Language Models are Causal Knowledge Extractors for Zero-shot Video Question Answering

Anonymous CVPR submission

Paper ID 40

Abstract

Causal Video Question Answering (CVidQA) queries not only association or temporal relations but also causal relations in a video. Existing question synthesis methods pretrained question generation (QG) systems on reading comprehension datasets with text descriptions as inputs. However, QG models only learn to ask association questions (e.g., "what is someone doing...") and result in inferior performance due to the poor transfer of association knowledge to CVidQA, which focuses on causal questions like "why is someone doing ...". Observing this, we proposed to exploit causal knowledge to generate question-answer pairs, and proposed a novel framework, Causal Knowledge Extraction from Language Models (CaKE-LM), leveraging causal commonsense knowledge from language models to tackle CVidQA. To extract knowledge from LMs, CaKE-LM generates causal questions containing two events with one triggering another (e.g., "score a goal" triggers "soccer player kicking ball") by prompting LM with the action (soccer player kicking ball) to retrieve the intention (to score a goal). CaKE-LM significantly outperforms conventional methods by 4% to 6% of zero-shot CVidQA accuracy on NExT-QA and Causal-VidQA datasets. We also conduct comprehensive analyses and provide key findings for future research.

1. Introduction

Video Question Answering (VidQA), which queries about a video clip with a natural language question, is a fundamental task connecting natural language processing with computer vision [29, 35, 41]. VidQA requires QA systems to understand both natural language questions and their corresponding visual contents, and further figure out the relations between entities or actions. Recent studies have advanced beyond VidQA to focus on Causal Video Question Answering (CVidQA), which focuses on **causal relations** between events [18, 32], where one event triggers another event. For example, "*why is the man in the video running*" asks about not only entity (*i.e.*, "man"), action (*i.e.*, "running"), and temporal relation (*i.e.*, what *happened before* "running"), but also the *intention* (*i.e.*, what *caused* "running"). These questions are more challenging because the model needs to distinguish causation from association for intention understanding. Recent VidQA methods [3, 33, 34] require the high annotation quality and massive quantity of question-answer pairs to exhaust the causal relations. However, collecting the annotations about causality is expensive, and it is difficult to cover all the scenarios in the ever-changing world. Therefore, improving the quality and quantity of QA annotations becomes a critical bottleneck for training CVidQA systems.

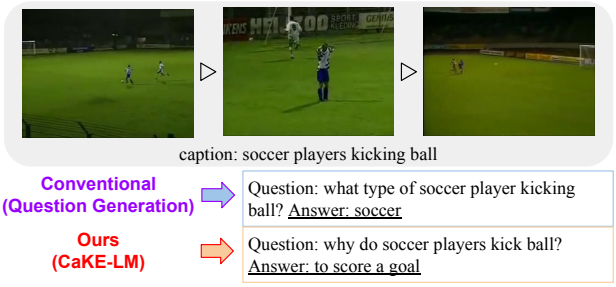

Figure 1. **Motivation.** Conventional question generation methods ask *association* questions (*e.g.*, "what type of playing kicking ball") that focus on entity linking, event recognition, or relation detection, which can hardly adapt to Causal Video Question Answering. Our CaKE-LM acquires *causal* commonsense knowledge by prompting language models (*e.g.*, "why do soccer players kick ball").

Recent VidQA studies tackled the annotation issue by utilizing question generation systems [11, 23] to automatically synthesize training question-answers based on video descriptions [19, 35, 37, 38, 41]. Among these works, Yang et al. [37] leveraged the SQuAD dataset [24] to train the question generator (QG) with implicit factoid QA style. Given captions as inputs, QG models generate *association* questions such as "*what type of player kicking ball?*" However, association knowledge does not imply causal knowledge and results in poor adaptability to CVidQA. Although QA pairs

generated by conventional question generation strategies improve the performance of VidQA models by large margins, these models failed to achieve improvements on causality-related questions as good as other types of question (*i.e.*, 6% lower than the overall accuracy, as shown in Table 1). This observation motivates us to ask the following question: *what is the required knowledge for causal video question answering, and how to obtain it*?

In this paper, we aim to generate question-answer data containing causal commonsense knowledge for CVidQA, which pinpoints the event by understanding cross-event causal relations. Unlike conventional association knowledge that focuses on entity linking, event recognition, and relation detection, causal commonsense knowledge addresses the intention-action relation across events, where one event (intention) triggers another (action), as shown in Figure 1. Causal commonsense knowledge stems from massive and general observation of events and cause-effect mapping between them. Thanks to the success of large-scale pre-training on tremendous and domain-generic language data, language models are witnessed to capture general knowledge from immense data and can be utilized in understanding chat-bot user intentions with dialogues [26, 27, 42] or classify cause-effect relations [12]. However, unlike supervised trained question generators, LMs do not provide a direct interface to map video descriptions to question-answers for CVidQA training. Therefore, extracting causal knowledge from language models for CVidQA still presents a challenge.

To tackle this challenge, we proposed a question-answer generation framework, Causal Knowledge Extraction from Language Models (CaKE-LM), to extract causal commonsense knowledge from LMs for causal video question answering. Noticing that a CVidQA question typically consists of two events with a causal relation, *e.g.*, an event X "*soccer players kicking ball*" and another event Y "*to score a goal*" that motivates the event X can be transferred to a QA pair: *why do soccer players kick ball?* and *A: to score a goal.* Observing this structure, we utilize the video description as event X and acquire event Y by prompting LMs for the intention of event X. Subsequently, we convert events X and Y into a question and answer, respectively, and generate distractors by sampling from other answers to create a multiple-choice question for CVidQA training.

We further conducted experiments on two large-scale CVidQA datasets, NExT-QA [32], and Causal-VidQA [18]. Experimental results demonstrate 4% to 6% of accuracy improvement on zero-shot causal questions. We also explore several key findings: (1) Causal knowledge for CVidQA is distillable between LMs with a straightforward approach. (2) Smaller LMs (GPT-Neo) with significantly fewer parameters (below 1.0% vs. GPT-3) are capable of few-shot CVidQA (less than 1.0% accuracy gap), and (3) Bridging the information gap between videos and text descriptions will potentially further improve CVidQA performance. These experimental results and findings suggest a novel perspective for leveraging LMs for visual reasoning tasks.

Our main contributions are summarized as follows.

- The first use of causal commonsense knowledge from LMs for zero-shot CVidQA.
- A novel framework, CaKE-LM, for extracting knowledge from LMs and generating QAs for CVidQA training by decomposing QA and prompting LMs.
- Improving results of up to 6% compared to traditional methods.
- Comprehensive analysis and key findings for future research.

2. Related Work

2.1. Causal Video Question Answering

Video question answering (VidQA) has been a crucial multi-modal task bridging natural language processing and computer vision. Early days VidQA [35, 41] mainly focused on querying objects or actions according to referential or spatial relations. Subsequent VidQA datasets [16, 17, 40] stepped forward to temporal relations of successive events. Compared to above datasets, Causal Video Question Answering (CVidQA) [18, 32] is more challenging due to the requirement of understanding causation beyond temporal association. Thanks to the large-scale datasets, several recent models [12, 33, 34] were proposed by graph reasoning with object-level representation. However, the collection of high quality and quantity datasets obstacles the employment of CVidQA models in the ever-changing world. Therefore, an adaptive and scalable data generation method is necessary for CVidQA application development.

2.2. QA Generation for Video QA

Automatic QA generation appears to be a desired solution for CVidQA applications as it avoids the prohibitive cost. Although several prior works [28, 30] have generated question-answer pairs with videos, a major limitation is that they require annotated QA pairs to train the question generation systems. Therefore, many researchers utilize descriptions associated with videos and text question generation systems. Several datasets [35, 41] were collected in this manner by rule-based QA generation systems [11], but the diversity of QA was limited by pre-defined templates. Recent studies [37, 38] have employed neural question generation models, such as T5 [23], which have been pre-trained on large-scale human-labeled datasets (e.g., SQuAD [24]), resulting in significant performance gains over traditional rule-based methods. Nevertheless, this approach is limited by the

patterns present in the pre-training dataset, which are primarily associative, resulting in poor adaptability to CVidQA applications. Unlike previous work, our approach extracts causal commonsense knowledge from language models that are not explicitly trained on a specific dataset. Consequently, our method excels in generating CVidQA QAs without relying on human-annotated data or model fine-tuning.

2.3. Language Models Adaptation

LMs pre-trained with causal language modeling CLM, which predicts the next word based on the previous context, are fundamental backbones of natural language processing. Large-scale Language Models (LLMs) [1, 2, 22] leverage the self-supervised property of CLM and vast text data on the web to acquire superior general reasoning capability. Recent studies suggested that LMs could be utilized to obtain procedural knowledge [20], understand user intention in dialogues [26, 27, 42] or learn from context to predict future events [31]. Another line of research [12] fine-tuned LMs to tackle cause-effect classification task. Inspired by recent studies in NLP, we leverage LMs for multi-modal CVidQA tasks by generating questions with prompting. LMs have been utilized for knowledge VQA tasks [8, 39] by serving as an external knowledge base. These methods used in-context learning to perform knowledge extraction with few examples. We move a step forward to tackle CVidQA by acquiring causal commonsense knowledge.

3. Approach

Causal Knowledge Extraction from Language Models (CaKE-LM) aims to extract causal commonsense knowledge for CVidQA by inquiring about intentions based on provided events from LMs without the need for data allocation or model fine-tuning, as shown in Figure 2. CaKE-LM is composed of Knowledge Source (Section 3.1), Knowledge Extraction by prompting (Section 3.2) and Question Generation (Section 3.3). We also investigate the knowledge transferability between different LMs by Distillation with LM Answer (Section 3.4).

3.1. Causal Knowledge from Language Models

Intuitively, LMs acquire causal commonsense knowledge through training on causal language modeling (CLM) objective, as understanding the cause-and-effect relationships improves next word prediction. Recent literature also shows the causal reasoning ability of LMs carrying causal commonsense knowledge for downstream tasks. For example, [13] utilized GPT-3 for intention prediction based on the causal relation, *e.g.*, X gets X's car repaired *because* X wanted $\rightarrow$ to maintain the car. This aligns with our objective of acquiring knowledge for CVidQA in the format of "Q: why is X getting X's car repaired? A: to maintain the car." As such, we utilize Language Models (LMs) as our source of causal knowledge.

3.2. Knowledge Extraction by Prompting

Different from conventional question generators, LMs are trained with CLM and do not provide an interface to generate questions for CVidQA, and fine-tuning the model requires further data annotation and is not scalable. Therefore, we cope with this challenge by decomposing a CVidQA question into two causal-associated events E_x and E_y, where E_x is triggered by E_y. Next, for a video Vid, we leverage the associated caption Cap as E_x, and inquire E_y by prompting for n possible intentions of E_x based on causal knowledge of the language model

$$Response = LM(Prompt(Cap)), \tag{1}$$

where $Prompt$ is the prompt function and $Responses$ represents the response generated by LM. To investigate LM behaviors, we define zero-shot and few-shot prompts as follows:

Zero-shot Prompting. GPT-2 demonstrated a noteworthy ability for zero-shot conversational question answering [25], outperforming several fully-supervised models [22] even though it was not pre-trained on this particular QA format. This motivates us to prompt the LMs in a question format of *"what is the intention of* $\{Cap\}$*?"* in order to extract causal knowledge from LMs.

Few-shot Prompting. Recent research [8, 39] has demonstrated that using in-context learning with a limited number of examples presented in the prompt to language models can yield promising results. We take advantage of this concept and define the few-shot prompting with k example inputs $Input = \{I_1, I_2 \cdots I_k\}$ and outputs $Output = \{O_1, O_2 \cdots O_k\}$ as:

"Input: $\{I_1\}$
Output: $\{O_1\}$
$\cdots$
Input: $\{I_k\}$
Output: $\{O_k\}$
Input: $\{Cap\}$
Output:".

3.3. Question-Answer Generation

With the caption Cap and the causal-associated response $Response$, we then generate the questions in multi-choice format with a question, the answer, and other options as the distractors. As $Response$ is the LM predicted intention of Cap, we utilize $Response$ as the correct answer Ans and transfer Cap from declared sentence to interrogative sentence to obtain the question. Specifically, we first randomly select a question prefix Pre from *why is, why did, why does*, and concatenate the prefix with the caption

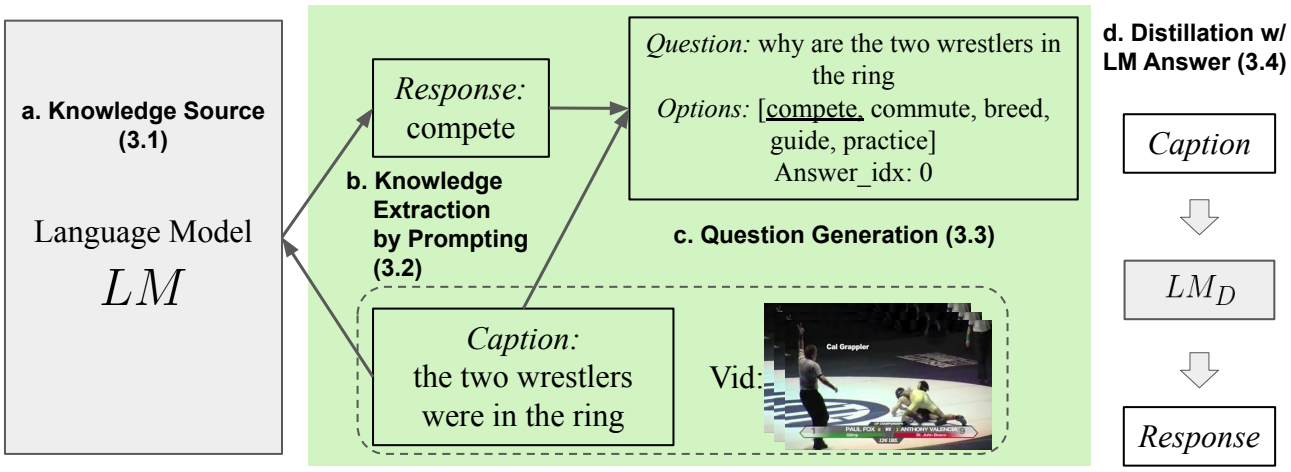

Figure 2. **CaKE-LM.** To acquire causal commonsense knowledge from LMs (a., Section 3.1), we prompt the LM to retrieve potential intentions (b., 3.2). Finally, the LM response and the caption are transformed into a multi-choice question for CVidQA (c., Section 3.3). We also investigate the potential of transferring this knowledge to another LM (d., Section 3.4).

as $Q_0 = \{Pre\}\{Cap\}$. Then, we pass Q_0 into a grammar correction system[1] GC and generate corrected question $Q = GC(Q_0)$.

Next, to obtain contextually similar distractors, we cluster all responses into option pools and sample the distractors of an answer from the pool where the answer belongs to. In particular, we cluster responses $\{response_1, response_2 \cdots response_{|R|}\}$ into $|P|$ pools according to their RoBerTa[2] [21] features:

$$remb_i = RoBerTa(response_i), \tag{2}$$

$$P_1, P_2 \cdots P_{|P|} = Cluster(remb_1, remb_2, \cdots remb_{|R|}), \tag{3}$$

where P_i represents the i^{th} pool.

Next, for an answer Ans belongs to pool p_j, we sample $|D|$ distractors $d_1, d_2, \cdots, d_{|D|}$ from p_j. The answers and the distractors $Ans, d_1, d_2, \cdots, d_{|D|}$ are then merged and shuffled as $Options$. Finally, we have $Vid, Q, Options$ to train models for CVidQA.

3.4. Distillation with LM Answer

Since LMs may include redundant information for CVidQA that increases computation cost, and future research might want to explore the opportunity of combining knowledge from multiple LMs, we further investigate the feasibility of distilling causal knowledge from one LM to another. Therefore, we inspect a straightforward way to distill LM knowledge: Training with LM-generated responses. Specifically, for a caption Cap and an response $Response$, we train another LM LM_D and the mapping from Cap to $Response$:

$$Response = LM_D(\theta, Cap) \tag{4}$$

[1]https://huggingface.co/vennify/t5-base-grammar-correction

[2]https://huggingface.co/roberta-large

In practice, LM_D can have significantly fewer parameters and knowledge from multiple LMs. Future research is encouraged to explore effective methods for distilling knowledge from multiple sources to enhance performance.

4. Experiments

4.1. Setup

Datasets Our approach is evaluated on two large-scale datasets: NExT-QA with 8,564 test questions[3] [32] and Causal-VidQA [18] with 10,760 test questions[4]. Both datasets include both causal and non-causal questions and we primarily focus on the performance of causal questions. For NExT-QA, we mainly examine on *Causal* category, including *why* (e.g., *why was the toddler in red crying at the end of the video*) and *how* (e.g., *how did the lady help the toddler who fell at the end?*). For Causal-VidQA, *Explanation* (e.g., *why did [p1] hold tight on the rope*), *Prediction* (e.g., *where will [p2] go*), and *Counterfactual* (e.g., *what would happen if the rope broke*) are considered.

Video QA Models We evaluate our approach by training Video QA models with the QAs generated by CaKE-LM pipeline. To test the robustness of our approach, we evaluate it using multiple models, including non-graph-based CoMem [6] and graph-based HGA [14]. CoMem generates multi-level contextual representations using appearance features (CNN) and motion features (3D-CNN) with the attention mechanism. HGA represents videos and questions as graphs

[3]48% of them are causal questions

[4]75% of them are causal questions. Since we do not have access to the test set, we evaluate our approach on the val set and do not use it for tuning our model parameters or selecting the best checkpoint.

		Causal	Why	How	All	Diff
Zero-shot	Random	20.00	20.00	20.00	20.00	0.00
	Just-Ask [37]	31.87	30.43	35.98	38.38	-6.51
	CaKE-LM + HGA (Ours)	35.28	34.71	**36.89**	34.77	+0.51
	CaKE-LM + CoMem (Ours)	**35.70**	**35.31**	36.81	34.85	+0.85
10% Semi-supervised	HGA	33.01	32.83	33.50	35.64	-2.63
	CoMem	30.63	29.68	33.33	32.70	-2.07
100% Supervised	HGA	47.56	47.85	46.72	49.82	-2.26
	CoMem	44.42	44.01	45.61	46.93	-2.51

Table 1. **NExt-QA results**. "Diff" denotes the performance difference between Causal and All. Our method surpass zero-shot Just-Ask by over 4% in Causal accuracy on the NEXT-QA dataset, while Just-Ask suffers from a notable decline in Causal category. Our method not only outperforms semi-supervised methods but also achieves superior Diff compared to Just-Ask and supervised methods. This suggests a more effective retrieval of causal knowledge for CVidQA. (Section 4.2)

and uses graph convolutional networks for reasoning.

Baselines We compare with the state-of-the-art approach [37] of traditional QA generation. We use the best checkpoint provided by the official implementation[5], which is pre-trained on the HowToVA69M and WebVidVQA3M datasets. We also compare our results with the oracle case, where a small portion (10%) or all (100%) of the training data is used as references.

Video Descriptions We use captions from the MSRVTT [36] dataset. To evaluate the effectiveness of the distillation approach outlined in Section 3.4, we split MSRVTT into GPT-10K (w/ 10,000 captions) and T5-130K (w/ 130,000 captions). We generate answers using a pre-trained LM LM and train another LM LM_D with GPT-10K. We then generate answers using distilled model LM_D with T5-130K. Our full model uses both GPT-10K and T5-130K for CVidQA training.

Language Models We use GPT-2[6] [22], GPT-Neo [1] (1.3B[7] and 2.7B[8]) trained on the Pile [7], and GPT-3[9] [2]. Apart from setting *temperature* to 0.7, max_len to 20, and top_k to 5 [10] , we use the default hyper-parameters. We provide examples of few-shot prompting by randomly sampling 5 QAs from NExT-QA and transferring the question to the declared sentence. For the distillation experiments (Section 3.4), we distill from GPT-3 to T5-large[11] model.

Video QA training In all of our experiments, we followed the NExT-QA [32] video preprocessing method, where we uniformly sampled eight segments of 16 consecutive frames. For visual features, we used Resnet101 [10] pre-trained on ImageNet [4] and inflated 3D ResNeXt-101 [9] pre-trained on Kinetics [15] as our feature extractors. For question and answer features, we pre-trained BERT [5] on our generated training set and extracted QA features adhering to the NExT-QA setting. To adapt an open-ended QA model for multiple-choice QA, we concatenated each candidate answer with the question and optimized the model with Hinge Loss, following the NExT-QA implementation.

For video QA training, we employ the default NeXT-QA implementation[12] with the exception of setting the *patience* in the *ReduceLROnPlateau* to 2 instead of 5, and the maximum number of epochs to 25 instead of 50, as we observed a faster convergence during training. We conduct the training on a single NVIDIA TITAN RTX GPU, and each experiment takes around 18 to 24 hours at most.

4.2. Video QA Performance

Baseline Comparison As shown in Table 1, CaKE-LM significantly outperform state-of-the-art QA generation model Just-Ask [37] by 4% of accuracy on NExT-QA benchmark. CaKE-LM also surpasses the semi-supervised trained model. Note that Just-Ask generated 72 million QAs and was pre-trained with 100K high quality, human-labeled questions (SQuAD). Additionally, Just-Ask uses a more advanced VidQA backbone, outperforming HGA by 4-6% on MSVD-QA and MSRVTT-QA *without* pre-training[13]. Conversely,

[5] https://github.com/antoyang/just-ask

[6] https://huggingface.co/gpt2

[7] https://huggingface.co/EleutherAI/gpt-neo-1.3B

[8] https://huggingface.co/EleutherAI/gpt-neo-2.7B

[9] OpenAI text-davinci-003 API https://openai.com/api/

[10] Our empirical findings indicate that adjusting the max_len and top_k parameters in the prompt for GPT-3 yields favorable results. As such, we utilize the default max_len and top_k settings and use prompts in the format of "what is the intention of {Cap}? Provide $\{top\_k\}$ answers within $\{max\_len\}$".

[11] https://huggingface.co/t5-large

[12] https://github.com/doc-doc/NExT-QA

[13] Just-Ask paper [37] Table 4 and 5

CaKE-LM generates only 140K QAs, and the performance is expected to improve with more data. This showcases the superior causal knowledge extraction capability of CaKE-LM. Additionally, Just-Ask's overall accuracy decreases by 6.5% in the Causal category, indicating challenges in adapting to CVidQA but showing the promising adaptability of CaKE-LM.

	Causal	Why	How	All
GPT-10K	33.45	33.17	34.24	33.19
T5-130K	35.14	35.10	35.27	34.64
GPT-10K + T5-130K	35.28	34.71	36.89	34.77

Table 2. **Ablation study** shows that LM causal knowledge is transferrable from GPT-3 (175B) to T5 (770M) with minimal performance loss (0.13%) despite the significant difference in parameters. (Section 4.2)

Knowledge Transferability Table 2 illustrates the transfer of causal knowledge to another LM by training with LM responses. T5-large (770M) has only 0.44% of GPT-3's (175B) parameters, yet its performance when trained with only T5-130K is only 0.13% worse than GPT-10K + T5-130K. Moreover, training with T5-130K yields better performance than training with GPT-10K alone. This approach not only reduces computation cost and hardware requirements, but also demonstrates the transferability of causal knowledge with such a straightforward way, indicating the potential for distilling knowledge from multiple LMs.

Causal-VidQA Results CaKE outperforms Just-Ask by 6% causal accuracy as shown in Table 3. Causal-VidQA evaluates models' understanding of both the answer (A) and the reason behind it (R) in prediction and counterfactual categories. We compare CaKE and Just-Ask to supervised methods that are trained with only answers for fair comparison. Furthermore, when compared to supervised methods, Just-Ask exhibits less than half the performance in causal categories and a drop in performance of about 1/4 in the description category. These results suggest that while Just-Ask performs well in traditional video QA tasks, it is less effective for CVidQA tasks. Contrarily, CaKE-LM, which uses causal commonsense knowledge, significantly outperforms Just-Ask in Explanation (by 13%) and Counterfactual (by 7%) tasks. Both Just-Ask and CaKE-LM face challenges when it comes to prediction tasks. We hypothesize that intention knowledge obtained from LMs may not transfer easily to prediction tasks, and suggest future research to address this issue by using more diverse prompting techniques, such as using inquiry LM to obtain results according to actions.

4.3. Language Model Analysis

Do We Need Extremely Large LMs? Even smaller language models like GPT-Neo (which has only 1-2% of the parameters of GPT-3) can serve as effective QA generators for CVidQA. As demonstrated in Table 2, GPT-Neo-2.7B (with just 5 examples) outperforms Just-Ask (31.87, see Table 1) in terms of causal performance, achieving a score of 32.52. In contrast, extremely large GPT-3 performs well even with fewer examples provided. In our experiments, all LMs except for GPT-3 experienced a drastic drop in performance by 1.5% to 3%.

Zero-shot vs Few-shot Prompting Table 4 shows that providing non-GPT-3 LMs with few examples improves QA performance. In addition, 1 shot is only comparable and sometimes even worse than 0 shot, while increasing the examples from 1 to 5 notably improves the performance. Nonetheless, GPT-3 does not derive significant benefit from few-shot prompting. This indicates that pre-trained GPT-3 can be prompted by human instructions, whereas other variants of LMs require examples to effectively extract causal commonsense knowledge for CVidQA.

Frequent words in LM answers As shown in Table 5. Few-shot LMs tend to generate common words from input datasets. This tendency may occur due to contextual similarity among examples, such as references to the same entity. For zero-shot LMs, GPT-3 can generate abstract summaries such as "enterain", while non-GPT-3 LMs may produce irrelevant words such as "think,", "like," "question," "know." These irrelevant words come from context-irrelevant answers like "I don't know" or "I mean." (4.4 for more discussion) In addition, GPT-3 generates more verbs in all settings. This finding suggests different behavior among LMs despite similar CVidQA performance.

4.4. Error Analysis

Context Irrelevant Error A LM generates irrelevant text when it fails to understand the prompt, as seen in Figure 3. We hypothesize that the model responds based on the distribution of the training corpus. Context-irrelevant answers can skew data but have related mild impact on performance in multi-choice settings, as they are uncommon in Video QA datasets. However, in open-ended scenarios, these generated responses may pose a challenge and require additional processing.

Non-causal Correlation Error Sometimes LMs generate answers that are relevant to the caption but do not accurately reflect the intention , such as the second example in Figure 3. In this case, the model may assume correlation implies cause-and-effect. This could influence CVidQA training as

	Acc_D	Acc_E	Acc_P			Acc_C			Causal	All
			A	R	AR	A	R	AR		
Random	20.00	20.00	20.00	20.00	4.00	20.00	20.00	4.00	9.33	12.00
Just-Ask	48.11	35.75	29.81	30.55	10.43	35.12	35.97	14.00	20.06	27.07
Ours + HGA	43.21	49.29	26.00	23.99	8.76	41.26	43.22	21.17	**26.41**	30.61
Ours + CoMem	41.95	49.44	23.47	22.84	7.50	40.88	43.30	21.09	**26.01**	30.00
Supervised (HGA)	66.82	64.18	46.27	48.57	27.52	54.51	54.25	35.80	42.50	48.57
Supervised (CoMem)	64.92	62.44	46.60	47.09	27.33	54.21	53.17	34.24	41.34	47.23

Table 3. Our method outperforms Just-Ask in a significant margin on Causal-VidQA dataset. D: Description, E: Explanation, P: Prediction, C: Counterfactual. Causal: D, P, C. A represents the answer and R represents the reason, while AR means correctly outputting both the answer and the reason. (Section 4.2)

LM	0 shot				1 shot				5 shot			
	Cau.	Why	How	All	Cau.	Why	How	All	Cau.	Why	How	All
GPT-2	28.46	27.96	29.89	29.21	28.64	28.68	28.52	29.21	30.09	29.62	31.43	31.41
GPT-Neo-1.3B	29.47	28.89	31.08	30.28	28.49	28.14	29.46	29.27	31.38	30.79	33.05	32.48
GPT-Neo-2.7B	29.53	28.83	31.51	30.41	29.71	29.29	30.91	30.57	**32.52**	31.73	34.76	32.64
GPT-3	**33.45**	33.17	34.24	33.19	**33.65**	33.71	33.48	33.27	**33.79**	33.50	34.59	34.52

Table 4. Cau.: Causal. LMs significantly smaller than GPT-3 also outperform JustAsk (31.87 causal accuracy) with merely 5 examples provided. Also, providing several examples improves LMs. (Section 4.3)

Caption: the man pulls himself up by the ropes

Context Irrelevant Error: is it no longer a matter of that? it is rather a matter of doing what the man would
Non-causal Correlation Error: the man is on the floor
Visual Mismatch Error: climbing

Figure 3. Three errors observed in LM-generated responses. Context Irrelevant Error: generated response is completely unrelated to the input caption. Non-causal Correlation Error: Relevant but not causal response. Visual Mismatch Error: Causal response according to the caption but is not represented in the video. (Section 4.4)

distractors are usually relevant but non-causal. Therefore, addressing this error is crucial for improving performance in CVidQA.

Visual Mismatch Error Sometimes LMs provide a reasonable intention based on the prompt, but the answer is not aligned with the video, as shown in Figure 3. Captions may not contain all the information in a video, making it difficult for LMs to understand the video content accurately. To address this, solutions such as incorporating additional information, like object detection, can be explored. This error is a key challenge in not only CVidQA but in general LM-based visual reasoning, and further research should aim to address it.

	GPT-2	GPT-Neo (1.3B)	GPT-Neo (2.7B)	GPT-3
0 Shot	one think like people man could know question idea	think like question know mean video people game one	think man like question mean one say know people	entertain fun show entertainment creating man inform viewers
1 Shot	man girl output, video playing one baby get boy	man video baby boy game girl play woman playing	man woman boy baby girl video one talking playing	playing singing music game ask show fun making laughing
5 Shot	man girl one video boy playing woman baby get	man play video playing game baby boy get dance	man playing play game baby talking video woman song	singing playing making music showing enjoying fun dancing show
Inputs: man video cartoon talking playing game woman two people singing boy stage				

Table 5. Top 9 generated words of each LM and the top 15 words in input captions. Red: overlapped nouns, Blue: overlapped verbs. (Section 4.3)

4.5. Potential Extensions

While we already demonstrated promising results in challenging Causal Video Question Answering, there are various exciting directions to further improve or extend our work to other tasks. We try to discuss some of them to pave paths for future research.

Visual-aware QA Generation As discussed in Section 4.4, visual mismatch is one of gaps in utilizing LMs. Bridging this gap can not only improve CaKE-LM but also enhance the utility of LMs in visual reasoning tasks as transferring visual signals into text tokens is a common and straightforward practice to leverage LM knowledge. There are two challenges for CVidQA: (1) ensuring the completeness of video descriptions provided to LMs, and (2) evaluating the alignment of LM responses with the video content. Challenge 1 can be addressed by obtaining more information on different levels in video through tools such as object detection, action recognition, or scenegraph generation. These tools can complement captions and provide rich context to guide LMs for more concise responses. Meanwhile, applying a visual-language similarity between generated responses and the videos can help to filter out reasonable but visually mismatched responses and alleviate challenge 2.

Different QA Applications Our study demonstrates the effectiveness of leveraging LMs to tackle the challenging task of CVidQA in VidQA. Our CaKE-LM pipeline can be extended to address other QA applications, such as temporal prediction, by using different prompts or examples. For instance, by prompting LMs with the question *"What could be the result of {Cap}?"*, we can extract LM predictions based on commonsense causal knowledge. We can also prompt LMs not only with the prediction or intention but also with the reason, which not only enhances the performance of the Causal-VidQA dataset but also improves explainability. By prompting from different angles, such as intention, result, reason, or counterfactual, we can extract causal commonsense knowledge from LMs to further improve CVidQA models. LMs can enhance traditional VidQA tasks by providing associations based on commonsense about associated objects or actions. For instance, we can prompt LMs with questions like *"What objects could be present in the video during {Cap}?"* to obtain relevant associations.

5. Conclusion

In this work, we investigate the utilization of causal commonsense knowledge in LMs for zero-shot CVidQA. We propose a novel framework, CaKE, for extracting causal commonsense knowledge from LMs by prompting and generating QAs to train CVidQA models. Results show a 4% to 6% improvement compared to previous methods on two large-scale benchmarks. We also conduct comprehensive analyses and provide key findings for future research.

References

[1] Sid Black, Leo Gao, Phil Wang, Connor Leahy, and Stella Biderman. GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow, Mar. 2021. If you use this software, please cite it using these metadata. 3, 5

[2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020. 3, 5

[3] Anoop Cherian, Chiori Hori, Tim K Marks, and Jonathan Le Roux. (2.5+ 1) d spatio-temporal scene graphs for video question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 444–453, 2022. 1

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 5

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. 5

[6] J. Gao, R. Ge, K. Chen, and R. Nevatia. Motion-appearance co-memory networks for video question answering. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6576–6585, Los Alamitos, CA, USA, jun 2018. IEEE Computer Society. 4

[7] Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020. 5

[8] Liangke Gui, Borui Wang, Qiuyuan Huang, Alex Hauptmann, Yonatan Bisk, and Jianfeng Gao. Kat: A knowledge augmented transformer for vision-and-language. In *NAACL*, 2022. 3

[9] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6546–6555, 2018. 5

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '16, pages 770–778. IEEE, June 2016. 5

[11] Michael Heilman and Noah A Smith. Question generation via overgenerating transformations and ranking. Technical report, Carnegie-Mellon Univ Pittsburgh pa language technologies insT, 2009. 1, 2

[12] Pedram Hosseini, David A. Broniatowski, and Mona Diab. Knowledge-augmented language models for cause-effect relation classification. In *Proceedings of the First Workshop on Commonsense Representation and Reasoning (CSRR 2022)*, pages 43–48, Dublin, Ireland, May 2022. Association for Computational Linguistics. 2, 3

[13] Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. Comet-atomic 2020: On symbolic and neural commonsense knowledge graphs. In *AAAI*, 2021. 3

[14] Pin Jiang and Yahong Han. Reasoning with heterogeneous graph alignment for video question answering. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07):11109–11116, Apr. 2020. 4

[15] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 5

[16] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara Berg. TVQA: Localized, compositional video question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018. 2

[17] Jie Lei, Licheng Yu, Tamara Berg, and Mohit Bansal. TVQA+: Spatio-temporal grounding for video question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8211–8225, Online, July 2020. Association for Computational Linguistics. 2

[18] Jiangtong Li, Li Niu, and Liqing Zhang. From representation to reasoning: Towards both evidence and commonsense reasoning for video question-answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022. 1, 2, 4

[19] Xudong Lin, Gedas Bertasius, Jue Wang, Shih-Fu Chang, Devi Parikh, and Lorenzo Torresani. Vx2text: End-to-end learning of video-based text generation from multimodal inputs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7005–7015, 2021. 1

[20] Xudong Lin, Fabio Petroni, Gedas Bertasius, Marcus Rohrbach, Shih-Fu Chang, and Lorenzo Torresani. Learning to recognize procedural activities with distant supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13853–13863, 2022. 3

[21] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 4

[22] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019. 3, 5

[23] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. 1, 2

[24] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas, Nov. 2016. Association for Computational Linguistics. 1, 2

[25] Siva Reddy, Danqi Chen, and Christopher D. Manning. CoQA: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266, 2019. 3

[26] Andy Rosenbaum, Saleh Soltan, Wael Hamza, Yannick Versley, and Markus Boese. LINGUIST: Language model instruction tuning to generate annotated utterances for intent classification and slot tagging. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 218–241, Gyeongju, Republic of Korea, Oct. 2022. International Committee on Computational Linguistics. 2, 3

[27] Gaurav Sahu, Pau Rodriguez, Issam Laradji, Parmida Atighehchian, David Vazquez, and Dzmitry Bahdanau. Data augmentation for intent classification with off-the-shelf large language models. In *Proceedings of the 4th Workshop on NLP for Conversational AI*, pages 47–57, Dublin, Ireland, May 2022. Association for Computational Linguistics. 2, 3

[28] Hung-Ting Su, Chen-Hsi Chang, Po-Wei Shen, Yu-Siang Wang, Ya-Liang Chang, Yu-Cheng Chang, Pu-Jen Cheng, and Winston H. Hsu. End-to-end video question-answer generation with generator-pretester network. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(11):4497–4507, 2021. 2

[29] Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. MovieQA: Understanding stories in movies through question-answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1

[30] Yu-Siang Wang, Hung-Ting Su, Chen-Hsi Chang, Zhe-Yu Liu, and Winston H. Hsu. Video question generation via semantic rich cross-modal self-attention networks learning. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2423–2427, 2020. 2

[31] Zhenhailong Wang, Manling Li, Ruochen Xu, Luowei Zhou, Jie Lei, Xudong Lin, Shuohang Wang, Ziyi Yang, Chenguang Zhu, Derek Hoiem, et al. Language models with image descriptors are strong few-shot video-language learners. *arXiv preprint arXiv:2205.10747*, 2022. 3

[32] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. NExT-QA: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1, 2, 4, 5

[33] Junbin Xiao, Angela Yao, Zhiyuan Liu, Yicong Li, Wei Ji, and Tat-Seng Chua. Video as conditional graph hierarchy for multi-granular question answering. In *Proceedings of the 36th AAAI Conference on Artificial Intelligence (AAAI)*, pages 2804–2812, 2022. 1, 2

[34] Junbin Xiao, Pan Zhou, Tat-Seng Chua, and Shuicheng Yan. Video graph transformer for video question answering. In *European Conference on Computer Vision*, pages 39–58. Springer, 2022. 1, 2

[35] Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. Video question answering via gradually refined attention over appearance and motion. In *ACM Multimedia*, 2017. 1, 2

[36] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), June 2016. 5

[37] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Just ask: Learning to answer questions from millions of narrated videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1686–1697, 2021. 1, 2, 5

[38] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Learning to answer visual questions from web videos. *IEEE transactions on pattern analysis and machine intelligence*, PP, 2022. 1, 2

[39] Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Yumao Lu, Zicheng Liu, and Lijuan Wang. An empirical study of gpt-3 for few-shot knowledge-based vqa. In *AAAI*, 2022. 3

[40] Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. ActivityNet-QA: A dataset for understanding complex web videos via question answering. 2019. 2

[41] Kuo-Hao Zeng, Tseng-Hung Chen, Ching-Yao Chuang, Yuan-Hong Liao, Juan Carlos Niebles, and Min Sun. Leveraging video descriptions to learn video question answering. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1), Feb. 2017. 1, 2

[42] Haode Zhang, Haowen Liang, Yuwei Zhang, Li-Ming Zhan, Xiao-Ming Wu, Xiaolei Lu, and Albert Lam. Fine-tuning pre-trained language models for few-shot intent detection: Supervised pre-training and isotropization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 532–542, Seattle, United States, July 2022. Association for Computational Linguistics. 2, 3