledge label as input, and generates corresponding inferential knowledge in an autoregressive manner. E.g., given a sentence *"She is drunk"* and a knowledge label *"<xWant>"* as input, it generates the inferential knowledge *"to go to sleep"*. It enables such inference ability by training on a large-scale knowledge graph containing vari-

*Corresponding author.
‡This work was done during internship at Baidu Inc.

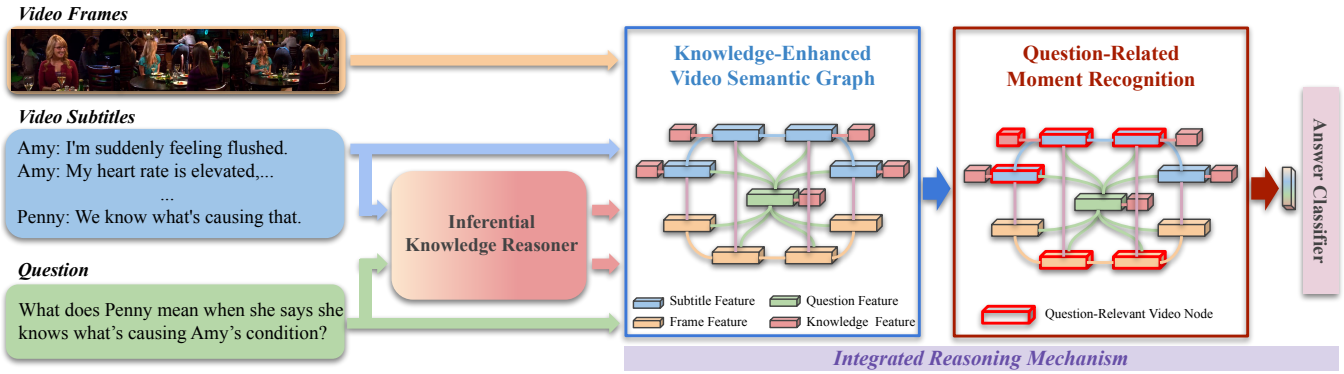


Figure 1: An overview of the Inferential Knowledge-Enhanced Integrated Reasoning Model.

ous inferential knowledge. Table 1 illustrates all the knowledge labels in the training corpus and their specific meanings. When selecting different labels, it generates different types of inferential knowledge, including causes, effects, mental states, etc. With this implicit information, the intelligent system can understand the meanings beyond words to better understand video content as humans do. **Second, conducting integrated reasoning, which leverages the inferential knowledge to enhance video content understanding and answer reasoning.** To this end, we construct a Knowledge-Enhanced Video Semantic Graph consisting of video frames, video subtitles, questions, candidate answers, and inferential knowledge. Then, we apply the graph message passing mechanism to update the representation of each element in the graph, which enables the inter and intra-modality information interaction to be more effective. Finally, the question-relevant video information and inferential knowledge are recognized and aggregated to predict the answer by the proposed Question-Relevant Moment Recognition task. Compared with the general paradigm (Lei et al. 2018; Kim, Tang, and Bansal 2020; Kim et al. 2021) of the video question answering model, it alleviates the problem of local information bias (Kim et al. 2021) by directly modeling the entire video multi-modal information and enhances the understanding of linguistic modality by the inferential knowledge.

We evaluate our method on two video question answering datasets that contains character dialogues. Experimental results show that our method achieves significant improvements on both datasets. Furthermore, we provide ablation results to demonstrate the effectiveness of the Inferential Knowledge and Integrated Reasoning Mechanism.

To sum up, the contributions of our work are as follows:

- We design the Inferential Knowledge Reasoner to imitate the inference behavior of humans when watching videos to reveal the implicit semantics of linguistic inputs.
- We propose the Integrated Reasoning Mechanism that leverages the generated inferential knowledge to enhance video content understanding and answer reasoning.
- We conduct experiments on two video question answering datasets, demonstrating the superiority of the proposed model compared to prior works.

Label	Meaning
<xIntent>	PersonX's intention of sth
<xReaction>	PersonX's reaction after sth happened
<oReaction>	Other's reaction of sth
<xNeed>	PersonX need to do sth
<xEffect>	Effect on PersonX
<xWant>	PersonX wants to do sth
<oWant>	Other wants to do sth
<oEffect>	Effect on Other people
<xAttribute>	PersonX's attribute

Table 1: All knowledge labels in the training corpus (ATOMIC) of the Inferential Knowledge Reasoner.

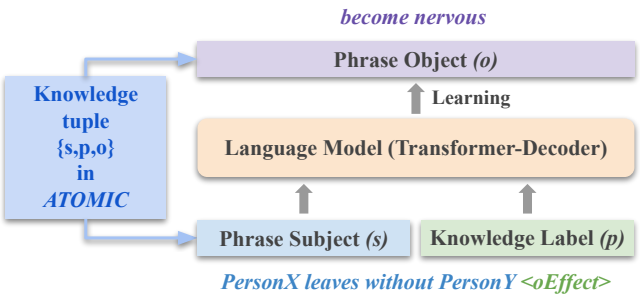
Method

Figure 1 illustrates the overall architecture of the proposed method, which consists of two components: (1) Inferential Knowledge Reasoner, which can generate inferential knowledge for linguistic inputs (subtitles-based dialogues, questions, and candidate answers). (2) Integrated Reasoning mechanism, which leverages the generated inferential knowledge to enhance video content understanding and answer reasoning.

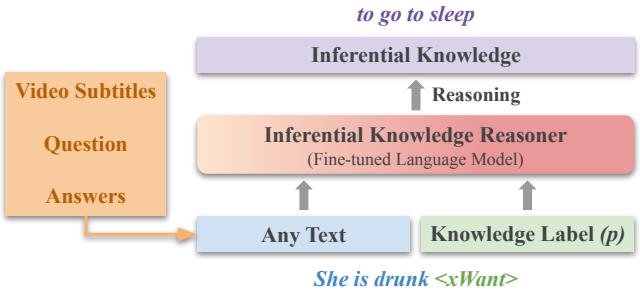
Inferential Knowledge Reasoner

Inferential Knowledge Reasoner (IKR) aims to imitate the inference behaviour of humans when watching videos. It generates inferential knowledge for linguistic modality information to reveal the implicit causes, effects, mental states, etc., which are not explicitly present in the linguistic inputs. With the help of inferential knowledge, the intelligent system can better understand the video content as humans do.

The IKR is a transformer-decoder-based (Vaswani et al. 2017) language model. As Figure 2 shows, at the training stage, it learns such inference ability by training on the ATOMIC (Sap et al. 2019), a large-scale knowledge graph that contains various inferential knowledge tuples. More specifically, e.g., given a knowledge tuple $\{s, p, o\}$ in the ATOMIC, where s ="*PersonX leaves without PersonY*" is the phrase subject of the tuple, p ="*<oEffect>*" is the



(a) Training Stage



(b) Inference Stage

Figure 2: The training and inference process of the Inferential Knowledge Reasoner.

knowledge label of the tuple, and o ="*become nervous*" is the phrase object of the tuple, the IKR is trained to predict the phrase object o given the phrase subject s and knowledge label p. To this end, the object function $\mathcal{O}$ is the conditional likelihood of predicting the phrase object tokens:

$$\mathcal{O}_{ikr} = \prod_{t=|s|+|p|}^{|s|+|p|+|o|} P(x_t|x_{<t}) \quad (1)$$

where $|s|$,$|p|$, and $|o|$ are the number of tokens in the subject phrase, knowledge label, and object phrase, respectively. The model parameters are fine-tuned by minimizing the loss function $\mathcal{L}_{ikr}$, which is the negative log of the objective function:

$$\mathcal{L}_{ikr} = -\sum_{t=|s|+|p|}^{|s|+|p|+|o|} logP(x_t|x_{<t}) \quad (2)$$

At the inference stage, given a sentence as s, and a knowledge label p defined in ATOMIC (Sap et al. 2019) as input, the IKR generates corresponding inferential knowledge in an autoregressive manner. We use it to generate inferential knowledge for video subtitles, questions, and candidate answers to enhance the understanding of the linguistic modality as humans do. We generate nine kinds of inferential knowledge for each sentence and choose the one with the highest confidence.

Training Corpus ATOMIC (Sap et al. 2019) is a knowledge graph of everyday commonsense reasoning, including 877K textual descriptions of inferential knowledge and nine knowledge labels, covering the laws of daily social life. Table 1 shows all knowledge labels in ATOMIC. In contrast to existing knowledge graph (Speer, Chin, and Havasi 2017) centering around taxonomic knowledge, it focuses on commonsense inferential knowledge organized as typed if-then relations with variables, e.g., given a tuple s ="*PersonX leaves without PersonY*", p ="*<oEffect>*", o ="*become nervous*", which means if PersonX leaves without PersonY, then PersonY is likely to become nervous.

Integrated Reasoning Mechanism

This section illustrates the Integrated Reasoning Mechanism, which leverages the generated inferential knowledge to enhance video content understanding and answer reasoning. First, we construct a **Knowledge-Enhanced Video Semantic Graph** to organize video multi-modal information and the generated inferential knowledge in a unified form. Then, we apply the graph message passing mechanism to update the representations of each element in the graph, which enable the inter and intra-modality information interaction. Finally, we recognize and aggregate **Question-Relevant Information** to reason the answer.

Feature Representation Given a video comprised of frame and subtitle sequence. We use ResNet-101 (He et al. 2016) trained on ImageNet (Deng et al. 2009) to extract frame features $F = \{f_l\}_{l=1}^{L}$, where $f_l \in \mathbb{R}^{d^v}$. Then, we split the subtitle sequence into T sentences, we use BERT (Devlin et al. 2018) to extract subtitles representations $S = \{s_t\}_{t=1}^{T}$, where $s_t \in \mathbb{R}^{d^l}$. More Specifically, we first add a special token $[CLS]$ at the beginning of the sentence as input. Then, we take the hidden state of $[CLS]$ token at the last layer as the representation of the sentence. To obtain the representations of question and candidate answers, we follow the Kim et al. (Kim, Tang, and Bansal 2020) approach to create N hypotheses H by concatenating each candidate answers with question. Then, we use BERT to extract hypotheses representations $H = \{h_n\}_{n=1}^{N}$, where $h_n \in \mathbb{R}^{d^l}$. For each sentence in the video subtitles and hypotheses, the Inferential Knowledge Reasoner will reason out K inferential knowledge, where K is the number of knowledge labels defined in ATMOIC (Sap et al. 2019). we also use BERT to extract the inferential knowledge representations $I = \{i_{m,k}\}_{m,k=1}^{T+N,K}$, where $i_{m,k} \in \mathbb{R}^{d^l}$. In addition, to be consistent with linguistic representation dimension d^l, we apply a linear transformation on visual representation to let d^v equal to d^l.

Knowledge-Enhanced Video Semantic Graph The graph is designed to enhance the video content understanding and answer reasoning by leveraging inferential knowledge, and alleviate the local information bias problem (Kim et al. 2021) by graph message passing mechanism. As Figure 3 (b) shows, it organizes video multi-modal information (video frames, video subtitles, question, candidate answers) and the generated inferential knowledge as a graph, which allows inter and intra-modality information interaction by graph message passing mechanism. In contrast, as the Figure 3 (a) shows, the general paradigm

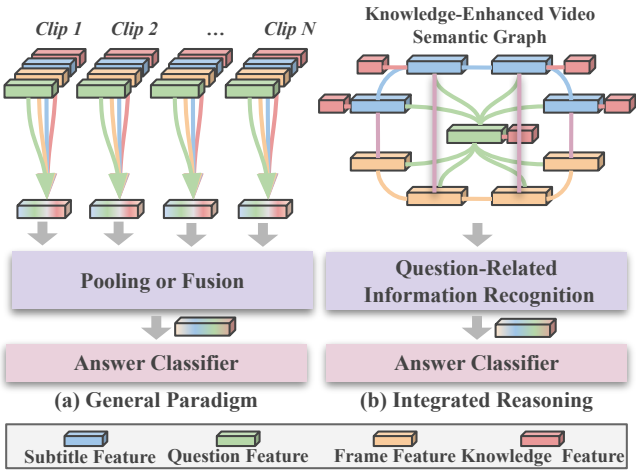


Figure 3: Our Integrated Reasoning mechanism and general paradigm of video question answering.

usually aggregates local information by the pooling method (Lei et al. 2018, 2020; Kim, Tang, and Bansal 2020; Kim et al. 2021), which may easily be affected by the local information bias (Kim et al. 2021). The graph is represented by $\mathcal{G} = \{F, S, H, I\}$. All frames nodes in F are connected to each other, and all subtitle nodes in S are also connected to each other for intra-modality information interaction. In addition, each frame node is connected to its temporal-aligned subtitle nodes for inter-modality information interaction. The hypothesis node in H is connected to all video nodes. Furthermore, both hypothesis and subtitle nodes are connected to corresponding inferential knowledge nodes in I.

Each node needs to further interact with its connected nodes for inter and intra-modality information interaction and enhance the understanding of linguistic modality with the inferential knowledge. To this end, the representation of each node $\mathbf{n_i}$ in $\mathcal{G}$ is updated according to its all neighbor nodes iteratively:

$$\mathbf{n}_i = a_{ii}\mathbf{W_g}\mathbf{n}_i + \sum_{j \in \mathcal{N}_i} a_{ij}\mathbf{W_g}\mathbf{n}_j \tag{3}$$

where $\mathbf{W_g}$ is a weight matrix, $\mathbf{A}_i = \{a_{ij}\}_{j=i}^{\mathcal{N}_i}$ is the attention weight of node $\mathbf{n}_i$ with its neighbor nodes, $\mathcal{N}_i$ is the neighbors of the node $\mathbf{n_i}$ in the graph. We obtain the weight $\mathbf{A_i}$ by computing the scaled dot-product of the $\mathbf{n_i}$ with its all neighbor nodes, divided by the dimension $\sqrt{d_{n_i}}$ of $\mathbf{n_i}$, and apply a softmax function to obtain the weights:

$$\mathbf{A}_i = \text{softmax}(\frac{\mathbf{n_i}\mathcal{N}_\mathbf{i}^\top}{\sqrt{d_{n_i}}}) \tag{4}$$

Finally, we update the representation of each node using a fully connected feed-forward network and employ a residual connection.

$$\mathbf{n}_i = \phi(\mathbf{n_i}\mathbf{W_f} + b_f) + \mathbf{n_i} \tag{5}$$

where $\phi(\cdot)$ is the ReLU activation function and $\mathbf{W_f}$ is a weight matrix.

Question-Relevant Information Recognition Given a video, answering different questions usually requires focusing on different moments within this video (Lei et al. 2018). To recognize nodes relevant to the given question in the Knowledge-Enhanced Video Semantic Graph, we design a binary classification task termed Question-Relevant Information Recognition. We first use the human importance annotations provided in the datasets as the supervisory signal, which are time spans (start-end point pairs) annotators think needed for answering the given question. If a video node (frame or subtitle) is within the time span, we regard it as relevant to the given question and have $y = 1$ as its label; otherwise, we get $y = 0$ as its label. After graph representation learning, we use a fully-connected layer with a sigmoid activation function as the classifiers to predict the probability $\hat{y}_i$ for node n_i:

$$\hat{y}_i = \text{sigmoid}(\mathbf{W_r}\mathbf{n_i} + \text{b}_\text{r}) \tag{6}$$

where $\mathbf{W_r}$ is a weight matrix and b_r is bias. We use binary cross-entropy as loss function:

$$\mathcal{L}_{qrir} = -\sum_{i}^{L+T}[y_i \cdot log(\hat{y}_i) + (1 - y_i) \cdot log(1 - \hat{y}_i)] \tag{7}$$

Meanwhile, we aggregate the video information by multiplying the node representations $n_v = \{F, S\}$ with the question-relevant probability y_v:

$$V = n_v \odot y_v \tag{8}$$

The hypotheses representations and aggregated video representations, are then concatenated and fed to the classifier to obtain the logits s for each candidate answer:

$$s = \text{classifier}([\text{H}; \text{V}]) \tag{9}$$

where the classifiers consist of two fully-connected layers, then we use the softmax function to obtain a probability distribution of each candidate answer and apply cross-entropy loss:

$$\hat{y}^{qa} = \text{softmax}(\text{s}) \tag{10}$$

$$\mathcal{L}_{qa} = -\sum_{i=1}^{N} y_i^{qa} log \hat{y}_i^{qa} \tag{11}$$

Finally, the total loss is:

$$\mathcal{L} = \mathcal{L}_{qrir} + \mathcal{L}_{qa} \tag{12}$$

Experiments

In this section, we evaluate the proposed method with previous works on two multi-choice video question answering datasets that contains character dialogues. We use standard Train/Val/Test-public splits and accuracy to measure the performance.

Model	Val(%)	Test-public(%)						
		BBT	Fridens	Himym	Grey	House	Castle	All
multi-stream (Lei et al. 2018)	65.85	70.25	65.78	64.02	67.20	66.84	63.96	66.46
PAMN (Kim et al. 2019b)	66.38	-	-	-	-	-	-	66.77
Multi-task (Kim et al. 2019a)	66.22	-	-	-	-	-	-	67.05
CA-RN (Geng et al. 2020)	68.90	71.43	65.78	67.20	70.62	69.10	69.14	68.77
STAGE (Lei et al. 2020)	70.50	-	-	-	-	-	-	70.23
MSAN (Kim et al. 2020)	70.79	-	-	-	-	-	-	71.13
DenseCap (Kim, Tang, and Bansal 2020)	74.20	74.04	73.03	74.34	73.44	74.68	74.86	74.09
iPerceive (Chadha, Arora, and Kaloty 2020)	76.97	75.32	74.22	75.14	74.42	75.22	75.77	75.15
Vx2Text (Lin et al. 2021)	74.90	-	-	-	-	-	-	75.00
SS-CRL (Kim et al. 2021)	76.23	77.43	73.24	76.72	74.04	76.94	77.86	76.15
IKEIR (Ours)	**77.82**	**79.06**	**75.69**	**78.30**	**76.05**	**78.07**	**79.38**	**77.88**

Table 2: Comparison with the Non-Multi-Modal Pre-trained models on TVQA dataset.

Model	Test-public(%)	Pre-training Dataset
HERO (Li et al. 2020)	74.80	TV + HowTo100M
DuKG (Li, He, and Feng 2021)	75.45	TV + HowTo100M
IKEIR (Ours)	77.88	/
MERLOT (Zellers et al. 2021)	78.70	YT-Temporal-180M
FrozenBiLM (Yang et al. 2022a)	82.00	WebVid10M
MERLOT RESERVE (Zellers et al. 2022)	86.10	YT-Temporal-1B

Table 3: Comparison with the Multi-Modal Pre-trained models, which first use large-scale multi-modal datasets for pre-training and then fine-tuned on TVQA datasets.

Model	Test(%)
ROCK (Garcia et al. 2020b)	65.40
ROLL (Garcia and Nakashima 2020b)	71.50
KTL (Wu et al. 2021)	73.10
DiagSumQA (Engin et al. 2021)	78.10
i-Code (w/o Pre-training) (Yang et al. 2022b)	78.10
i-Code (w/ Pre-training) (Yang et al. 2022b)	80.50
IKEIR (Ours)	**81.40**

Table 4: Comparison with the state-of-the-art methods on KnowIT VQA dataset. Notably, our method outperforms i-Code (w/ Pre-training), which is first pre-trained on the YT-Temporal-180M dataset and then fine-tuned on the KnowIT VQA dataset.

Datasets

TVQA TVQA is a widely used multi-choice video question answering dataset. It is constructed based on six TV shows: The Big Bang Theory, How I Met Your Mother, Friends, Grey's Anatomy, House, and Castle. It consists of 152,545 human-written QA pairs from 21,793 clips (60-90 seconds), spanning over 460 hours of video, covering various real-life scenarios. Each video clip is associated with 7 questions, with 5 candidate answers (1 correct) for each question. The questions are designed to be compositional, requiring systems to jointly localize relevant moments within a clip, comprehend subtitles-based dialogue, and recognize relevant visual concepts.

KnowIT VQA KnowIt VQA is another popular multi-choice video question answering dataset. It contains 24,282 human-generated question-answer pairs. Each question with 4 candidate answers (1 correct). The dataset combines visual, textual, and temporal coherence reasoning with knowledge-based questions.

Implementation Details

We select ResNet-101 trained on ImageNet as the vision encoder and BERT as the language encoder. Then dimensions of the visual and linguistic feature d^v and d^l are set to 2048 and 768, respectively. For Integrated Reasoning, we set batch size as 16 and use AdamW optimizer with an initial learning rate of 0.00005. About the Inferential Knowledge Reasoner, the model parameters are initialized with the pre-trained parameters from the GPT. We set batch size as 128, and use AdamW optimizer with an initial learning rate of 0.00005, and set beam search size as 5, and set the maximum decoding step as 35.

Comparison with State-of-the-Arts

We compare our method with other state-of-the-art methods on TVQA and KnowIT VQA datasets. Previous work can be split into two categories according to the model type: (1) Non-Multi-Modal Pre-trained Models, which are not pre-trained on extra large-scale multi-modal data. (2) Multi-

Modal Pre-trained Models, which are first pre-trained on large-scale multi-modal data and then fine-tuned on the target datasets. Our model belongs to the first category.

Result on TVQA For TVQA, our method achieves state-of-the-art against Non-Multi-Modal Pre-trained models and achieves competitive performance with Multi-Modal Pre-trained models.

Comparison with Non-Multi-Modal Pre-trained Models As Table 2 shows, our method outperforms the previous Non-Multi-Modal Pre-trained state-of-the-art models on TVQA val and test sets and gets the best performance across all 6 TV shows.

Comparison with Multi-Modal Pre-trained Models As Table 3 shows, we also compare our method with Multi-Modal Pre-trained models, which first use large-scale multi-modal data for pre-training and then fine-tuned on TVQA datasets. Prior work has demonstrated that amazing improvements by increasing datasets size for pre-training. Li et al. (Li et al. 2020) proposed a transformer-based hierarchical model and use HowTo100M (Miech et al. 2019) datasets for pre-training. Li et al. (Li, He, and Feng 2021) incorporate a large-scale image-text pre-trained model CLIP (Radford et al. 2021) into HERO. Zellers et al. (Zellers et al. 2021) used YT-Temporal-180M (Zellers et al. 2021) for pre-training to learn multimodal script knowledge. Zellers et al. (Zellers et al. 2022) introduced a new datasets, named YT-Temporal-1B (Zellers et al. 2022), and used it for pre-training.

- **HowTo100M**, which contains 136 million video clips with captions sourced from 1.2 million Youtube videos.
- **YT-Temporal-180M**, which contains 180 million segments from 6 million public YouTube videos.
- **YT-Temporal-1B**, which contains 20 million videos and 1 billion frames.
- **WebVid10M**, which contains 10 million of video-text pairs scraped from the Shutterstock website.

It is worth noting that our model still achieves competitive performance without pre-trained on any extra large-scale multi-modal data.

Result on KnowIT VQA As Table 4 shows, our method achieves state-of-the-art on the KnowIT VQA dataset. It is worth noting that our method outperforms i-Code (Yang et al. 2022b), which first uses a large-scale multi-modal dataset (YT-Temporal-180M) for pre-training and then fine-tuned on the KnowIT VQA dataset.

Ablation Study

Inferential Knowledge In this part, we explore the effectiveness of inferential knowledge. Figure 4 shows an example of inferential knowledge generated by the Inferential Knowledge Reasoner. The inferential knowledge reveals deeper semantics that we humans know but is not explicitly present in the text. With the help of the knowledge, the model can better understand the linguistic modality information as humans do. As Table 5 shows, without inferential knowledge, the model performance decreased, demonstrating the superiority of inferential knowledge on linguistic modality information understanding.

Model	TVQA Val(%)
IKEIR w/o KE-VSG	75.83
IKEIR w/o Inferential Knowledge	76.54
IKEIR w/o QRIR	76.68
IKEIR (Ours)	**77.82**

Table 5: Ablation studies of different components. The Integrated Reasoning Mechanism consists of a Knowledge-Enhanced Video Semantic Graph (KE-VSG) and Question-Relevant Information Recognition task (QRIR).

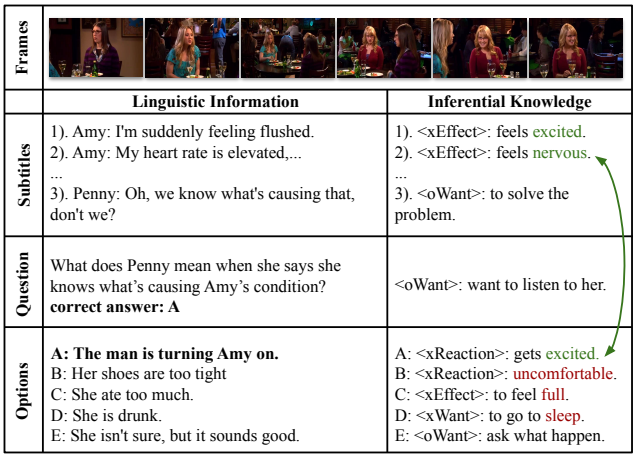


Figure 4: An example of the TVQA dataset. The generated inferential knowledge helps the model better understand the video content and predict the answer.

Integrated Reasoning Mechanism We proposed two key improvements for video multi-modal information modeling and answer reasoning in the integrated reasoning mechanism: (1) Knowledge-Enhanced Video Semantic Graph (KE-VSG). (2) Question-Relevant Information Recognition. To explore the effectiveness of the two key improvements, we remove them from our model, respectively. As Table 5 shows, when removing the Knowledge-Enhanced Video Semantic Graph, the IKEIR models all the information with the general paradigm (Fig. 3). We find that the performance decreased, demonstrating the superiority of KE-VSG on inter and intra-modality information modeling. The Question-Relevant Information Recognition task helps the model recognize and aggregate the question-relevant video information to answer the question. When removing it, the performance also decreased.

Related Work

Video Question Answering

Video question answering can be split into two categories according to the modality of video information needed to answer the given question. (1) VQ-VideoQA (Jang et al. 2017; Xu et al. 2017), in which the question can be answered by visual modality information. (2) VLQ-VideoQA (Lei et al. 2018; Garcia et al. 2020a), where both the visual and linguistic modality information is needed for answering the questions. Our work is relevant to the VLQ-VideoQA task.

VQ-VideoQA VQ-VideoQA demands a deep understanding of video visual modality information and spatio-temporal information. To tackle this problem, Jang et al. (Jang et al. 2017) used LSTMs for sequential learning and proposed the Spatio-Temporal attention mechanism to recognize the question-relevant visual information. Gao et al. (Gao et al. 2018) proposed a co-memory network to learn the correlation among the motion, appearance, and question for answer prediction. Fan et al. (Fan et al. 2019) proposed a heterogeneous memory network to learn global context information from appearance and motion, and design a question memory to understand the complex semantics of the question. Huang et al. (Huang et al. 2020) proposed a location-aware graph to model the location and relations among object interactions that occurred in videos. Le et al. (Le et al. 2020) used a reusable Conditional Relation Network to construct more sophisticated structures for representation and reasoning over video. Mao et al. (Mao et al. 2022) proposed a dynamic multistep reasoning mechanism to predict the answer based on structural video semantic representations.

VLQ-VideoQA Compared to the VQ-VideoQA, VLQ-VideoQA is more challenging, which requires the system to understand the visual modality information and raises higher demands on a deep understanding of linguistic modality information. The video linguistic modality information refers to the subtitles-based dialogue. Prior works on VLQ-VideoQA mainly adopted the paradigm of cross-modal encoding and matching (Lei et al. 2018, 2020; Kim et al. 2019b, 2021). Most of the following researchers focused on optimizing video and language understanding (Kim, Tang, and Bansal 2020; Li et al. 2020; Zellers et al. 2021; Garcia and Nakashima 2020a). Specifically, for the understanding of visual modality, Lei et al. (Lei et al. 2018) used LSTMs to encode the multi-modal information and adopted context-matching attention to fuse the features for answer prediction. Kim et al. (Kim, Tang, and Bansal 2020) resorted to object detection (Ren et al. 2015) and dense image captioning (Johnson, Karpathy, and Fei-Fei 2016) to obtain a fine-grained understanding of visual modality information. Chadha et al. (Chadha, Arora, and Kaloty 2020) explored an unsupervised visual causal inference method to correct common-sense errors while generating dense image captioning. Garcia et al. (Garcia and Nakashima 2020a) adopted symbolized video scene graphs to model the visual semantic information. For the understanding of linguistic modality, Engin et al. (Engin et al. 2021) retrieved video-relevant plot summary as external knowledge. Kim et al. (Kim et al. 2021) adopted contrastive learning to promote the representation learning of linguistic modality. Recently, large-scale multi-modal pre-trained models (Li et al. 2020; Li, He, and Feng 2021; Zellers et al. 2021; Yang et al. 2022a; Zellers et al. 2022; Yang et al. 2022b) have achieved significant progress in unified representation learning by leveraging the consistency and complementary of different modality. In addition, Lin et al. (Lin et al. 2021) explored a unified model to learn knowledge of different modalities and used a generative method to solve multi-choice video question answering tasks.

Unlike previous methods, our approach has two distinctive characteristics: (1) It enhances the understanding of linguistic modality by imitating the inference behavior of humans. (2) It explores an Integrated Reasoning Mechanism to enhance video content understanding by leveraging inferential knowledge, and alleviate the problem of local information bias by graph message passing mechanism.

Knowledge-Enhanced Reasoning

Various studies have assessed the efficacy of external knowledge in natural language processing tasks, such as commonsense question answering (Chen et al. 2020) and machine reading comprehension (Pan et al. 2019; Qiu et al. 2019). Researchers have also introduced external knowledge in other tasks such as language generation (Ji et al. 2020). Song et al. (Song et al. 2021) retrieved entity-based knowledge from ConceptNet (Speer, Chin, and Havasi 2017) for visual commonsense reasoning. Garcia et al. (Garcia and Nakashima 2020a) retrieved video-relevant plot summary as external knowledge in a weakly supervised fashion for video question answering.

Unlike previous methods that mainly adopted taxonomic knowledge, our approach explores inferential knowledge to enhance the understanding of linguistic modality information.

Conclusion

This paper proposes the Inferential Knowledge-Enhanced Integrated Reasoning method for video question answering. First, we design the Inferential Knowledge Reasoner to generate inferential knowledge for linguistic inputs by imitating the inference behavior of humans when watching videos. The inferential knowledge reveals the implicit semantics, including the causes and effects of the event and the character's mental states. Then, we propose the Integrated Reasoning Mechanism to enhance the video content understanding and answer reasoning by leveraging the generated inferential knowledge, and alleviate the problem of local information bias by graph message passing mechanism. The experimental results on TVQA and KnowIT VQA datasets demonstrate the superiority of our method. The ablation results further demonstrate the effectiveness of each component of our approach.

Acknowledgements

Our work is supported by the National Natural Science Foundation of China (62276250). This work is also sup-

ported by Baidu and ICTCAS Joint Project. We would like to thank the anonymous reviewers for their valuable feedback.

References

Chadha, A.; Arora, G.; and Kaloty, N. 2020. iPerceive: Applying common-sense reasoning to multi-modal dense video captioning and video question answering. *arXiv preprint arXiv:2011.07735*.

Chen, Q.; Ji, F.; Chen, H.; and Zhang, Y. 2020. Improving Commonsense Question Answering by Graph-based Iterative Retrieval over Multiple Knowledge Sources. In *Proc. of the 28th International Conference on Computational Linguistics*, 2583–2594.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on CVPR*, 248–255. Ieee.

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*.

Engin, D.; Schnitzler, F.; Duong, N. Q.; and Avrithis, Y. 2021. On the hidden treasure of dialog in video question answering. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2044–2053. IEEE Computer Society.

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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1999–2007.

Gao, J.; Ge, R.; Chen, K.; and Nevatia, R. 2018. Motion-appearance co-memory networks for video question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6576–6585.

Garcia, N.; and Nakashima, Y. 2020a. Knowledge-based video question answering with unsupervised scene descriptions. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16*, 581–598. Springer.

Garcia, N.; and Nakashima, Y. 2020b. Knowledge-based video question answering with unsupervised scene descriptions. In *European Conference on Computer Vision*, 581–598. Springer.

Garcia, N.; Otani, M.; Chu, C.; and Nakashima, Y. 2020a. KnowIT VQA: Answering knowledge-based questions about videos. In *Proc. of the AAAI*, volume 34, 10826–10834.

Garcia, N.; Otani, M.; Chu, C.; and Nakashima, Y. 2020b. KnowIT VQA: Answering Knowledge-Based Questions about Videos. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence*.

Geng, S.; Zhang, J.; Fu, Z.; Gao, P.; Zhang, H.; and de Melo, G. 2020. Character matters: Video story understanding with character-aware relations. *arXiv preprint arXiv:2005.08646*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proc. of the IEEE conference on CVPR*, 770–778.

Huang, D.; Chen, P.; Zeng, R.; Du, Q.; Tan, M.; and Gan, C. 2020. Location-aware graph convolutional networks for video question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 11021–11028.

Jang, Y.; Song, Y.; Yu, Y.; Kim, Y.; and Kim, G. 2017. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *Proc. of the IEEE conference on CVPR*, 2758–2766.

Ji, H.; Ke, P.; Huang, S.; Wei, F.; and Huang, M. 2020. Generating Commonsense Explanation by Extracting Bridge Concepts from Reasoning Paths. In *Proc. of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, 248–257.

Johnson, J.; Karpathy, A.; and Fei-Fei, L. 2016. Densecap: Fully convolutional localization networks for dense captioning. In *Proc. of the IEEE conference on CVPR*, 4565–4574.

Kim, H.; Tang, Z.; and Bansal, M. 2020. Dense-Caption Matching and Frame-Selection Gating for Temporal Localization in VideoQA. In *Proc. of the 58th ACL*, 4812–4822.

Kim, J.; Ma, M.; Kim, K.; Kim, S.; and Chang, D. Y. 2019a. Gaining Extra Supervision via Multi-task learning for Multi-Modal Video Question Answering. In *2019 IJCNN*.

Kim, J.; Ma, M.; Kim, K.; Kim, S.; and Yoo, C. D. 2019b. Progressive Attention Memory Network for Movie Story Question Answering. In *Proc. of the IEEE/CVF CVPR*.

Kim, J.; Ma, M.; Pham, T.; Kim, K.; and Yoo, C. D. 2020. Modality shifting attention network for multi-modal video question answering. In *Proc. of the IEEE/CVF CVPR*, 10106–10115.

Kim, S.; Jeong, S.; Kim, E.; Kang, I.; and Kwak, N. 2021. Self-supervised Pre-training and Contrastive Representation Learning for Multiple-choice Video QA. In *Proc. of the AAAI*, volume 35, 13171–13179.

Le, T. M.; Le, V.; Venkatesh, S.; and Tran, T. 2020. Hierarchical conditional relation networks for video question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9972–9981.

Lei, J.; Yu, L.; Bansal, M.; and Berg, T. 2018. TVQA: Localized, Compositional Video Question Answering. In *Proc. of the 2018 Conference on EMNLP*, 1369–1379.

Lei, J.; Yu, L.; Berg, T.; and Bansal, M. 2020. TVQA+: Spatio-Temporal Grounding for Video Question Answering. In *Proc. of the 58th ACL*, 8211–8225.

Li, G.; He, F.; and Feng, Z. 2021. A CLIP-Enhanced Method for Video-Language Understanding. *arXiv preprint arXiv:2110.07137*.

Li, L.; Chen, Y.-C.; Cheng, Y.; Gan, Z.; Yu, L.; and Liu, J. 2020. Hero: Hierarchical Encoder for Video+ Language Omni-representation Pre-training. In *Proc. of the 2020 Conference on EMNLP*, 2046–2065.

Lin, X.; Bertasius, G.; Wang, J.; Chang, S.-F.; Parikh, D.; and Torresani, L. 2021. VX2TEXT: End-to-End Learning of

Video-Based Text Generation From Multimodal Inputs. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 7001–7011. IEEE Computer Society.

Mao, J.; Jiang, W.; Wang, X.; Feng, Z.; Lyu, Y.; Liu, H.; and Zhu, Y. 2022. Dynamic Multistep Reasoning based on Video Scene Graph for Video Question Answering. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 3894–3904.

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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2630–2640.

Pan, X.; Sun, K.; Yu, D.; Chen, J.; Ji, H.; Cardie, C.; and Yu, D. 2019. Improving Question Answering with External Knowledge. In *Proc. of the 2nd Workshop on Machine Reading for Question Answering*, 27–37.

Qiu, D.; Zhang, Y.; Feng, X.; Liao, X.; Jiang, W.; Lyu, Y.; Liu, K.; and Zhao, J. 2019. Machine reading comprehension using structural knowledge graph-aware network. In *Proc. of the 2019 Conference on EMNLP-IJCNLP*, 5896–5901.

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 *International Conference on Machine Learning*, 8748–8763. PMLR.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28: 91–99.

Sap, M.; Le Bras, R.; Allaway, E.; Bhagavatula, C.; Lourie, N.; Rashkin, H.; Roof, B.; Smith, N. A.; and Choi, Y. 2019. Atomic: An atlas of machine commonsense for if-then reasoning. In *Proc. of the AAAI*, volume 33, 3027–3035.

Song, D.; Ma, S.; Sun, Z.; Yang, S.; and Liao, L. 2021. Kvlbert: Knowledge enhanced visual-and-linguistic bert for visual commonsense reasoning. *Knowledge-Based Systems*, 230: 107408.

Speer, R.; Chin, J.; and Havasi, C. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proc. of the Thirty-first AAAI*.

Tapaswi, M.; Zhu, Y.; Stiefelhagen, R.; Torralba, A.; Urtasun, R.; and Fidler, S. 2016. Movieqa: Understanding stories in movies through question-answering. In *Proc. of the IEEE conference on CVPR*, 4631–4640.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in neural information processing systems*, 5998–6008.

Wu, T.; Garcia, N.; Otani, M.; Chu, C.; Nakashima, Y.; and Takemura, H. 2021. Transferring Domain-Agnostic Knowledge in Video Question Answering. *arXiv preprint arXiv:2110.13395*.

Xu, D.; Zhao, Z.; Xiao, J.; Wu, F.; Zhang, H.; He, X.; and Zhuang, Y. 2017. Video question answering via gradually refined attention over appearance and motion. In *Proceedings of the 25th ACM international conference on Multimedia*, 1645–1653.

Yang, A.; Miech, A.; Sivic, J.; Laptev, I.; and Schmid, C. 2022a. Zero-Shot Video Question Answering via Frozen Bidirectional Language Models. *arXiv preprint arXiv:2206.08155*.

Yang, Z.; Fang, Y.; Zhu, C.; Pryzant, R.; Chen, D.; Shi, Y.; Xu, Y.; Qian, Y.; Gao, M.; Chen, Y.-L.; et al. 2022b. i-Code: An Integrative and Composable Multimodal Learning Framework. *arXiv preprint arXiv:2205.01818*.

Zellers, R.; Lu, J.; Lu, X.; Yu, Y.; Zhao, Y.; Salehi, M.; Kusupati, A.; Hessel, J.; Farhadi, A.; and Choi, Y. 2022. MERLOT Reserve: Neural Script Knowledge through Vision and Language and Sound. *arXiv preprint arXiv:2201.02639*.

Zellers, R.; Lu, X.; Hessel, J.; Yu, Y.; Park, J. S.; Cao, J.; Farhadi, A.; and Choi, Y. 2021. MERLOT: Multimodal Neural Script Knowledge Models. *arXiv preprint arXiv:2106.02636*.