

DeepMaven: Deep Question Answering on Long-Distance Movie/TV Show Videos with Multimedia Knowledge Extraction and Synthesis

Yi Fung^{1*}, Han Wang², Tong Wang²,
Ali Kebarighotbi², Mohit Bansal², Heng Ji², Prem Natarajan²

¹University of Illinois at Urbana Champaign, ²Amazon Alexa

yifung2@illinois.edu

{wngn, tonwng, alikeba, mobansal, jihj, premknat}@amazon.com

Abstract

Long video content understanding poses a challenging set of research questions as it involves long-distance, cross-media reasoning and knowledge awareness. In this paper, we present a new benchmark for this problem domain, targeting the task of deep movie/TV question answering (QA) beyond previous work’s focus on simple plot summary and short video moment settings. We define several baselines based on direct retrieval of relevant context for long-distance movie QA. Observing that real-world QAs may require higher-order multi-hop inferences, we further propose a novel framework, called the **DEEPMAVEN**, which extracts events, entities, and relations from the rich multimedia content in long videos to preconstruct movie knowledge graphs (movieKGs), and at the time of QA inference, complements general semantics with structured knowledge for more effective information retrieval and knowledge reasoning. We also introduce our recently collected *DeepMovieQA* dataset, including 1,000 long-form QA pairs from 41 hours of videos, to serve as a new and useful resource for future work. Empirical results show the DeepMaven performs competitively for both the new *DeepMovieQA* and the pre-existing *MovieQA* dataset.¹

1 Introduction

Our world tells an evolving story of people, objects, and their interactions. This storytelling may exist in various forms, from textual summaries and spoken dialogues, to accompanying images and videos. Because of its dynamically evolving and multi-media nature, long-distance video question answering on movies/TV shows provides a useful setting for studying the computational understanding of interconnected stories and events that aligns closely with real-world application scenarios. Yet,

current intelligent systems still struggle with adequately processing the rich multimedia content in long videos and fail to answer many common inquiries that humans are interested in. If we ask a virtual assistant about the relationship between two main characters in a well-known movie/TV show, it likely defaults to some null response such as “*hmm... I don’t know this one*”.

The research challenge of long video content understanding and question answering frameworks stems from the need to support information probing on certain specific details over a large multimedia context space. Propagating information across long-distance has been a well-known challenge due to the vanishing gradient and memory loss problem (Hochreiter, 1998). Long videos also can not fit its entire data in-memory typically for end-to-end feature extraction and neural network training. Meanwhile, retrieval-based methodologies, which first find sections in the video relevant to a query through semantic cues from the corresponding dialogues and/or frames, and then return the relevant textual dialogue and/or visual frame context for question answering, tend to narrowly isolate media instances semantically similar to the query inputs. Query inputs are short and succinct in nature, so simply considering data points that match the query semantics will overlook the larger picture behind the selected media instances *i.e.* how they relate to each other as well as to other relevant media components that are initially missed out from retrieval as they involve additional reasoning.

For example, a question about “the hobby interests of *Midge’s husband*” in TV show “The Marvelous Mrs. Maisel” would likely be missed by text retrieval if the dialogue mostly uses the name of *Midge’s husband - Joel*. A question about “*Midge and Joel’s marriage breakdown*” not only needs to identify the point of confrontation between *Midge* and *Joel*, but also other *past, concurrent, or future* events connected to these two person entities,

^{*}Work done as an intern at Amazon Alexa AI.

¹See <https://www.amazon.science/publications> for update of information about code and resources.






Category	Question (Q)	Natural Language Answer (A)	Visual Evidence
Direct 41% (long-distance)	Who bailed Midge out of jail after she was arrested for public nudity? –“The Marvelous Mrs. Maisel”	Susie Meyerson, the manager of the Gaslight, bailed Midge out.	
Cross-Media 33% (text+visual)	Why does Joel punch someone at Gaslight? –“The Marvelous Mrs. Maisel”	That person insulted Maisel during her act. Joel punches him and asserts that Maisel is good.	
Multi-Scene 11% (≥ 2 scene, tot. min. > 1)	How did Midge and Joel’s marriage fall apart? –“The Marvelous Mrs. Maisel”	Joel has been cheating on Midge, with his secretary Penny. One day, Joel got angry at Midge for giving suggestions that ended in a failed stand-up comedy performance for him. He packed up and left Midge.	
Multi-Hop 11% (higher reasoning through missing links/info)	What does the Mayor oversee on the market? –“Les Miserables”	The mayor oversees which vendors are allowed on the market, what they are allowed to sell, how big their space is and what kind of stands should be added to the market. He charges the vendors his fees and have the police talk to vendors who act up.	
Background Knowledge 10% (facts, Wikipedia)	In real history, what ultimately happened to Malcolm X’s friend, singer Sam Cooke? –“One Night in Miami”	He was shot to death by a motel manager for attempting to molest a woman, but his death was controversial and involved conspiracy theories.	

Table 1: Types of data annotated in our new *DeepMovieQA* benchmark, along with their statistics. The category pertains to the critical reasoning used in understanding the question, picking up relevant details from textual and/or visual media in the movie/TV show, and deriving to an answer.

Midge and Joel, relevant to their marriage dynamics, such as infidelity, pent-up frustration, and aftermaths. Previous work tends to focus on very high level questions with answers coming only from text summary, such as whether a character is good/bad, or focus on very superficial questions such as what object is behind this person, whereas we want to enable more interesting QAs such as from mimicking MovieClub conversations.

So in this work, we define a new task of deep movie question answering (**DeepMovieQA**), which gauges the content understanding of long movie/TV show videos that align with generic human inquiries. In contrast to previous work, the DeepMovieQA task consists of questions that naturally arise from watching the full length of a movie/TV show asset (as opposed to plot summary or short minute-long clips), and involve a more challenging set of reasoning summarized in Table 1, which make use of many different data modalities: plot summary, textual dialogues, visual frames, and background knowledge. DeepMovieQA additionally involves localizing the time frames that provide evidence for the question answering.

To tackle DeepMovieQA, we propose a novel framework, *DeepMaven*, which extracts context from multimedia video asset semantically rele-

vant to the queries and generates a coherent answer conditioned on the retrieved context using a transformer-based backbone, with added benefits of transparency and explainability through this two-stage process. Observing that the new DeepMovieQA task involves more challenging types of reasoning, we further leverage a structured approach to incorporate wider context and better handle multi-hop queries, such as bridging the connection between the mention of *Midge’s husband* in a query and the canonical *Joel* entity.

Our key novel contributions can be summarized as follows: 1) we present a new research task of **DeepMovieQA** in the long-distance multimedia video understanding domain that involves reasoning beyond surface-level understanding of short moments and summaries; 2) we present **DEEP-MAVEN**, a novel multimedia approach that leverages complementary information from (local) dialogue passages and visual frames and (wider-context) movieKG subgraph to guide deeper content understanding and question answering, achieving 10+% absolute gain in answer extraction and relative gain in answer generation over single modality baselines; and finally, 3) we contribute a DeepMovieQA dataset that includes 1000 QA annotations, to serve as a new resource for studying

and evaluating this challenging but exciting task.

2 DeepMovieQA Data Collection

Pre-existing video/movie/TV show QA datasets fall into two extremes of either focusing on too narrow and minor scene-specific details such as *an object on the couch* that is irrelevant to the storyline development (Lei et al., 2018), or overly brief and simple details from summaries (Tapaswi et al., 2016) that align poorly with real-world needs such as *the year in which Hook’s squad is sent to Belfast*, which is written directly in the short plot summary. In comparison, we aim for QAs closer to the conversations in movie clubs or what we talk about when we step out of the movie theatre i.e. the detailed yet crucial information to main event developments. This means we want question annotations that involve longer context of video clips and answer length, compared to prior work as reported in Table 2.

With this motivation in mind, we collect our dataset, *DeepMovieQA*, on 4 movies and 5 TV shows listed in Table 3, chosen based on genre diversity. We instruct annotators to come up with QA pairs that arise naturally as they watch the movie/TV shows, through two rounds. In the initial pass, we provide ten annotators movie/TV show assets split into 20 minute video clips for focused attention on detailed plot content. In the second pass, we ask the annotators to re-watch the entire movie/TV show asset, review QAs previously annotated, and come up with interesting QAs that involve piecing information together from several scenes or sources – such as discussion of themes, messages, empathetic reactions², and societal background. This two-pass QA labeling process allows annotators to become more familiar with the movie/TV show content and come up with challenging QAs that better align with genuine human audience interests. *DeepMovieQA* provides the first long-distance, multi-scene QA benchmark in the multimedia setting, with long-form answers labeled to promote elaborative movie QA model capabilities.

3 Methodology

3.1 Direct Retrieval from Source Data

Given a movie or TV show $V = \{V_f, V_{dial}, V_S\}$, consisting of the video frames, dialogues, and sum-

²Similar to <https://teachwithmovies.org/discussion-questions-for-use-with-any-film-that-is-a-work-of-fiction/>

Benchmark/ Dataset	Total Hrs	# Min. per Video Clip	# Tokens in Answer
MovieQA (Tapaswi et al., 2016)	280	-	5.6
TVQA (Lei et al., 2018)	461	1.3	5.1
TVQA+ (Lei et al., 2020a)	461	1.3	5.1
TVR (Lei et al., 2020b)	461	1.3	-
KnowIT VQA (Garcia et al., 2020)	69	0.3	4.5
DeepMovieQA (ours)	41	51.7	26.5

Table 2: Comparison of DeepMovieQA with respect to pre-existing datasets in the movie/TV domain. The # of mins and # of tokens listed are averaged numbers.

Video Asset	# Eps	# Min.
Manchester by the Sea (MBTS)	-	137
Without Remorse (WORE)	-	109
One Night in Miami (ONIM)	-	114
Les Miserables (LMIS)	-	105
Marvelous Mrs. Maisel (MRSM), S1	8	419
Jack Ryan (JKRY), S1	8	400
The Boys (TBYZ), S1	8	361
Transparent (TP), S1	10	287
The Family Man (FMAN), S1	10	448
Total	-	2480

Table 3: The movies/TV shows in DeepMovieQA.

mary, and a question, q , a natural first step is to select relevant context through semantic matching of local features. On the text side, we draw inspirations from previous research (Qu et al., 2020; Mossad et al., 2020) to encode the query, q , and candidate passages V_{dial_i} , truncated from every $n = 5$ utterance exchange (a tweakable hyperparameter), using *BERT* (Devlin et al., 2019), which are expressive bidirectional transformers for capturing latent language representations, followed by two separate linear layers, ϕ_q and ϕ_p . Then, we compute the retriever matching score through the cosine similarity of the encoded representations.

$$h_q = \phi_q(\text{ReLU}(\text{BERT}(q)))$$

$$h_p = \phi_p(\text{ReLU}(\text{BERT}(V_{dial_i})))$$

$$s_{q, V_{dial_i}} = \text{cos_sim}(h_q, h_p)$$

For visual retrieval, we utilize the powerful multimedia CLiP encoder release (Radford et al., 2021), which consists of a textual encoder component *CLiP_t* based on GPT (Radford et al., 2019) and a visual encoder component *CLiP_v* based on ViT

(Dosovitskiy et al., 2020), pretrained through contrastive learning on 400 million image caption pairs. We compare for semantic proximity between q , and a set of candidate visual frames, V_{f_i} through cosine similarity, similar to text retrieval. But note that as each candidate V_{f_i} is actually a set of visual frames corresponding to the time frame under V_{dial_i} , we take the mean of encoded image representations as the overall feature for V_{f_i} .

$$h_q = CLiP_t(q)$$

$$h_{V_{f_i}} = \frac{1}{|V_{f_i}|} \sum_{j \in 1..|V_{f_i}|} CLiP_v(V_{f_j})$$

$$s_{q, V_{f_i}} = \text{cos_sim}(h_q, h_{V_{f_i}})$$

As dialogue exchanges and visual scenes may contain cues that complement each other in signaling whether a section of the video is relevant to the query, we utilize a linear combination of the textual and visual retrieval scores for multimedia retrieval:

$$s_{q, (V_{f_i}, V_{dial_i})} = a * s_{q, V_{dial_i}} + b * s_{q, V_{f_i}}$$

We select the V_{dial_i} and/or V_{f_i} with top k semantic matching score for query q as the relevant context.

3.2 Retrieval from Structured Knowledge

Yet, structured knowledge provides benefits for long-distance and multi-hop information since events and interactions at separate time points (see Table 1 for examples) may then be directly connected through grounded nodes such as character entities.

Multimedia KG Construction To incorporate structured knowledge, we pre-construct *movieKG* using the open-source IE pipeline from (Wen et al., 2021) to extract events/entities/relations (Lin et al., 2020) from movie summary and dialogue, link entities to background knowledge base (Pan et al., 2015) where applicable, and perform event/entity coreference resolution (Lai et al., 2021). This leads to an initial sparse *KG*, in which the nodes consist of events (\mathcal{N}_v) and entities (\mathcal{N}_n) while the edges consist of argument roles and relations (E_r), following a pre-defined ontology which inadequately covers the open-domain in diverse film genres, so we augment ontological-guided IE with Abstract Meaning Representation (AMR) parsing (Fernandez Astudillo et al., 2020; Zhang and Ji, 2021) on the movie dialogue and summaries. For example, let’s consider a subgraph extracted from IE containing events such as “<Midge, arrested

(by), the police>”. IE has the advantage of performing entity/event linking, so the entity nodes have other direct connections in the constructed *MovieKG* as well, such as “<Midge, located in, Manhattan>” and “<police, (also) arrest, comedian Larry Bruce>”. However, other important events may not be captured by the IE ontology due to its more abstract or rarer occurrence in daily life and important news events, such as “<Midge, bailed out by, Gaslight manager Susie Meyerson>” and “<Gaslight manager Susie Meyerson, recognize, Midge’s talent>”, even though these information triplets can be directly extracted as the noun/verb-form concept nodes from AMR parsing on plot summary and dialogue transcripts. Hence, we add subgraphs from AMR parsing to the *movieKG* initially constructed from IE where there exists a coreferential event or entity node, such as “Midge”.

To further enrich this *KG* with visual information, we perform event extraction using grounded image situation recognition and localization (Pratt et al., 2020), which extracts the verb in action from the visual frame, as well as the semantic roles of objects detected (agent, item, destination, place, etc). We also perform character mapping for agents that play a role in visual events by finding the closest match from a bank of character profile images³, using visual features extracted by an iResNet model (Duta et al., 2021) following Meng et al. (2021). Finally, we merge objects and events that occur across textual and visual media based on embedding similarity in a multimedia common semantic space computed from CLiP, introduced in Sec 4.1.

Knowledge Subgraph Retrieval Now, we cannot simply select the neighborhood surrounding seed nodes that best match question q as the relevant subgraph because recurring entities contain many dense connections in the long story-telling domain. So instead, we select relevant subgraph based on context-aware saliency. Given a natural language question, we first represent it as a query graph that has undergone knowledge extraction for closer comparability with the *movieKG*. For instance, a question about “*the initial encounter between Midge and Larry Bruce*” becomes a graph with two connections: [Midge–the initial encounter, the initial encounter–Larry Bruce]. At the time of probing, we compute contextualized embeddings of node mentions, from BERT

³These can be manually identified from video frames or automatically collected from www.imdb.com.

encoders with awareness of the sentence that they occur in, as the local features for both the query graph and the movieKG, concatenate with wider-context knowledge embeddings from a two-hop neighborhood aggregation of maximally aligning neighbor nodes’ local embedding with respect to the query graph nodes. For query graph nodes, we concatenate the sentence-level BERT embedding as their wider context features. We score and rank $KG_{subgraph}$ selection based on the semantic (cosine) similarity of these knowledge embeddings between nodes in the movieKG and query graph.

3.3 Combining Context for Answer Fusion

In this work, we regard extractive QA as the task of selecting the relevant V_{dial_i} and V_{f_i} (as well as summary sentences V_{S_i} and $movieKG_{subgraph}$ though these lack labels for evaluation) from a movie/TV show, given question q . We formulate abstractive QA as the task of natural language answer generation, conditioned on these relevant V_{dial_i} , V_{f_i} , $movieKG_{subgraph}$, and V_{S_i} , providing complementary context to each other. Though it is tempting to learn an answer generation model that intakes context retrieved from different modalities through a straightforward common semantic space, we note the low-resource setting of our challenging task (in which annotations are expensive and time-consuming to obtain). Thus, we take advantage of pretrained conditional text generation transformers, in particular, BART, which has an encoder to extract context information and a generative decoder for sequential token generation, as a suitable backbone for answer generation (Lewis et al., 2020; Khashabi et al., 2020). Moreover, we aim to match the format of input that robust and high-performing conditional text generators have been pretrained on, which is natural language text. While q , V_{dial_i} , and V_{S_i} directly fit this desirable input format, retrieved V_{f_i} and $movieKG_{subgraph}$ should be projected into a textual semantic space for optimal alignment with our pretrained backbone answer generator. Hence, for the retrieved $movieKG_{subgraph}$, we take a stringified representation of its structured connections following (Ribeiro et al., 2020). For the retrieved video frames, we make sure that knowledge elements in these images are included in or concatenated to the stringified movieKG subgraph representation. Although our transformer-based answer fusion mechanism may be relatively simple and straightforward,

we observe it can handle QAs such as the ones in Table 6, detailed further in Sec 4. Figure 1 provides a walk-through illustration of our overall framework, which we refer to as the DEEPMAVEN. It is worthy to note that by nature of our information extraction-based QA reasoning, our textual answer generation is inherently supplemented with grounding to the visual context retrieved. This includes bounding box localization of the actions and participants (e.g., event – ‘drinking’, person – ‘Midge Maisel’, etc.) from the video frames.

4 Experiments

4.1 Benchmark and Dataset

DeepMovieQA Corpus This is our newly constructed dataset from 4 TV shows and 4 movies. The QAs may involve deep content beyond plot summary information, background knowledge, and higher-order reasoning across different scenes and events, as well as cross-media inferences. The answers are designed to be conversational friendly in nature and have an average token length of 19. A separate expert annotator manually checked 100 random QA pairs and judged all of them as accurate, informative, and comprehensive. We used a 8:1:1 train/val/test data split due to the corpus size.

Shallow MovieQA Corpus This is a multiple choice QA dataset (Tapaswi et al., 2016) annotated solely from movie plot summaries, with a pre-established 66-13-21% train/val/test data split.

4.2 Experimental Setting

For retrieval (extractive question answering), we include single modality approaches as the natural baselines to our proposed model, DEEPMAVEN.

- **Baseline 1: Dialogue Component Only**
Here, we simply retrieve the relevant dialogue sections from BERT embedding similarity with respect to the input question, similar to the top performing approach (Mossad et al., 2020) in MovieQA leaderboard.
- **Baseline 2: Visual Component Only**
Similarly, here, we use the features from CLiP for semantic matching between the question and visual frames.
- **Baseline 3: Textual & Visual Component w/o Structured Knowledge**
We further perform comparison with a simplified version of our DeepMaven framework that focuses on the cross-media dialogue and

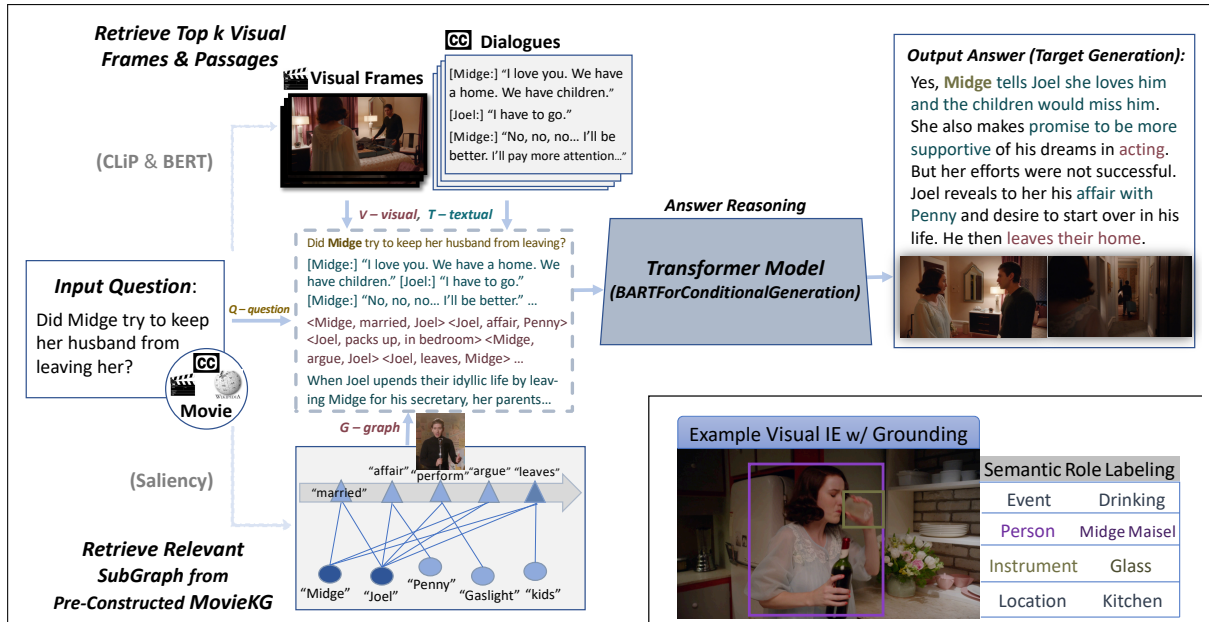


Figure 1: Our DEEPMAVEN model architecture takes as input the question, concatenated with relevant dialogues, visual frames, movieKG_{subgraph}, and summary sentences (illustrated by the dotted box), for answer stitching through a pretrained conditional text generator backbone. Note: due to low-resource data annotations, we convert the visual frame and KG_{subgraph} into string format to better align with the semantic space of the answer generator.

visual frames without structured knowledge from the movie KG_{subgraph}.

For abstractive question answer generation, we consider the base variant of DEEPMAVEN, which fine-tunes BART (Lewis et al., 2020), using the textual dialogue retrieved from movies/TV shows as input context only. We also include the baseline of UniVL (Luo et al., 2020), which has been pre-trained on video and language multimodal feature representation and text generation.

4.3 Evaluation Metrics

For long-distance retrieval, we report the hit@K metrics of whether the selected section of the dialogue or visual frames fall under the source time interval from the QA annotation. For natural language answer generation, we compute the ROUGE-L (Lin, 2004) F-scores⁴. However, there might be many variants of words and phrases with different level of granularities that can be used to describe the same answer in the same scenes. Take the question about “how Midge’s parent react to the news that Joel has left Midge” in “The Mavelous Mrs Maisel” for example. The ground truth answer annotation provides detailed descriptions on how “Midge’s dad starts to play the piano frantically and later blames Midge for marrying a weak man against his advice” while “Midge’s mother

starts crying and whining heavily”. We observed that our system generated answers are more general but still correct, outputting “the way the parents are reacting is very upset and blaming each other”. A strict measurement of n-gram overlaps from ROUGE would not credit such an answer generation sufficiently. Therefore, we also include a semantic-based BLEURT⁵ metric (Sellam et al., 2020). In addition, we conduct a human assessment on the extracted answers as well. For each pair of system extracted answer and ground-truth answer, we ask human assessors to judge whether the system answer is completely correct, partially correct or incorrect.

4.4 Quantitative Results and Analysis

Extractive Question Answer Retrieval As shown in Table 4, retrieval is a non-trivial task in the movie setting, which contains lengthy plot content, and a random approach has a very low hit@K=5 of 0.08. Single modality retrieval baselines perform similarly, with text retrieval using BERT and visual retrieval using CLiP having hit@K=5 of 0.32 and 0.28 respectively, while cross-media retrieval achieves a noticeable boost to hit@K=5 of 0.41, suggesting that textual and visual features offer similar levels of useful signals that complement each other for making sense of

⁴www.pyqi.org/project/pycocoevalcap/

⁵We use the BLEURT-20 scorer model from <https://github.com/google-research/bleurt>.

Approach	Hit@K=5
Random	0.08
Textual (BERT)*	0.32
Visual (CLiP)*	0.28
Textual+Visual Δ	0.41
Textual+Visual+KG priming	0.49

Table 4: These are the video frame retrieval results, with * as the baselines, Δ as the simplified DEEPMAVEN without structured knowledge, and **bold** as the full DEEPMAVEN retrieval approach.

question inputs from long movie/TV show content. Finally, movieKG-guided structured information on top of semantic matching from direct source data, using our DEEPMAVEN framework, unlocks the best retrieval hit rate (to near 50%).

Abstractive Question Answer Generation Natural language answer generation requires an additional decoder module for outputting tokens. Our DEEPMAVEN model achieves a Rouge-L of 61.6% in answer generation on the MovieQA benchmark. Model performance for the new, challenging Deep-MovieQA benchmark is reported in Table 5 below.

	Rouge-1	Rouge-L	BLEURT
UniVL	16.8	15.1	22.3
DeepMaven			
- dial. only	13.9	11.1	22.1
- all	21.6	17.4	31.7

Table 5: Answer generation (%) scores on our Deep-MovieQA dataset.

4.5 Qualitative Analysis

Table 6 shows answer generation results while Table 7 shows content selection examples, performed by DEEPMAVEN. We additionally conduct a human assessment on the DeepMovieQA answer generations from held-out data, rating answers on a likert scale of 0 – 5 (with 0 being nonsense and 5 being perfect). Thirteen percent of the generated answers scored ≥ 4 , with a Kappa coefficient of 0.42, reflecting moderate inter-annotator agreement.

4.6 Remaining Challenges

In general, we found that DEEPMAVEN retrieval performs worse on QAs that involve commonsense reasoning and reference to objects/entities that may not be easily picked up by our structure-guided cross-media semantic matching system. Some

	Questions	Answers
MovieQA	Who comes to the officers' rescue? – <i>World Trade Center</i>	Two United States marines, Dave Karnes and Jason Thomas. Two United States marines, Dave Karnes and Jason Thomas.
	Where are Stigman and Trench taken after being captured? – <i>2 Guns</i>	Mexico To Greco's farm in Mexico
DeepMovieQA	How does Joel leave the apartment after they fell asleep the night before? – <i>MRSM</i>	Joel gets out of bed and drives to work. Joel sneaks out Midge's bedroom window the next morning, just like he did when they first dated.
	What were in the boxes in the back of the truck Sajid was driving? – <i>FMAN</i>	They were filled with nerve gas. The boxes were filled with canisters of nerve gas that were going to be used to attack New Delhi.

Table 6: Example DEEPMAVEN abstractive QA results, with the **generated** and **grouth truth** answers.

deeper movie QAs may also be more suitable for the conversational question answering setting due to insufficiently detailed question wording and open-ended answer form. Finally, we observed that long-form answer generation with a limited-sized training data is more prone to hallucination. Given questions such as “What did Susie promise to one of the waiters at the Copacabana Club in return for smuggling her into the club for free?”, the language model generator backbone tends to fill in some plausible pre-trained knowledge such as “Susie promises to buy one of the men a drink after every act” that may not match the actual detailed answer of “Susie promised him a prime slot for his singing act at the Gaslight Café for 2 weeks. Unfortunately Susie thinks his act is awful but keeps her word even since they had to watch the show from the kitchen”. Minimizing hallucination in long-form low-resource question answer generation is an important issue that merits future investigation.

5 Related Work

Text QA: QA has been a popular task (Rajpurkar et al., 2016), with significant advances made recently by attention-guided transformers (Qu et al., 2019). Several corpora have been proposed for QA in story comprehension. FriendsQA (Yang and Choi, 2019) bases QA on short dialogue exchanges. MovieQA (Tapaswi et al., 2016) and NarrativeQA (Kocisky et al., 2018) are annotated from plot summaries. (Zhou et al., 2018; Moon et al., 2019; Zhou

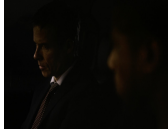


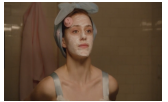


	Cue Category	Question	Retrieved Dialogue Passages and	Visual Frames
Impressive Cases	Dialogue Content	What did Kelly threaten Clay with after forcing him into his car? <i>-WORE</i>	"Why my family? My wife? My daughter? Why kill my team? Rykov was all in, so I'd like a little bit of what you sold him... Know where we driving? You got a farm out in West Virginia, right? Your daughter should be home from college. Wesleyan. That's a good school. Your wife. Your son too. So I suggest you start talking."	
	Visual Scenes	Who takes care of Lee after the fight? <i>-MBTS</i>	"Settle down, all right? Are we cool? Get off. You'll kill him... Should he go to the hospital? I don't think so. Nothing's broken."	
	Implicit Cross-Media Entity Grounding	What does Midge's mother want to talk about after Midge arrives at her parent's apartment? <i>-MRSM</i>	"Thanks for taking the kids last night. Were they okay? - We need to talk about the baby. Why? What's the matter with her? That forehead is not improving. - What? Are you sure?"	
	Implicit Multi-hop Event Reasoning	What does Midge do after Joel falls asleep and before he wakes up again? <i>-MRSM</i>	"Good night, Gracie. Hey. Good morning. - Did the alarm go off? - It sure did. Wow. I didn't hear it at all. You never do. - Good morning, Jerry. - Good morning, Mrs. Maisel."	
Erroneous Cases	Failure to Identify Certain Visual Objects	What does Midge suggest in the taxi to improve Joel's performance? <i>-MRSM</i>	"Maybe you should write a beginning, something that says who you are or something. What do you think? Good evening. What a nice... Good evening, ladies and gentlemen. Thank you for the nice... nice... 'Nice' is a bad, bad word. All that applause for me? What am I, putting out after?..."	
	Mismatch in Level of Detail Between Query and Context	What did Midge do while her friend Imogene is visiting to make sure she's still in shape? <i>-MRSM</i>	"She's going on and on about this miracle treatment she had done in Mexico. It involved goat's milk and avocados. Right ankle 8, left ankle 8. They smear it on your face, wrap a hot towel around your head..."	

Table 7: Impressive and Erroneous Examples of Retrieval Results.

et al., 2020) annotated open-ended conversations with groundings to either Wikipedia pages, or pre-existing structured knowledge graphs such as Freebase (Bollacker et al., 2008) and XLORE (Wang et al., 2013). These benchmarks overlook the full content from the original data, which we address. **Visual QA:** The visual question answering (VQA) task (Antol et al., 2015) aims to predict a natural language answer, given a natural language question and image(s) without other textual context. Various datasets have been constructed for this task, such as VQA 2.0 (Goyal et al., 2017), VCR (Zellers et al., 2019), and scientific PlotQA (Methani et al., 2020), in the setting of single images; as well as MSR-VTT-QA (Xu et al., 2016), MovieFIB (Maharaj et al., 2017), and VideoQA (Zhu et al., 2017), in the setting of visual frame sequences. The answers are typically multiple choices or from a predefined vocabulary. Simple baseline methods that only use question understanding (Kazemi and Elqursh, 2017) or sentiment analysis on answer options (Manjunatha et al., 2019) have proven surprisingly well on datasets such as VQA and VQA 2.0 but are unlikely to provide good answers for understanding complex events and person interactions. **Multimedia QA:** There are increasing interest

nowadays in using information from multiple modalities for answering questions, such as reasoning through text, images, and tables in Many-ModalQA (Hannan et al., 2020), MultiModalQA (Talmor et al., 2021), and MuMuQA (Reddy et al., 2022). The (movie/TV) video domain is a further step that involves dynamic events from dialogues and corresponding visual frames. TVQA (Lei et al., 2018) presents a multimedia QA dataset grounded on minute-long video clip snippets from popular TV show, with frame localization added in TVQA+ (Lei et al., 2020a). However, their QAs involve highly scene-specific discussion points, trivial to the deeper contents around the central plotline, e.g. "*Q: What is on the couch behind Joey when he is at the counter? A: A soccer ball.*". In contrast, our work highlights deeper knowledge and multi-frame information synthesis. **Other Story-based Video Understanding:** Benchmarks such as MovieNet (Huang et al., 2020) and LVU (Wu and Krahenbuhl, 2021) gauge a variety of content classification tasks related to the 'scene/place', 'cinematic style', 'genre', 'produced year', 'popularity', etc. Moreover, Lei et al. (2020b); Huang et al. (2020); Bain et al. (2020) explore retrieval based on sentence-length

scene description while Liu et al. (2020) study video-language inference, but these are all less challenging than the natural question answering setting our work focuses on.

6 Conclusions and Future Work

This work is the first to study long-distance movie question answering. Through guidance from multimedia source context and structured knowledge retrieval, our proposed model, DEEPMAVEN, is shown to perform well for extractive question answering, as well as abstractive answer generation for the questions in pre-existing MovieQA benchmarks. But improvements are needed for generating answers to the more challenging questions presented by our new DeepMovieQA benchmark, which serve as a better reflection of real-world application settings with more reasoning involved. We aim to kickstart an extractive movie question answering interface for human users, and through this, naturally acquire genuine movie questions for more effectively expanding DeepMovieQA annotation, and later extend this into interactive conversational AI settings.

7 Limitations

Our work focuses on video question answering that aims to be detailed and engaging. It is not interactive in nature like conversation exchanges. Finally, we bear in mind that automatically generated answers may contain the risk of insensitive phrasing, and mitigating language model bias deserves further exploration and efforts.

References

- Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.
- Max Bain, Arsha Nagrani, Andrew Brown, and Andrew Zisserman. 2020. Condensed movies: Story based retrieval with contextual embeddings. In *Proceedings of the Asian Conference on Computer Vision*.
- Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1247–1250.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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. Association for Computational Linguistics.
- Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.
- Ionut Cosmin Duta, Li Liu, Fan Zhu, and Ling Shao. 2021. Improved residual networks for image and video recognition. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 9415–9422. IEEE.
- Ramón Fernandez Astudillo, Miguel Ballesteros, Tahira Naseem, Austin Blodgett, and Radu Florian. 2020. Transition-based parsing with stack-transformers. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1001–1007, Online. Association for Computational Linguistics.
- Noa Garcia, Mayu Otani, Chenhui Chu, and Yuta Nakashima. 2020. Knowit vqa: Answering knowledge-based questions about videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10826–10834.
- Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6904–6913.
- Darryl Hannan, Akshay Jain, and Mohit Bansal. 2020. Mnymodalqa: Modality disambiguation and qa over diverse inputs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7879–7886.
- Sepp Hochreiter. 1998. The vanishing gradient problem during learning recurrent neural nets and problem solutions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 6(02):107–116.
- Qingqiu Huang, Yu Xiong, Anyi Rao, Jiaze Wang, and Dahua Lin. 2020. Movienet: A holistic dataset for movie understanding. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 709–727. Springer.
- Vahid Kazemi and Ali Elqursh. 2017. Show, ask, attend, and answer: A strong baseline for visual question answering. *arXiv preprint arXiv:1704.03162*.

- Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hananeh Hajishirzi. 2020. Unifiedqa: Crossing format boundaries with a single qa system. *arXiv preprint arXiv:2005.00700*.
- Tomas Kocisky, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gabor Melis, and Edward Grefenstette. 2018. The narrativeqa reading comprehension challenge. In *Transactions of the Association for Computational Linguistics*, volume 6, pages 317–328. MIT Press.
- Tuan Lai, Heng Ji, Trung Bui, Quan Hung Tran, Franck Dernoncourt, and Walter Chang. 2021. [A context-dependent gated module for incorporating symbolic semantics into event coreference resolution](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3491–3499, Online. Association for Computational Linguistics.
- Jie Lei, Licheng Yu, Mohit Bansal, and Tamara Berg. 2018. [TVQA: Localized, compositional video question answering](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1369–1379, Brussels, Belgium. Association for Computational Linguistics.
- Jie Lei, Licheng Yu, Tamara Berg, and Mohit Bansal. 2020a. [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. Association for Computational Linguistics.
- Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal. 2020b. [Tvr: A large-scale dataset for video-subtitle moment retrieval](#). In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16*, pages 447–463. Springer.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. [BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.
- Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.
- Ying Lin, Heng Ji, Fei Huang, and Lingfei Wu. 2020. A joint neural model for information extraction with global features. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7999–8009.
- Jingzhou Liu, Wenhui Chen, Yu Cheng, Zhe Gan, Licheng Yu, Yiming Yang, and Jingjing Liu. 2020. Violin: A large-scale dataset for video-and-language inference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10900–10910.
- Huaishao Luo, Lei Ji, Botian Shi, Haoyang Huang, Nan Duan, Tianrui Li, Jason Li, Taroon Bharti, and Ming Zhou. 2020. Univl: A unified video and language pre-training model for multimodal understanding and generation. *arXiv preprint arXiv:2002.06353*.
- Tegan Maharaj, Nicolas Ballas, Anna Rohrbach, Aaron Courville, and Christopher Pal. 2017. A dataset and exploration of models for understanding video data through fill-in-the-blank question-answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6884–6893.
- Varun Manjunatha, Nirat Saini, and Larry S Davis. 2019. Explicit bias discovery in visual question answering models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9562–9571.
- Qiang Meng, Shichao Zhao, Zhida Huang, and Feng Zhou. 2021. [Magface: A universal representation for face recognition and quality assessment](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14225–14234.
- Nitesh Methani, Pritha Ganguly, Mitesh M Khapra, and Pratyush Kumar. 2020. [Plotqa: Reasoning over scientific plots](#). In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1527–1536.
- Seungwhan Moon, Pararth Shah, Anuj Kumar, and Rajen Subba. 2019. [Opendialkg: Explainable conversational reasoning with attention-based walks over knowledge graphs](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 845–854.
- Omar Mossad, Amgad Ahmed, Anandharaju Raju, Hari Karthikeyan, and Zayed Ahmed. 2020. [Fat albert: Finding answers in large texts using semantic similarity attention layer based on bert](#). *arXiv preprint arXiv:2009.01004*.
- Xiaoman Pan, Taylor Cassidy, Ulf Hermjakob, Heng Ji, and Kevin Knight. 2015. Unsupervised entity linking with abstract meaning representation. In *Proceedings of the 2015 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pages 1130–1139.
- Sarah Pratt, Mark Yatskar, Luca Weihs, Ali Farhadi, and Aniruddha Kembhavi. 2020. Grounded situation recognition. In *European Conference on Computer Vision*, pages 314–332. Springer.
- Chen Qu, Liu Yang, Cen Chen, Minghui Qiu, W Bruce Croft, and Mohit Iyyer. 2020. Open-retrieval conversational question answering. In *Proceedings of*

- the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pages 539–548.
- Chen Qu, Liu Yang, Minghui Qiu, W Bruce Croft, Yongfeng Zhang, and Mohit Iyyer. 2019. Bert with history answer embedding for conversational question answering. In *Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval*, pages 1133–1136.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. [Learning transferable visual models from natural language supervision](#). In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.
- Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. [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. Association for Computational Linguistics.
- Revanth Gangi Reddy, Xilin Rui, Manling Li, Xudong Lin, Haoyang Wen, Jaemin Cho, Lifu Huang, Mohit Bansal, Avi Sil, Shih-Fu Chang, Alexander Schwing, and Heng Ji. 2022. Mumuqa: Multi-media multi-hop news question answering based on cross-media grounding. In *Proc. Thirty-Sixth AAAI Conference on Artificial Intelligence (AAAI2022)*.
- Leonardo FR Ribeiro, Martin Schmitt, Hinrich Schütze, and Iryna Gurevych. 2020. Investigating pretrained language models for graph-to-text generation. *arXiv preprint arXiv:2007.08426*.
- Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. Bleurt: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892.
- Alon Talmor, Ori Yoran, Amnon Catav, Dan Lahav, Yizhong Wang, Akari Asai, Gabriel Ilharco, Hananeh Hajishirzi, and Jonathan Berant. 2021. Multimodalqa: Complex question answering over text, tables and images. *arXiv preprint arXiv:2104.06039*.
- Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. 2016. Movieqa: Understanding stories in movies through question-answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4631–4640.
- Zhigang Wang, Juanzi Li, Zhichun Wang, Shuangjie Li, Mingyang Li, Dongsheng Zhang, Yao Shi, Yongbin Liu, Peng Zhang, and Jie Tang. 2013. Xlore: A large-scale english-chinese bilingual knowledge graph. In *International semantic web conference (Posters & Demos)*, volume 1035, pages 121–124.
- Haoyang Wen, Ying Lin, Tuan Lai, Xiaoman Pan, Sha Li, Xudong Lin, Ben Zhou, Manling Li, Haoyu Wang, Hongming Zhang, et al. 2021. Resin: A dockerized schema-guided cross-document cross-lingual cross-media information extraction and event tracking system. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations*, pages 133–143.
- Chao-Yuan Wu and Philipp Krahenbuhl. 2021. Towards long-form video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1884–1894.
- Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296.
- Zhengzhe Yang and Jinho D Choi. 2019. Friendsqa: Open-domain question answering on tv show transcripts. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 188–197.
- Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6720–6731.
- Zixuan Zhang and Heng Ji. 2021. Abstract meaning representation guided graph encoding and decoding for joint information extraction. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 39–49.
- Hao Zhou, Chujie Zheng, Kaili Huang, Minlie Huang, and Xiaoyan Zhu. 2020. Kdconv: A chinese multi-domain dialogue dataset towards multi-turn knowledge-driven conversation. *arXiv preprint arXiv:2004.04100*.
- Kangyan Zhou, Shrimai Prabhunoye, and Alan W Black. 2018. A dataset for document grounded conversations. *arXiv preprint arXiv:1809.07358*.
- Linchao Zhu, Zhongwen Xu, Yi Yang, and Alexander G Hauptmann. 2017. Uncovering the temporal context for video question answering. *International Journal of Computer Vision*, 124(3):409–421.