VideoChain: A Transformer-Based Framework for Multi-hop Video Question Generation

Arpan Phukan, Anupam Pandey, Deepjyoti Bodo and Asif Ekbal
Department of Computer Science and Engineering, IIT Patna
arpan_2121cs33@iitp.ac.in/iitpainlpmlresourcerequest@gmail.com,
anupam_2311cs31@iitp.ac.in, deepjyoti_2311ai65@iitp.ac.in, asif@iitp.ac.in

Abstract

Multi-hop Question Generation (QG) effectively evaluates reasoning but remains confined to text; Video Question Generation (VideoQG) is limited to zero-hop questions over single segments. To address this, we introduce VideoChain, a novel Multi-hop Video Question Generation (MVQG) framework designed to generate questions that require reasoning across multiple, temporally separated video segments. VideoChain features a modular architecture built on a modified BART backbone enhanced with video embeddings, capturing textual and visual dependencies. Using the TVQA+ dataset, we automatically construct the large-scale MVQ-60 dataset by merging zero-hop QA pairs, ensuring scalability and diversity. Evaluations show VideoChain strong performance across standard generation metrics: ROUGE-L (0.6454), ROUGE-1 (0.6854), BLEU-1 (0.6711), BERTScore-F1 (0.7967), and semantic similarity (0.8110). These results highlight the model's ability to generate coherent, contextually grounded, and reasoning-intensive questions. To facilitate future research, we publicly release our code and dataset[1].

1 Introduction

The field of Question Generation (QG) has garnered substantial attention for its potential to create interactive and informative learning environments, educational tools, and intelligent systems. Traditionally, QG systems focused on generating questions based on textual passages, with applications spanning from educational quizzes (Krishna et al., 2015) to interview questions and clarification prompts (Kumar and Black, 2020). However, the exploration of QG in video-based content remains significantly underdeveloped compared to its text-based counterpart.

Video Question Generation (VideoQG) involves generating questions based on the visual and textual information available in video content. This task is particularly important for assessing a model's ability to understand and reason over dynamic and temporal data, as videos often present information across multiple frames, segments, and visual cues. Despite its potential, existing VideoQG research has mostly focused on generating zero-hop questions from transcripts or basic visual elements in videos, limiting the scope of reasoning required from the model. In contrast to zero-hop questions, multi-hop question generation requires reasoning across multiple, often non-contiguous, segments of data. Although this has been extensively explored in the text domain through datasets like HotpotQA (Yang et al., 2018b), it remains relatively underexplored in video-based tasks. Multi-hop questions challenge the model to synthesize information across several frames or segments in a video, thus requiring a more in-depth understanding of both visual and textual modalities.

To address this gap in the literature, we introduce the task of Multi-hop Video Question Generation (MVQG), where the goal is to generate questions that require reasoning over multiple, temporally-separated segments of a video. To facilitate this task, we construct a new MVQG dataset (MVQ-60) by merging zero-hop questions from the TVQA+ dataset (Lei et al., 2020a). Inspired by the MusiQue paper (Trivedi et al., 2022), which demonstrated the effectiveness of merging simple questions to create multi-hop questions in the text domain, we extend this concept to videos. This dataset is designed to challenge models to process and integrate information from different parts of a video, combining both visual frames and textual transcripts. Additionally, we fine-tune the BART (Lewis et al., 2020), enhanced with video embeddings, to generate coherent, contextually rich multi-hop questions that span multiple video segments. The fine-

[1] https://github.com/AnupamPandey199949/VideoChain

Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 2246–2266

tuned models developed in this work demonstrate a significant improvement in generating complex, multi-hop video-based questions both in terms of automatic evaluation metrics and human judgment.

We summarize the contributions of our work as: **(1)** We contribute the first MVQG system by fine-tuning a customized BART architecture, incorporating text and video embeddings, to generate multi-hop questions requiring reasoning across video segments. **(2)** We introduce MVQ-60, the first dataset, specifically designed for MVQG, created by merging zero-hop questions from the TVQA+ dataset, enabling complex reasoning over multiple video segments.

2 Related Work

2.1 Text-based Question Generation

Question Generation (QG) in text has been extensively studied, with works focusing on various levels of textual granularity. These include document-level QG (Pan et al., 2020a), paragraph-level QG (Zhang et al., 2017), sentence-level QG (Ali et al., 2010), and keyword-based QG (Pan et al., 2020b). Early works employed rule-based approaches, but recent advancements leverage deep learning models, particularly sequence-to-sequence architectures and pre-trained transformers (Pan et al., 2019). Techniques, such as semantic parsing and reinforcement learning have further enhanced the quality of generated questions (Chatterjee et al., 2020). While text-based QG has matured significantly, the shift towards multimodal domains presents new challenges, particularly in reasoning over both visual and temporal modalities.

2.2 Visual Question Generation

Visual Question Generation (VQG), introduced by (Mostafazadeh et al., 2016), generates questions from images and encompasses three types: Visually Grounded (answerable from the image (Antol et al., 2015)), Commonsense-Based (requiring external knowledge (Wang et al., 2017)), and World Knowledge-Based (integrating factual knowledge bases (Shah et al., 2019)). Proposed methods include encoder-decoder (Mostafazadeh et al., 2016), compositional (Liu et al., 2018), and generative models (Jain et al., 2017), enhanced by reinforcement learning (Yang et al., 2018a) and bilinear pooling (Fukui et al., 2016). Domain-specific applications (e.g., medical imaging, education (Mehta et al., 2024)) exist, but challenges persist in visual grounding, multi-object reasoning, and extending these challenges to video.

2.3 Video Question Generation

VideoQG is inherently more challenging than text or visual QG due to the temporal structure and multimodal nature of videos. Early works (Yang et al., 2021) primarily focused on generating questions based on video transcripts or static object and attribute descriptions (Gupta and Gupta, 2022), but fell short of addressing more complex reasoning requirements. Multi-hop reasoning in video QG presents unique challenges: Contextual Integration: Generating self-contained questions that require linking temporally distant events in a video. Entity-Action Mapping: Associating visual entities with their respective actions or interactions in a coherent manner. Multimodal Fusion: Effectively leveraging signals from various modalities (e.g., video frames, audio, and textual subtitles) to generate questions that reflect comprehensive reasoning. Existing VideoQG datasets (Gupta and Gupta, 2022; Acharya et al., 2019) target zero-hop question generation, lacking support for reasoning across multiple video segments. While recent video-language models like Flamingo (Alayrac et al., 2022) and Vid2Seq (Yang et al., 2023) advance video understanding, they remain limited for multi-hop question generation evaluation.

Our work addresses these gaps by introducing a novel dataset MVQ-60 and developing method specifically designed for multi-hop reasoning over videos (see table 3).

3 Datasets

VideoQA progress stems from datasets with distinct challenges. Existing datasets, such as MSR-VTT (Xu et al., 2016b) (open-domain video captioning), HowTo100M (Miech et al., 2019b) (instructional videos), TVQA+ (Lei et al., 2020a) (narrative comprehension), ActivityNet-QA (Yu et al., 2019b)(grasping complex videos), and others have been useful in evaluating the performance of VideoQA models. While existing datasets focus on zero-hop questions (answerable from single events), multi-hop reasoning across video segments remains underexplored.

3.1 Dataset Creation

Recognizing the lack of multi-hop VideoQA datasets, we opted to create a new dataset. Given

the scalability and reproducibility challenges of manual annotation, we pursued the development of an automated process to generate multi-hop questions by using existing datasets. This decision was inspired by the MUSIQUE dataset (Trivedi et al., 2022), which successfully generated textual multi-hop questions through automated merging techniques. Our methodology involved merging two zero-hop video questions to form a multi-hop video question. Among the 40 datasets reviewed, TVQA+ (Lei et al., 2020a) consisting of 152,545 QA pairs from 21,793 clips, spanning over 460 hours of videos based on 6 popular TV shows(The Big Bang Theory, How I Met Your Mother, Friends, House M.D., Grey's Anatomy, and Castle), emerged as the optimal base for dataset construction. Its advantages include:

High Annotation Quality: Questions and answers are manually crafted, ensuring accuracy and relevance. **Contextual Richness:** Includes subtitles, video frames, and metadata (e.g., episodes, seasons). **Widespread Use:** Recognized as a benchmark in VideoQA research, ensuring broad compatibility with existing methods. the TVQA+'s extensive coverage of temporally rich, real-world scenarios provided an ideal foundation for our multi-hop dataset.

3.2 Automated multi-hop Question Generation

To generate multi-hop questions, we developed a rule-based merging algorithm inspired by the MUSIQUE dataset's textual question generation strategy (Trivedi et al., 2022), that combines pairs of zero-hop questions into coherent multi-hop questions. The key steps in the algorithm are:

Question Filtering: Short questions with concise answers were prioritized to maintain readability and prevent excessive length in merged multi-hop questions. Based on the empirical distribution of the TVQA+ dataset and experimental validation, we set the length thresholds to 15 words for questions and 3 words for answers. These thresholds ensured broad coverage of the dataset while avoiding verbosity, and led to the most effective generation of coherent multi-hop questions.

Temporal and Contextual Matching: The questions were grouped based on the shared metadata, specifically the episode. Let,$\mathcal{M} = \{m_1, m_2, \ldots, m_k\}$ represent the metadata attributes, where m includes the episode (e) and segment (s). Two questions q_i and q_j are considered

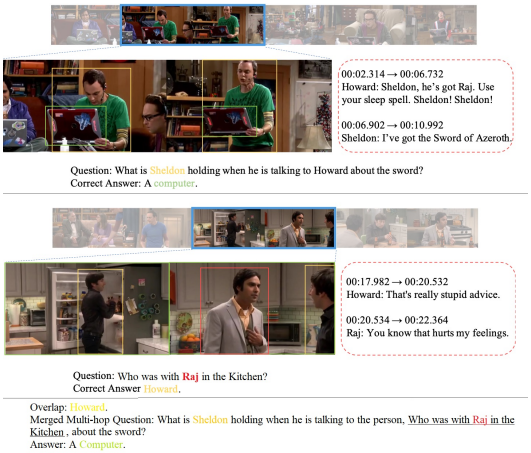

Figure 1: Example: Merged Multi-hop Question

Temporally and contextually aligned if they share the same episode but have different segments:

$$\text{Match}(q_i, q_j) = \begin{cases} 1, & \text{if } e_i = e_j \text{ and } s_i \neq s_j \\ 0, & \text{otherwise} \end{cases}$$

This ensures that only questions referring to the same episode but different segments are considered for merging.

Overlap Detection: Overlap between two questions is defined as instances where the answer to one question forms a semantic part of another question. Let q_1 and q_2 be two questions with answers a_1 and a_2, respectively. Overlap is defined as:

$$\text{Overlap}(q_1, a_2) = \begin{cases} 1, & \text{if } a_2 \in q_1 \\ 0, & \text{otherwise} \end{cases}$$

Question Pairing and Merging: Pairs of overlapping questions were merged to create multi-hop questions by replacing the overlap in first question q_1 and second answer a_2 with the second question q_2. The merged question q_{merged} is defined as:

$$q_{\text{merged}} = q_1 \setminus a_2 + q_2$$

where $q_1 \setminus a_2$ denotes the replacement of a_2 in q_1 with q_2's context. An example of this process is shown in Figure 1.

3.3 Quality Evaluation Metrics

To assess the quality of the generated questions, we evaluate them using a range of metrics: **Fluency:** Grammatical correctness and natural language quality. **Multi-Hop Reasoning:** Complexity of reasoning required to answer the question. **Video Relevance:** Degree of relevance to one or both videos.

Engagingness: How captivating and interesting the question is. **Factual Correctness:** Logical and factual accuracy of the question-answer pairs. **Inclusiveness:** How well the questions covered diverse aspects of the video content. Each metric was scored on a scale of 0 to 3, where 3 indicated the highest quality. For example:

Fluency: "What is Chandler's wife cooking?" → Score: 3 (Excellent) Relevance: "Compare the styles of dance by people in both videos." → Score: 3 (Highly Relevant)

For scalability, we conducted evaluation on a random sample of 200 questions. According to statistical sampling theory, such a subset can provide reliable estimates of overall dataset characteristics. To further validate the dataset's quality and ensure its appropriateness for multi-hop video QA tasks, we performed manual human evaluation on the sampled set. Annotators scored each question across the above six criteria using predefined rubrics. The average scores were: Fluency(2.92), Multi-hop Reasoning(3.00), Engagingness(2.80), Factual Correctness(3.00). The high scores, particularly in reasoning and factual correctness, are attributed in part to the use of high-quality human annotated TVQA base questions, which were merged to construct MVQ-60. All annotations were done by three trained annotators following strict rubrics which ensured high objectivity and achieved an inter-rater agreement (Cohen's κ) of **0.72**. These results reinforce the validity and challenge-level of MVQ-60 for future multi-hop video reasoning research. Finally, we introduce MVQ60, the first large-scale dataset of multi-hop (two-hop) video questions, consisting of over 60,000 questions based on six popular TV shows (Friends, The Big Bang Theory, How I Met Your Mother, House M.D., Grey's Anatomy, Castle), with an average question length of 27 words.

4 Methodology

Video Embedding Generation: A key aspect of our methodology was the generation of expressive video embeddings to encode spatio-temporal dynamics. We used VideoMAE, a masked autoencoder for self-supervised video representation learning (Wang et al., 2023). VideoMAE embeddings effectively capture both motion and appearance features while maintaining computational efficiency. These embeddings served as the foundational representation for all the subsequent experiments and models.

Initial Exploration with Video-Based Models: We began by finetuning state-of-the-art video-based models, pretrained on tasks, such as video captioning and VideoQA, to adapt them for MVQG. These models were trained using VideoMAE embeddings and textual inputs (e.g., video transcripts). This approach was inspired by works like (Lei et al., 2020a), which emphasized the use of spatio-temporal features for question answering, and (Yu et al., 2019b), which demonstrated the benefits of video pretraining for understanding complex narratives. While these models showed promise in zero-hop reasoning tasks, they struggled with multi-hop questions. Their large sizes and monolithic architectures resulted in slow processing times and difficulty scaling to higher-hop reasoning. These limitations prompted us to explore lightweight, text-based alternatives.

Transition to Text-Based Models: Inspired by (Phukan et al., 2024) that text-based models can generate high-quality video questions with video embeddings, we explored using T5 (Raffel et al., 2020), BART (Lewis et al., 2020), and similar models. We modified these text-based models to accept VideoMAE embeddings as additional input alongside textual data. During finetuning, the models were provided with both embeddings and transcripts. While this setup improved efficiency, the generated questions occasionally failed to integrate information across video segments coherently. These challenges, highlighted the need for explicit architectural modifications to handle multi-hop reasoning. **Modular Two-Component Architecture:** To address the observed limitations, we develop a model with two-component architecture tailored for multi-hop question generation. This modular design was inspired by the principles of (Andreas et al., 2016), which demonstrated the benefits of task decomposition for reasoning tasks, and (Trivedi et al., 2022), which emphasized modular frameworks for multistep reasoning. The first component generates zero-hop questions from individual video segments. It accepts as input the VideoMAE embedding of a video clip, its corresponding transcript, and a prompt that guides the question generation process. The model outputs a concise, contextually grounded question. For example, given a clip where Monica is cooking, this component generates the question "What is Monica cooking?". This step isolates relevant information from each video segment, forming a foundation for

multi-hop reasoning. The second component integrates information from multiple video segments to generate multi-hop questions. It accepts as input the zero-hop question produced by the first component, along with the VideoMAE embedding and transcript of a second video segment. The model refines and expands the question, producing a multi-hop question, such as "What is Chandler's wife cooking?". This process can be repeated iteratively, allowing the model to generate higher-hop questions (e.g., three-hop or four-hop questions) by forwarding the output question to subsequent iterations with new inputs.

Two-Component Approach: The modular nature of the two-component architecture offers several key benefits. By decoupling zero-hop and multi-hop reasoning, the system scales efficiently to higher-hop questions without overwhelming model capacity. Each component specializes in a specific task, improving both precision and contextual grounding.

5 Model Architecture

The proposed architecture for MVQG (figure 2) centers around VideoChain, a version of the BART-large CNN model (Lewis et al., 2020) we modified to process both video and textual inputs as distinct modalities. It integrates the spatio-temporal dynamics of video content through VideoMAE embeddings (Wang et al., 2023) while preserving BART's generative capabilities for text-based reasoning. VideoChain processes two primary input types: video embeddings and text embeddings. Video embeddings are generated using VideoMAE, a self-supervised video representation learning model that encodes video segments into $\mathbb{R}^{1568\times1024}$ dimensional feature vectors. These embeddings capture both motion and appearance features, providing a compact representation of the video's spatio-temporal content. Text embeddings are derived from video transcripts and prompts. Notably, the prompts differ between components; the first component uses prompts to generate zero-hop questions (e.g., "Generate a question about this clip"), while the second component uses prompts to guide the generation of multi-hop questions (e.g., "Generate a multi-hop question based on the previous question and this clip"). The architecture builds upon BART-large CNN by introducing modifications that enable it to handle multimodal inputs effectively. First, the encoder is extended to include dual input streams for video and text embeddings. The video embeddings are processed through dedicated multi-head attention and feedforward layers designed for spatio-temporal data, while the text embeddings are processed through the standard BART encoder layers. A cross-modal attention mechanism is introduced to fuse the outputs of the video and text streams, enabling VideoChain to reason jointly over both modalities. The decoder attends to the fused multimodal representation, generating the output question token by token. We use this VideoChain in both modules of our architecture.

Module 1:Zero-hop Question Generation: The zero-hop question generation component constitutes the first stage of the architecture. This component generates a concise question based on the content of a single video segment. The video embeddings, transcripts, and prompts are processed independently through their respective streams in VideoChain's encoder. The cross-modal attention mechanism aligns the visual and textual representations, producing a unified multimodal encoding that informs the decoder's generation process. For example, given a video clip where Monica is cooking, the model generates the question "What is Monica cooking?". The training process for this component employs cross-entropy loss. By focusing on zero-hop reasoning, this component ensures that VideoChain can effectively extract and represent information from individual video segments.

Module 2:Multi-hop Question Composition: The second stage of the architecture, the composition of multi-hop questions, extends the zero-hop question by incorporating additional information from subsequent video segments. This component takes as input the zero-hop question generated by the first component, along with the video embedding, transcript, and a multi-hop-specific prompt corresponding to the second video segment. The encoder processes these inputs in their respective streams, and the cross-modal attention mechanism aligns the zero-hop question with the visual and textual context of the second segment. The decoder refines and expands the zero-hop question into a multi-hop question. For instance, if the zero-hop question is "What is Monica cooking?" and the second segment provides context about Monica's relationship with Chandler, Module-2 generates the multi-hop question "What is Chandler's wife cooking?". This iterative design enables the architecture to handle reasoning tasks of arbitrary complexity

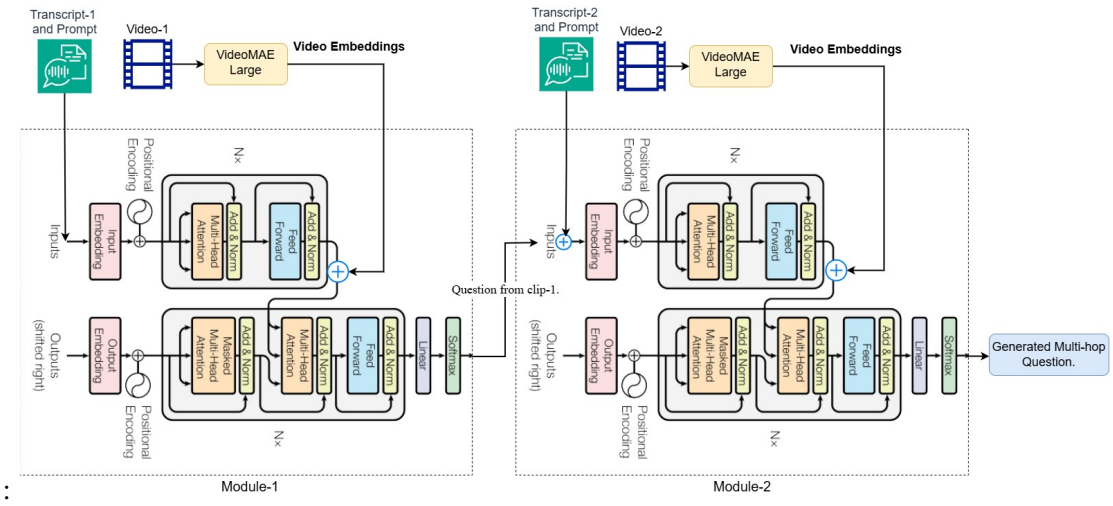

Figure 2: Proposed Model Architecture

by recursively invoking the second component with new inputs.

Training Strategy: Our model is trained in two stages to optimize its performance for zero-hop and multi-hop tasks. The zero-hop Question Generation component is trained on zero-hop question-answer pairs using cross-entropy loss, while the multi-hop Question Composition component is trained on multi-hop question-answer pairs, with ground-truth intermediate questions provided during training. A composite loss function is used for the second component, combining cross-entropy loss with alignment loss to ensure effective multimodal fusion. During inference, our model uses beam search to enhance the fluency and coherence of the generated questions, particularly for higher-hop reasoning tasks where maintaining logical consistency is critical. **Scalability and Adaptability:** Our model's modularity ensures scalability by isolating reasoning subtasks into distinct components. The recursive nature of the multi-hop component allows the system to handle increasingly complex tasks without requiring additional architectural changes. VideoChain's flexibility allows it to integrate with other pre-trained generative models. While BART-large CNN serves as the base in this implementation, the modifications applied to handle video and text inputs can be extended to models, such as T5 (Raffel et al., 2020) or mBART (Liu et al., 2020), enabling the framework to adapt to a wide range of datasets and applications. This adaptability is crucial for advancing multimodal reasoning in VideoQA (Zellers et al., 2019).

6 Experiments, Results and Analysis

6.1 Experiment Setup

For our experiments, we used the MVQ-60 dataset, which comprises 60,000 multi-hop questions paired with video segments and corresponding transcripts. The dataset was split into 80% training, 10% validation, and 10% test, ensuring no episode level overlap between these splits to prevent data leakage and overfitting. During the fine-tuning step, we used the input IDs and attention masks of the inputs, which consisted of concatenated prompts and transcripts representing the questions, and passed the video embeddings as a separate entity into our model. The model was trained on 2 Tesla T4 GPUs on kaggle for a total of 8 hours. *Hyperparameters:* a learning rate of 3e-5, a batch size (4)suitable for the available hardware, and gradient accumulation steps set to 4. Training was conducted for 50 epochs. We used a mixed precision training approach with FP16 enabled for compatibility with CUDA, and we monitored performance using the evaluation strategy set to run every 100 steps. The maximum gradient norm was clipped to 1.0 to ensure stable training, while unnecessary columns were removed to optimize memory usage. Despite being a compact architecture (406M parameters, using BART-large), our model delivers strong multi-hop reasoning performance with significantly lower training time and resource requirements compared to large-scale vision-language models. For example, Qwen2-VL-2B requires approximately 28 hours of training on similar data, whereas our VideoChain-based model achieves competitive results with just 8 hours of training, highlighting the

efficiency and scalability of our approach.

6.2 Experiments

As VideoChain represents the first dedicated architecture for MVQG, there are no prior models explicitly trained for this task. To establish meaningful baselines, we conducted zero-shot evaluations using recent general-purpose multimodal models: Qwen2-VL-2B-Instruct, SmolVLM, MGM-2B, and PaliGemma. These models were selected for their ability to process video and language inputs without task-specific fine-tuning. In addition, to assess the adaptability of existing VideoQA models, we finetuned the ECIS model (Phukan et al., 2024) on our MVQ-60 dataset. This provides a strong baseline from the VideoQA domain. This evaluation highlights the limitations of generic vision-language models and repurposed QA models in generating coherent, compositional video-grounded questions.

6.3 Evaluation Setup

For Human Evaluation: To ensure a robust and diverse evaluation, we recruited human annotators from various demographic backgrounds. Our annotators consisted of individuals proficient in English, with a mix of undergraduate and graduate students. Each evaluator was tasked with assessing a randomly sampled subset of multi-hop questions based on predefined quality metrics. To maintain fairness and ethical considerations, all the annotators were compensated at a competitive rate in accordance with the standard research compensation guidelines.

For zero-shot evaluation, we provided each model with pairs of video segments, corresponding transcripts, and prompts designed to elicit multi-hop reasoning. The prompts were standardized across models to ensure fair comparison. For example: *"Based on the two video segments and their transcripts, generate a question that requires integrating information from both videos."* The models' outputs were evaluated on both automated and human evaluation metrics, including *fluency*, *relevance*, *multi-hop reasoning*, *factual correctness*, *engagingness* and *inclusiveness*.

6.4 Results Analysis

As summarized in Tables 1 and 2, our proposed **VideoChain** model consistently outperforms recent multimodal baselines across both human and automatic evaluation metrics. In **human evaluation**, while all models demonstrate strong **fluency** and **engagingness** due to their pretrained language modeling capabilities,**VideoChain** achieves the highest scores in **relevance (2.91)**, **multi-hop reasoning (2.81)**, and **factual correctness (2.92)**. The fine-tuned ECIS model also shows strong performance, particularly in fluency (2.90) and correctness (2.84), but falls slightly short in relevance and multi-hop depth. The relatively low factual correctness scores of models, such as Qwen2-VL (2.56), PaliGemma (2.20), and SmolVLM (1.98) suggest hallucination and external knowledge leakage, likely caused by exposure to the same TV shows during pretraining. In contrast, VideoChain is explicitly trained on grounded supervision, encouraging content-aligned generation. The **inclusiveness** metric further highlights the strength of our architecture. As most baselines treat both video segments jointly, they often produce questions grounded in only one clip. VideoChain's modular dual-stage design processes each clip in separate stages, enabling more inclusive question generation (**2.78** compared to 2.05 for Qwen2-VL and 2.14 for PaliGemma). For **multi-hop reasoning**, baseline models frequently concatenate multiple independent zero-hop questions rather than forming a coherent multi-hop question. Our model, trained with explicit multi-hop supervision, demonstrates better temporal and semantic integration across clips. In **automatic evaluation**, VideoChain leads across all metrics, BERTScore F1 (0.7967), semantic similarity (0.8110), ROUGE-1 (0.6854), ROUGE-L (0.6454), BLEU-1 (0.6711), Distinct-1 (0.7911), and Distinct-2 (0.9850). These gains show that VideoChain generates fluent, diverse, and semantically aligned multi-hop questions. ECIS also shows competitive performance, especially in fluency-aligned metrics, confirming the benefits of fine-tuning on MVQ-60. while the other Baseline models often generate overly long questions (e.g., PaliGemma: 82.6 tokens, SmolVLM: 78.1), which negatively impacts lexical and semantic alignment, the stricter length-controlled finetuning of VideoChain and ECIS, helped them produce more concise and accurate questions.

6.5 Discussion

The zero-shot evaluation highlights the limitations of current pre-trained models in addressing multi-hop VideoQA tasks without specialized training. While models like Qwen2-VL-2B-Instruct

Table 1: Results on Human Evaluation

Model	Fluency	Relevance	Multi-Hop Reasoning	Engagingness	Factual Correctness	Inclusiveness
VideoChain (Ours)	2.89	**2.91**	**2.81**	**2.75**	**2.92**	**2.78**
ECIS (Finetuned)	2.90	2.85	2.72	2.63	2.84	2.65
Qwen2-VL-2B-Instruct	2.71	2.32	1.54	2.25	2.56	2.05
SmolVLM	2.51	2.09	1.24	1.82	1.98	2.24
MGM-2B	2.80	2.54	1.64	2.21	2.24	1.97
PaliGemma	**2.95**	2.60	1.77	2.41	2.18	2.14

Table 2: Results on Automatic Evaluation

Model	Bert score F1	Generation length	Semantic similarity	Rouge-1	Rouge-L	Bleu-1	Distinct-1	Distinct-2
VideoChain (Ours)	**0.7967**	**53.2**	**0.8110**	**0.6854**	**0.6454**	**0.6711**	**0.7911**	**0.9850**
ECIS (Finetuned)	0.6253	47.5	0.5291	0.4006	0.3174	0.4203	0.7430	0.9553
Qwen2-VL-2B-Instruct	0.5120	75.3	0.4547	0.2746	0.2451	0.3712	0.7230	0.9370
SmolVLM	0.4881	78.1	0.5154	0.2819	0.2482	0.3574	0.7115	0.9260
MGM-2B	0.5046	70.7	0.5627	0.3004	0.2627	0.3861	0.7250	0.9410
PaliGemma	0.5287	82.6	0.4987	0.2912	0.2519	0.4021	0.7340	0.9505

and PaliGemma show promise in handling general vision-language tasks, their performance on multi-hop reasoning remains suboptimal compared to our fine-tuned model. This suggests that while general-purpose multimodal models offer flexibility, task-specific architectures like VideoChain are essential to achieve state-of-the-art performance in complex reasoning tasks, such as MVQG.

6.6 Ablation Study

We assess the contribution of VideoChain's core components through two ablations: (1) Text-only variant: Removing video embeddings to isolate text reliance caused significant performance drops, particularly in video relevance (2.91 to 2.09) and multi-hop reasoning (2.85 to 1.54). This confirms visual grounding is essential for contextually rich question generation. (2) Single-component variant: Replacing the modular pipeline with direct multi-hop generation severely degraded reasoning capability (multi-hop: 1.24 vs. 2.85) and factual correctness (1.98 vs. 2.97), often yielding shallow or concatenated questions. Both variants showed substantial overall performance degradation (0.69 text-only; 0.76 single-component) versus the full model. These results validate the necessity of multimodal inputs for visual grounding and modular decomposition for complex reasoning. Detailed metrics and analysis are provided in the appendix section D.1.

6.7 Error Analysis

To better illustrate the strengths and limitations of our MVQG model, we present qualitative examples across the six TVQA+ shows in Table 6. Each row provides one example from a distinct TV show, categorized into four groups: (1) **Correct Generations**, (2) **Multi-Hop Reasoning Failures**, (3) **External Knowledge Leakage**, and (4) **Hallucination**. Detailed Error Analysis is discussed in the Appendix section E

7 Conclusion and Future Work

We introduced VideoChain, the first modular architecture for MVQG. We created MVQ-60, a large-scale multihop video question dataset spanning six TV shows. VideoChain's modular design ensures scalability for complex reasoning. Evaluations demonstrated its strong performance in generating fluent, relevant, and coherent video-grounded multihop questions, validating our approach. We evaluated the intrinsic quality of the generated questions (using automatic metrics and human judgements), but we did not perform an extrinsic evaluation, e.g., using the questions for a downstream task like training a VideoQA model or testing a model's reasoning capabilities. Future work includes expanding to diverse domains (e.g., education, surveillance), developing domain-specific MVQG systems, enabling multilingual generation, and integrating emerging vision-language models for enhanced reasoning and nuance.

8 Limitations

Despite the advancements demonstrated in this work, several limitations warrant further investigation. Expanding to diverse video domains (e.g., educational, surveillance) requires tackling distinct features and QA demands, likely needing domain-specific adaptations. Second, our system generates questions only in English. Enabling multilingual question generation is crucial for broader accessibility but involves complex cross-lingual understanding and generation. Third, the modular pipeline risks error propagation; factual errors

from Module-1 often persist despite Module-2's mitigation of grammatical/semantic issues mitigation. While not significantly impacting overall quality, explicit error correction or joint optimization could help. Finally, rapidly evolving vision-language models offer potential for more powerful representations, reasoning, and nuanced question generation. Future work should integrate these to achieve deeper video understanding. Furthermore, since our work does not explicitly leverage raw audio signals or audio features (beyond what is indirectly available through dialogue text), future work could include tri-modal input (video, text, audio) for better grounding. Additionally, explicit error correction or a joint optimization of both modules could be explored to mitigate the propagation of factual errors from Module-1 to Module-2.

References

Manoj Acharya, Kushal Kafle, and Christopher Kanan. 2019. Tallyqa: Answering complex counting questions. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 8076–8084.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, and 8 others. 2022. Flamingo: a visual language model for few-shot learning. *Preprint*, arXiv:2204.14198.

Husam Ali, Yllias Chali, and Sadid A Hasan. 2010. Automatic question generation from sentences. In *Actes de la 17e conférence sur le Traitement Automatique des Langues Naturelles. Articles courts*, pages 213–218.

Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. 2016. Neural module networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 39–48.

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.

Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, Thomas Unterthiner, Daniel Keysers, Skanda Koppula, Fangyu Liu, Adam Grycner, Alexey Gritsenko, Neil Houlsby, Manoj Kumar, Keran Rong, and 16 others. 2024. Paligemma: A versatile 3b vlm for transfer. *Preprint*, arXiv:2407.07726.

Santiago Castro, Naihao Deng, Pingxuan Huang, Mihai Burzo, and Rada Mihalcea. 2022. Wildqa: In-the-wild video question answering. *arXiv preprint arXiv:2209.06650*.

Ankush Chatterjee, Manish Gupta, and Puneet Agrawal. 2020. Faqaugmenter: suggesting questions for enterprise faq pages. In *Proceedings of the 13th international conference on web search and data mining*, pages 829–832.

Seongho Choi, Kyoung-Woon On, Yu-Jung Heo, Ahjeong Seo, Youwon Jang, Minsu Lee, and Byoung-Tak Zhang. 2021. Dramaqa: Character-centered video story understanding with hierarchical qa. In *Proceedings of the aaai conference on artificial intelligence*, volume 35, pages 1166–1174.

Anthony Colas, Seokhwan Kim, Franck Dernoncourt, Siddhesh Gupte, Daisy Zhe Wang, and Doo Soon Kim. 2019. Tutorialvqa: Question answering dataset for tutorial videos. *arXiv preprint arXiv:1912.01046*.

Daniel Cores, Michael Dorkenwald, Manuel Mucientes, Cees GM Snoek, and Yuki M Asano. 2024. Tvbench: Redesigning video-language evaluation. *arXiv preprint arXiv:2410.07752*.

Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. 2016. Multimodal compact bilinear pooling for visual question answering and visual grounding. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. 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.

Madeleine Grunde-McLaughlin, Ranjay Krishna, and Maneesh Agrawala. 2021. Agqa: A benchmark for compositional spatio-temporal reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11287–11297.

Pranay Gupta and Manish Gupta. 2022. Newskvqa: Knowledge-aware news video question answering. In *Pacific-asia conference on knowledge and data mining*, pages 3–15. Springer.

Unnat Jain, Ziyu Zhang, and Alexander G Schwing. 2017. Creativity: Generating diverse questions using variational autoencoders. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6485–6494.

Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. 2017. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2758–2766.

Baoxiong Jia, Ting Lei, Song-Chun Zhu, and Siyuan Huang. 2022. Egotaskqa: Understanding human tasks in egocentric videos. *Advances in Neural Information Processing Systems*, 35:3343–3360.

Amrith Krishna, Plaban Bhowmick, Krishnendu Ghosh, Archana Sahu, and Subhayan Roy. 2015. Automatic generation and insertion of assessment items in online video courses. In *Proceedings of the 20th International Conference on Intelligent User Interfaces Companion*, pages 1–4.

Vaibhav Kumar and Alan W Black. 2020. Clarq: A large-scale and diverse dataset for clarification question generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7296–7301.

J. Lei, L. Yu, M. Bansal, and T. L. Berg. 2021. Tvqa+: Spatio-temporal grounding for video question answering. In *Proceedings of CVPR 2021*.

J. Lei, Z. Yu, and M. Bansal. 2018. Tvqa: Localized, compositional video question answering. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2911–2921.

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.

Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal. 2020b. What is more likely to happen next? video-and-language future event prediction. *arXiv preprint arXiv:2010.07999*.

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.

Jiapeng Li, Ping Wei, Wenjuan Han, and Lifeng Fan. 2023a. Intentqa: Context-aware video intent reasoning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11963–11974.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, and 1 others. 2024. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22195–22206.

Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. 2020. Hero: Hierarchical encoder for video+ language omni-representation pre-training. *arXiv preprint arXiv:2005.00200*.

Linjie Li, Jie Lei, Zhe Gan, Licheng Yu, Yen-Chun Chen, Rohit Pillai, Yu Cheng, Luowei Zhou, Xin Eric Wang, William Yang Wang, and 1 others. 2021. Value: A multi-task benchmark for video-and-language understanding evaluation. *arXiv preprint arXiv:2106.04632*.

Yanwei Li, Yuechen Zhang, Chengyao Wang, Zhisheng Zhong, Yixin Chen, Ruihang Chu, Shaoteng Liu, and Jiaya Jia. 2023b. Mini-gemini: Mining the potential of multi-modality vision language models. *arXiv:2403.18814*.

Yuncheng Li, Yale Song, Liangliang Cao, Joel Tetreault, Larry Goldberg, Alejandro Jaimes, and Jiebo Luo. 2016. Tgif: A new dataset and benchmark on animated gif description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4641–4650.

Feng Liu, Tao Xiang, Timothy M Hospedales, Wankou Yang, and Changyin Sun. 2018. Ivqa: Inverse visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8611–8619.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Preprint*, arXiv:2001.08210.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. 2024a. Videogpt+: Integrating image and video encoders for enhanced video understanding. *arXiv preprint arXiv:2406.09418*.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. 2024b. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024)*.

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.

Andrés Marafioti, Orr Zohar, Miquel Farré, Merve Noyan, Elie Bakouch, Pedro Cuenca, Cyril Zakka, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, Vaibhav Srivastav, Joshua Lochner, Hugo Larcher, Mathieu Morlon, Lewis Tunstall, Leandro von Werra, and Thomas Wolf. 2025. Smolvlm: Redefining small and efficient multimodal models. Preprint.

Rahul Mehta, Bhavyajeet Singh, Vasudeva Varma, and Manish Gupta. 2024. Circuitvqa: A visual question answering dataset for electrical circuit images. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 440–460. Springer.

A. Miech, D. Zhukov, J. B. Alayrac, M. Tapaswi, I. Laptev, and J. Sivic. 2019a. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. *Proceedings of the IEEE International Conference on Computer Vision*, pages 2630–2639.

Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. 2019b. HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips. In *ICCV*.

Nasrin Mostafazadeh, Ishan Misra, Jacob Devlin, Margaret Mitchell, Xiaodong He, and Lucy Vanderwende. 2016. Generating natural questions about an image. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1802–1813.

Arsha Nagrani, Mingda Zhang, Ramin Mehran, Rachel Hornung, Nitesh Bharadwaj Gundavarapu, Nilpa Jha, Austin Myers, Xingyi Zhou, Boqing Gong, Cordelia Schmid, and 1 others. 2024. Neptune: The long orbit to benchmarking long video understanding. *arXiv preprint arXiv:2412.09582*.

Liangming Pan, Wenqiang Lei, Tat-Seng Chua, and Min-Yen Kan. 2019. Recent advances in neural question generation. *arXiv preprint arXiv:1905.08949*.

Liangming Pan, Yuxi Xie, Yansong Feng, Tat-Seng Chua, and Min-Yen Kan. 2020a. Semantic graphs for generating deep questions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1463–1475.

Youcheng Pan, Baotian Hu, Qingcai Chen, Yang Xiang, and Xiaolong Wang. 2020b. Learning to generate diverse questions from keywords. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8224–8228. IEEE.

Paritosh Parmar, Eric Peh, Ruirui Chen, Ting En Lam, Yuhan Chen, Elston Tan, and Basura Fernando. 2024. Causalchaos! dataset for comprehensive causal action question answering over longer causal chains grounded in dynamic visual scenes. *Advances in Neural Information Processing Systems*, 37:92769–92802.

Maitreya Patel, Tejas Gokhale, Chitta Baral, and Yezhou Yang. 2022. Cripp-vqa: Counterfactual reasoning about implicit physical properties via video question answering. *arXiv preprint arXiv:2211.03779*.

Viorica Patraucean, Lucas Smaira, Ankush Gupta, Adria Recasens, Larisa Markeeva, Dylan Banarse, Skanda Koppula, Mateusz Malinowski, Yi Yang, Carl Doersch, and 1 others. 2023. Perception test: A diagnostic benchmark for multimodal video models. *Advances in Neural Information Processing Systems*, 36:42748–42761.

Arpan Phukan, Manish Gupta, and Asif Ekbal. 2024. ECIS-VQG: Generation of entity-centric information-seeking questions from videos. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14411–14436, Miami, Florida, USA. Association for Computational Linguistics.

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

Ruchit Rawal, Khalid Saifullah, Miquel Farré, Ronen Basri, David Jacobs, Gowthami Somepalli, and Tom Goldstein. 2024. Cinepile: A long video question answering dataset and benchmark. *arXiv preprint arXiv:2405.08813*.

Sanket Shah, Anand Mishra, Naganand Yadati, and Partha Pratim Talukdar. 2019. Kvqa: Knowledge-aware visual question answering. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 8876–8884.

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.

George Tom, Minesh Mathew, Sergi Garcia-Bordils, Dimosthenis Karatzas, and CV Jawahar. 2023. Reading between the lanes: Text videoqa on the road. In *International Conference on Document Analysis and Recognition*, pages 137–154. Springer.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. Musique: Multi-hop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554.

Paul Voigtlaender, Soravit Changpinyo, Jordi Pont-Tuset, Radu Soricut, and Vittorio Ferrari. 2023. Connecting Vision and Language with Video Localized Narratives. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. 2023. Videomae v2: Scaling video masked autoencoders with dual masking. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14549–14560.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Peng Wang, Qi Wu, Chunhua Shen, Anthony Dick, and Anton Van Den Hengel. 2017. Fvqa: Fact-based visual question answering. *IEEE transactions on pattern analysis and machine intelligence*, 40(10):2413–2427.

Bo Wu, Shoubin Yu, Zhenfang Chen, Joshua B Tenenbaum, and Chuang Gan. 2024. Star: A benchmark for situated reasoning in real-world videos. *arXiv preprint arXiv:2405.09711*.

Y. Xiao, X. Li, and 1 others. 2021. Next-qa: A dataset for causal and temporal reasoning in video. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4093–4103.

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

J. Xu, T. Mei, and 1 others. 2016a. Msr-vtt: A large video description dataset for bridging video and language. *Proceedings of the 2016 ACM Conference on Multimedia Conference*, pages 1086–1094.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016b. Msr-vtt: A large video description dataset for bridging video and language. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5288–5296.

Li Xu, He Huang, and Jun Liu. 2021. Sutd-trafficqa: A question answering benchmark and an efficient network for video reasoning over traffic events. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9878–9888.

Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. 2021. 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.

Antoine Yang, Arsha Nagrani, Paul Hongsuck Seo, Antoine Miech, Jordi Pont-Tuset, Ivan Laptev, Josef Sivic, and Cordelia Schmid. 2023. Vid2seq: Large-scale pretraining of a visual language model for dense video captioning. In *CVPR*.

Dongjie Yang, Suyuan Huang, Chengqiang Lu, Xiaodong Han, Haoxin Zhang, Yan Gao, Yao Hu, and Hai Zhao. 2024. Vript: A video is worth thousands of words. *Advances in Neural Information Processing Systems*, 37:57240–57261.

Jianwei Yang, Jiasen Lu, Stefan Lee, Dhruv Batra, and Devi Parikh. 2018a. Visual curiosity: Learning to ask questions to learn visual recognition. In *Conference on Robot Learning*, PMLR, pages 63–80.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018b. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Y. Yu, J. Kim, and 1 others. 2019a. Activitynet-qa: A dataset for video question answering. *Proceedings of the 2019 IEEE International Conference on Computer Vision*, pages 6731–6740.

Youngjae Yu, Jongseok Kim, and Gunhee Kim. 2018. A joint sequence fusion model for video question answering and retrieval. In *Proceedings of the European conference on computer vision (ECCV)*, pages 471–487.

Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. 2019b. Activitynet-qa: A dataset for understanding complex web videos via question answering. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33:9127–9134.

Amir Zadeh, Michael Chan, Paul Pu Liang, Edmund Tong, and Louis-Philippe Morency. 2019. Social-iq: A question answering benchmark for artificial social intelligence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8807–8817.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From recognition to cognition: Visual commonsense reasoning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Shijie Zhang, Lizhen Qu, Shaodi You, Zhenglu Yang, and Jiawan Zhang. 2017. Automatic generation of grounded visual questions. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 4235–4243.

A Appendix

B Automated MVQ-60 Dataset Generation

We propose a rule-based algorithm for generating multi-hop questions by merging zero-hop question-answer pairs, inspired by MUSIQUE (Trivedi et al., 2022):

The algorithm processes filtered QA pairs (len(q) $\leq$15, len(a) $\leq$3) grouped by episode. For valid pairs from different segments where a_2 appears in q_1, it replaces a_2 in q_1 with q_2 to form multi-hop questions requiring cross-segment reasoning.

C Models Evaluated

We evaluate the following pretrained models for MVQG:

Qwen2-VL-2B-Instruct (Wang et al., 2024): A vision-language model optimized for instruction-following tasks, capable of integrating visual inputs with complex text prompts. It was tested using raw video frames and associated transcripts.

Algorithm 1 MVQ-60 Generation

Require: Set of zero-hop QA pairs $\{(q_i, a_i, m_i)\}$ with metadata $m_i = (e_i, s_i)$
Ensure: Set of multi-hop questions $\mathcal{Q}_{\text{multi}}$

```
Filter QA pairs: Q_filtered ← {(q, a) | len(q) ≤ 15, len(a) ≤ 3}
for each episode e do
    Group pairs: G_e ← {(q_i, a_i, s_i) | e_i = e}
    for each pair (q_1, a_1, s_1), (q_2, a_2, s_2) ∈ G_e × G_e do
        if s_1 ≠ s_2 and a_2 is substring of q_1 then
            q_merged ← replace(q_1, a_2, q_2)
            Q_multi ← Q_multi ∪ {q_merged}
        end if
    end for
end for
```

Table 3: Comparison of Different Approaches

Approach	Text	Video	Multihop Reasoning	Video Relevance
Text QG	✓	×	✓	×
Video QG	✓	✓	×	✓
Zeroshot	✓	✓	✓	×
Our Approach	✓	✓	✓	✓

SmolVLM (Marafioti et al., 2025): A lightweight multimodal large language model designed for efficient inference on vision-language tasks. Despite its compact size, SmolVLM demonstrated strong performance on zero-hop VQA tasks, but its ability to handle multi-hop reasoning remained untested prior to our evaluation.

MGM-2B (Li et al., 2023b): A 2-billion parameter multimodal generative model designed for cross-modal understanding and generation tasks. It processes video frame embeddings and textual data simultaneously, offering a comprehensive baseline for multimodal reasoning.

PaliGemma (Beyer et al., 2024): A recent multimodal model designed by Google DeepMind for VQA and vision-language reasoning tasks. Although It excels in vision-language alignment, its capacity for multi-hop reasoning with video content was tested in our experiments.

ECIS-VQG (Phukan et al., 2024): A VideoQG model, which was designed to produce entity-centric, information-seeking questions grounded in video content. Originally proposed for Zero-hop VideoQA tasks, we include ECIS-VQG to evaluate its capacity for generalizing to the multi-hop setting when finetuned on our MVQ-60 dataset.

D Datasets Explored

To contextualize the need for a dedicated multihop video question answering (MVQG) dataset and to inform our design choices, we conducted a comprehensive survey of existing video question answering (VideoQA) datasets. This exploration encompassed approximately 40 publicly available datasets, each with varying characteristics in terms of scale, domain, question type, and associated annotations. A summary of these datasets, including their approximate size, primary focus, and question types, is provided in Table 10.

While these datasets have significantly advanced the field of VideoQA, they primarily focus on zero-hop questions that can be answered by directly attending to specific segments or elements within a single video clip. As highlighted in **Table 10**, these datasets cover a diverse range of domains.

Our analysis of these existing resources revealed a critical gap: the absence of datasets specifically designed to evaluate and drive research in multihop video question answering. As detailed in the main body, multihop questions require reasoning across multiple temporal segments or understanding the relationships between different events or entities within and potentially across video clips. The existing datasets, while valuable for zero-hop VideoQA, do not adequately support the investigation of these more complex reasoning capabilities.

To address this limitation and facilitate research into MVQG, we undertook the creation of a novel dataset, leveraging the TVQA+ dataset as a foundation, as described in Section:dataset-creation in main paper. Our approach to automatically generating multihop questions aimed to create a scalable resource for evaluating models capable of performing temporal and relational reasoning across video content.

Dataset Evaluation

We prompted GPT5 to evaluate the dataset question quality. Due to cost considerations, we evaluated the same 200-question sample with 300 additional questions.

Prompt: You are an evaluator. Your task is to assess a generated question based on the context and the following criteria:

Fluency: Evaluates the grammatical correctness and naturalness of the generated question. 0: Poor (Grammatical errors and awkward phrasing). Example: "What doing is Chandler wife cooking?"

1: Fair (Some grammatical errors, but understandable). Example: "What Chandler wife cooking?" 2: Good (Grammatically correct). Example: "What is the wife of Chandler cooking?" 3: Excellent (Fluent and natural language with no errors). Example: "What is Chandler's wife cooking?"

Relevance: Assesses the extent to which the generated question pertains to the content of the provided video clips. 0: Irrelevant (Does not relate to the video). Example: "What is the capital of France?" 1: Slightly relevant (Partially relates to the video). Example: "Are there any people in the video?" 2: Mostly relevant (Mostly relates to the Videos). Example: "Are there people dancing in these videos?" 3: Highly relevant (Directly relates to the images). Example: "What is the connection between the people dancing in these Videos?"

Multi-Hop Reasoning: Evaluates the complexity of reasoning required to answer the generated question based on the provided video clips. 0: Single-hop (Only needs one Video for the answer). Example: "Is Monica dancing in the first video?" 1: Simple multi-hop (Requires basic information from both videos). Example: "Are People dancing in both Videos?" 2: Intermediate multi-hop (Requires more complex connections between images). Example: "Are the same people dancing in both videos?" 3: Advanced multi-hop (Involves detailed reasoning using both images). Example: "what is the relation between the common people dancing in both videos?"

Engagingness: Evaluates how interesting and captivating the generated question is to a human observer. 0: Not engaging (Boring or uninteresting). Example: "Are there hats in the videos?" 1: Slightly engaging (Mildly interesting). Example: "What colours are the hats in the videos?" 2: Moderately engaging (Interesting and engaging). Example: "How do the styles of hats in the videos differ?" 3: Highly engaging (Very interesting and captivating). Example: "What do the hats in the videos reveal about the event going on and time period of the scenes depicted?"

Factual Correctness: Evaluates whether the generated question contains any factual inaccuracies. 0: Incorrect (factually incorrect). Example: "Why does Chandler want to leave after hanging out with the group, which includes Amy and Emma?" (Emma and Amy both are wrong). 1: Mostly Incorrect (contains major factual errors, though a small part of the content may be accurate). Example: "Why does Chandler want to leave after hanging out with the group, which includes Joey and Amy?" (Joey is correct, Amy is wrong). 2: Partially Correct (factually accurate in the main aspect but contains a minor mistake or omission). Example: "Why does Chandler want to leave after hanging out with the group, which includes Joey?" 3: Factually correct. Example: "Why does Chandler want to leave after hanging out with the group, which includes Joey and Monica?"

Inclusiveness: Evaluates whether the generated question is inclusive and avoids any potentially biased or discriminatory language. 0: Not inclusive (The question contains biased or discriminatory language or assumptions). Example: "Why are the women in the video acting emotionally?" 1: Slightly inclusive (The question is mostly neutral but could be phrased more inclusively). Example: "What are the people in the video doing?" (If the context strongly implies a specific gender) 2: Moderately inclusive (The question attempts to use neutral language but might still have some underlying assumptions). Example: "What is the role of each person in the scene?" 3: Highly inclusive (The question uses neutral and respectful language, avoiding any biased or discriminatory assumptions about gender, race, age, etc.). Example: "What actions are the individuals performing in the video?"

***Question** **Context** Output Structure: Fluency: value, Relevance: value, Multi-Hop Reasoning: value, Engagingness: value, Factual Correctness: value, Inclusiveness: value,*

Human Evaluation Metrics

To assess the quality of the generated multihop video questions, we employed a comprehensive human evaluation protocol using the following set of metrics. Each metric was evaluated on a 4-point scale(0 to 3), with higher scores indicating better quality according to the specific criterion.

Fluency: Evaluates the grammatical correctness and naturalness of the generated question.

- **0**: Poor (Grammatical errors and awkward phrasing). *Example: "What doing is Chandler wife cooking?"*
- **1**: Fair (Some grammatical errors but understandable). *Example: "What Chandler wife cooking?"*
- **2**: Good (Grammatically correct). *Example: "What is the wife of Chandler cooking?"*

Table 4: Human and GPT Evaluation the Dataset, MVQ-60

Method	Fluency	Relevance	Multi-Hop Reasoning	Engagingness	Factual Correctness	Inclusiveness
Human Eval	**2.92**	**3**	**3**	**2.8**	**3**	**3**
gpt-5-nano-2025-08-07	2.88	3	2.82	2.66	2.74	2.96

- **3**: Excellent (Fluent and natural language with no errors). *Example: "What is Chandler's wife cooking?"*

Relevance: Assesses the extent to which the generated question pertains to the content of the provided video clips.

- **0**: Irrelevant (Does not relate to the video). *Example: "What is the capital of France?"*
- **1**: Slightly relevant (Partially relates to the video). *Example: "Are there any people in the video?"*
- **2**: Mostly relevant (Mostly relates to the Videos). *Example: "Are there people dancing in these videos?"*
- **3**: Highly relevant (Directly relates to the images). *Example: "What is the connection between the people dancing in these Videos?"*

Multi-Hop Reasoning: Evaluates the complexity of reasoning required to answer the generated question based on the provided video clips.

- **0**: Single-hop (Only needs one Video for the answer). *Example: "Is Monica dancing in the first video?"*
- **1**: Simple multi-hop (Requires basic information from both videos). *Example: "Are People dancing in both Videos?"*
- **2**: Intermediate multi-hop (Requires more complex connections between images). *Example: "Are the same people dancing in both videos?"*
- **3**: Advanced multi-hop (Involves detailed reasoning using both images). *Example: "what is the relation between the common people dancing in both videos?"*

Engagingness: Evaluates how interesting and captivating the generated question is to a human observer.

- **0**: Not engaging (Boring or uninteresting). *Example: "Are there hats in the videos?"*
- **1**: Slightly engaging (Mildly interesting). *Example: "What colours are the hats in the videos?"*
- **2**: Moderately engaging (Interesting and engaging). *Example: "How do the styles of hats in the videos differ?"*
- **3**: Highly engaging (Very interesting and captivating). *Example: "What do the hats in the videos reveal about the event going on and time period of the scenes depicted?"*

Factual Correctness: Evaluates whether the generated question contains any factual inaccuracies.

- **0**: Incorrect (factually incorrect). *Example: "Why does Chandler want to leave after hanging out with the group, which includes Amy and Emma?"*
- **3**: Factually correct. *Example: "Why does Chandler want to leave after hanging out with the group, which includes joey and monica?"*

Inclusiveness: Evaluates whether the generated question is inclusive and avoids any potentially biased or discriminatory language.

- **0**: Not inclusive (The question contains biased or discriminatory language or assumptions). *Example: "Why are the women in the video acting emotionally?"*
- **1**: Slightly inclusive (The question is mostly neutral but could be phrased more inclusively). *Example: "What are the people in the video doing?" (If the context strongly implies a specific gender)*
- **2**: Moderately inclusive (The question attempts to use neutral language but might still have some underlying assumptions). *Example: "What is the role of each person in the scene?"*
- **3**: Highly inclusive (The question uses neutral and respectful language, avoiding any biased or discriminatory assumptions about gender, race, age, etc.). *Example: "What actions are the individuals performing in the video?"*

Examples from Our Multihop Video Question Generation Dataset

To illustrate the characteristics of the multihop questions within our newly created dataset MVQ-60, we present a selection of examples in Table 5. Each row includes the constituent questions, their respective answers, the involved video clip names, and the resulting initial multihop question. It is important to note that these are the initial multihop (merged) questions generated by our system, after paraphrasing they often exhibit improved fluency and naturalness.

Note: Due to space constraints, we have presented a subset of the generated multihop questions. The full dataset contains a diverse range of questions requiring various forms of temporal and relational reasoning across different video segments from the TV show "Friends". The structure of these examples demonstrates how our dataset links information across potentially non-contiguous video clips through the composition of simpler questions

D.1 Ablation Study

To better understand the contributions of various components in **VideoChain**, we conducted an ablation study by modifying the model in two key ways: (1) removing the video embeddings to evaluate the reliance on textual information, and (2) simplifying the architecture into a single-component system to directly generate multi-hop questions. These experiments highlight the significance of both the multimodal input and the modular design in enhancing the model's performance on multi-hop question generation tasks. The results are shown in Table 8

Text-Only Model: In this configuration, we removed video embeddings and trained the model using only textual data—transcripts and prompts. This setup was designed to assess how much the model relies on visual information versus language alone when generating coherent and contextually grounded multi-hop questions. Compared to the full model, the text-only version showed a drop across all metrics. Most notably, video relevance fell from 2.91 $\rightarrow$ 2.09, and multi-hop reasoning dropped from 2.85 $\rightarrow$ 1.54. While fluency remained relatively high (2.81 $\rightarrow$ 2.66), the absence of visual grounding led to decreased factual correctness (2.97 $\rightarrow$ 2.36) and overall inclusiveness (2.78 $\rightarrow$ 2.05). This suggests that visual information plays a critical role in enabling richer, more contextually grounded question generation.

Single-Component Multi-Hop Generation (Direct Multihop Generation): In this variant, we removed the modular design and trained the model to generate multi-hop questions in a single stage, directly from paired video segments and transcripts. While this setting retained multimodal inputs, it lacked the iterative structure of the full model. As shown in Table 8, this led to a sharp drop in performance, particularly in *multi-hop reasoning* (1.24 vs. 2.85) and *factual correctness* (1.98 vs. 2.97). The model often produced concatenated or shallow questions, failing to perform genuine multi-step reasoning. This highlights the importance of structured decomposition for handling complex, compositional question generation.

Evaluation and Comparative Metrics: Both ablation settings were evaluated using the same metrics as the full model: *fluency*, *relevance*, *multi-hop reasoning complexity*, and *factual correctness*. Let S_{full}, $S_{\text{text-only}}$, and $S_{\text{direct-mh}}$ denote the average scores for the full model, the text-only model, and the single-component model, respectively. The relative performance drop ΔS for each ablation is computed as:

$$\Delta S_{\text{text-only}} = S_{\text{full}} - S_{\text{text-only}} \quad (1)$$

$$\Delta S_{\text{direct-mh}} = S_{\text{full}} - S_{\text{direct-mh}}. \quad (2)$$

We observe $\delta S_{\text{direct-mh}}$ to be 0.76 and $\delta S_{\text{text-only}}$ to be 0.69, the significant performance degradation in the **text-only** model, is due to tasks requiring visual grounding. Similarly, the **single-component model** is observed to underperform on tasks requiring complex multi-hop reasoning due to the absence of iterative refinement. The results of ablation study shown in table 8, underscores the critical role of **video embeddings** in grounding questions in visual context and highlights the effectiveness of our model's **modular design** in handling multi-hop reasoning.

E Qualitative Examples of MVQG Error Types

We conducted a detailed error analysis to understand the common failure modes in our MVQG framework. Table 9 summarizes the frequency of each error category before and after applying mitigation strategies. Below, we elaborate on the nature of each error, provide examples, and describe how each was addressed.

Multi-Hop Reasoning Failures: The model often concatenated two zero-hop questions using

Table 5: Examples from Our Multihop Video Question Generation Dataset

Question 1	Answer 1	Video Clip 1	Question 2	Answer 2	Video Clip 2	Merged Question
Who was talking on the phone before Joey picked up the phone the first time?	Ross	friends, s02e01, seg02, clip, 07	Who was Joey talking with when Ross went inside?	Joey was talking with his dad	friends, s02e01, seg02, clip, 21	Who was Joey talking with when , the person Who was talking on the phone before Joey picked up the phone the first time?, went inside?
Who was talking on the phone before Joey picked up the phone the first time?	Ross	friends, s02e01, seg02, clip, 07	Where did Ross went after the conversation with Rachel?	Ross went inside the house	friends, s02e01, seg02, clip, 21	Where did , the person Who was talking on the phone before Joey picked up the phone the first time?, went after the conversation with Rachel?
Who does Charlie disagree knows art when Ross mentions him/her?	Joey	friends, s09e21, seg02, clip, 18	Why does Joey joke with Ross after he gives suggestions for his date?	Joey jokes becasue Ross has detailed ideas specific to Joey's date's preferences.	friends, s09e21, seg02, clip, 08	Why does , the person Who does Charlie disagree knows art when Ross mentions him/her?, joke with Ross after he gives suggestions for his date?
Who does Charlie disagree knows art when Ross mentions him/her?	Joey	friends, s09e21, seg02, clip, 18	Why doesn't Joey know what he just said after getting asked by Ross?	His brain is thinking about monster trucks	friends, s09e21, seg02, clip, 12	Why doesn't , the person Who does Charlie disagree knows art when Ross mentions him/her?, know what he just said after getting asked by Ross?
Who came to the room when Castle was talking?	Ryan	castle, s06e21, seg02, clip, 16	Who comes looking for Ryan after he hangs up the phone?	Esposito comes looking for Ryan.	castle, s06e21, seg02, clip, 11	Who comes looking for , the person Who came to the room when Castle was talking?, after he hangs up the phone?
Who came to the room when Castle was talking?	Ryan	castle, s06e21, seg02, clip, 16	What is Lanie waving around in her hand when she is facing Ryan and Esposito?	A pen.	castle, s06e21, seg02, clip, 18	What is Lanie waving around in her hand when she is facing , the person Who came to the room when Castle was talking?, and Esposito?
Who follows beckett out of montgomerys office after she leaves montgomerys office?	Castle	castle, s03e22, seg02, clip, 03	What type of cup does Castle sit by when he clasps his hands?	Wine glass.	castle, s03e22, seg02, clip, 15	What type of cup does , the person Who follows beckett out of montgomerys office after she leaves montgomerys office?, sit by when he clasps his hands?
Who follows beckett out of montgomerys office after she leaves montgomerys office?	Castle	castle, s03e22, seg02, clip, 03	Who started jumping onto Beckett after Castle opened the door?	Seeger	castle, s03e22, seg02, clip, 22	Who started jumping onto Beckett after , the person Who follows beckett out of montgomerys office after she leaves montgomerys office?, opened the door?

"and" instead of forming a meaningful reasoning chain.

Incorrect: "What was Amy holding when she is talking to Penny **and** who was with Sheldon in the car?"

Expected: "What was Amy holding when she is talking to the girl who was in the car with Sheldon?"

We mitigated this by refining training prompts with clearer examples and post-processing rules, reducing the error rate from 24% to 6%.

Factual Inaccuracy: Some questions introduced incorrect or unsupported claims not grounded in the video.*Example*: "What did Rachel say when she got the job offer in Paris?" (scene not present).

We applied negative sampling and semantic filters to reduce such cases from 12% to 7%.

Semantic Drift: The model occasionally produced abstract or off-topic questions. *Example*: "How do characters reflect on their past experiences?"

Prompt refinement and focused sampling during fine-tuning reduced this issue from 11% to 5%.

Grammatical Errors: Syntax or fluency issues such as tense mismatch or fragmented clauses. *Example*: "What do Monica was cooking when Ross came in?"

Due to consistent training and BART's language generation strength, these errors decreased from 5% to 3%.

Redundancy: Some questions repeated the same concepts across hops. *Example*: "Who was with Rachel and who was talking to Rachel in the kitchen?"

We filtered such patterns post-merging, reducing redundancy from 9% to 4%.

Ambiguity or Vagueness: Questions lacked

Table 6: Qualitative examples of MVQG generations.

Series	Correct Generation	Multi-Hop Failure	External Knowledge Leakage	Hallucination
BBT	What did , the person who wishes Sheldon a happy Valentines Day after he opens the door?, do after Sheldon told her he was being selfish?	What does the person who is eating with Sheldon say and who knocks on the door when Sheldon is talking?	What does , the person who is talking to Amy about moving in together after returning from Princeton?, say when they discuss their living arrangements?	What does , the person who gives Sheldon a Star Wars gift during his birthday party?, say when he opens the present?
Friends	What is , the person who came into the apartment when Leonard was on the phone?, holding when she is talking to Leonard?	What is the person who is sitting next to Monica doing and what does Rachel say when she enters the room?	What does , the person who is talking to Rachel when she says she got off the plane?, do after hearing her decision?	What does , the person who is cooking Thanksgiving dinner with Monica?, do when Joey accidentally drops the turkey?
HIMYM	What does , the person who is in charge when Howard and Sheldon work together?, do when he is talking to Howard and Raj?	What does the person who gives Ted a drink say and who walks into the apartment when Barney is talking?	What does , the person who is sitting in front of Ted when he starts telling the story about how he met their mother?, say when he gives them something?	What does , the person who plays the guitar during Robin's farewell party?, say when Ted offers a toast?
Grey's Anatomy	What did , the person who is playing the piano when everyone is singing to Bernadette?, say when he was talking to Raj?	What does the patient who is lying on the bed do and who enters the room when Meredith is looking at the monitor?	What does , the person who is talking to Meredith before Derek's accident?, say when she expresses her concern?	What does , the person who argues with Cristina about the heart surgery?, do when the patient flatlines?
Castle	What did , the person who is wearing a brown trench coat when Beckett enters the alley?, say when she finds the second clue?	What is the person who found the evidence doing and what does Beckett say when she enters the office?	What does , the person who is confronting Castle when he finds out who killed his mother?, do when Castle reacts?	What does , the person who brings evidence to Beckett during the rooftop chase?, say when she finds the clue?
House	What does , the person who walks into the patient's room after House finishes speaking with the nurse?, do when he notices the charts are missing?	What does the nurse who checks the IV bag say and what is House doing when he talks to Wilson?	What does , the person who is inside the house when House crashes his car into it?, say when House approaches?	What does , the person who challenges House's diagnosis in the operating room?, do when the patient wakes up?

clarity or contained poorly grounded references. *Example*: "What did he do after she left?"
This was addressed through coreference-aware sampling, reducing such errors from 14% to 6%.

External Knowledge Leakage: The model occasionally used memorized facts beyond input scope. *Examples*:
"What does Sheldon say to Leonard before they move to their new apartment?", "How does Ross react when Rachel leaves for Paris?"
By constraining context and refining prompts, hallucination was reduced from 20% to 8%, and leakage from 32% to 8%.

Correct generations demonstrate well-formed multi-hop questions that refer to grounded visual-textual content and follow the long-form template described in Section 6.4. Multi-hop reasoning failures typically involve the incorrect use of conjunctions (e.g., "and") instead of forming compositional reasoning chains. External knowledge leakage reveals the model's tendency to rely on memorized facts from the broader TV series rather than the provided scene. Hallucination refers to entirely fabricated content not grounded in the input video or known series lore.

These examples highlight the diversity of errors our model encounters and the necessity for strong grounding and compositional reasoning mechanisms in MVQG systems.

E.1 Prompt Variants and Failure Cases

As part of our baseline prompt engineering study, we investigated the impact of structured prompting strategies inspired by chain-of-thought reasoning. While such approaches have proven effective in text-based reasoning tasks, we found that they were not beneficial in our setting due to the combination of a smaller base model (BART-large) and the inherent complexity of the MVQG task.

Despite our initial expectations, prompts designed to simulate step-by-step reasoning often resulted in incoherent, vague, or incomplete outputs. Table 7 shows representative examples highlighting these failure cases.

These findings informed our final choice of using clear and direct prompts, such as: *"Based on the two video segments and their transcripts, generate a question that requires integrating information from both videos."* This prompt yielded more stable and contextually grounded outputs, particularly when paired with a well-trained zero-hop module. We include these prompt examples and results in the supplementary material for transparency and reproducibility.

Table 7: Examples of structured prompts and their outputs. These structured prompts often confused the model or led to verbose, incoherent generation.

Prompt	**Generated Output**
"Think step by step. First, look at the first video. Then, look at the second. Now generate a question that connects both."	"Okay. First video: Sheldon is holding something. Second video: Raj is cooking. What is happening in both?" *(Fails to form a coherent multi-hop link)*
"Understand what happens in the first clip, then the second, and finally write a question that links them."	"In the first clip, someone is talking. In the second, there is a kitchen. Can you describe both?" *(Vague; no reasoning or question structure)*
"Let's think carefully. Break down the events and create a multi-hop question based on them."	"Hmm. First event: someone holding sword. Second event: Raj in kitchen. What's the question here?" *(Fails to produce a question; stays in meta-mode)*

Table 8: Performance Comparison of Ablations

Model	**Fluency**	**Video Relevance**	**Multi-Hop Reasoning**	**Engagingness**	**Factual Correctness**	**Inclusiveness**
Full Model	2.81	2.91	2.85	2.75	2.92	2.78
Without Video	2.66	2.09	1.54	2.25	2.36	2.05
Single Component	2.31	2.39	1.24	2.32	1.98	2.24

Table 9: Improvements in Error categories

Error Category	**Error before improvement (%)**	**Error After improvement (%)**
Multihop Reasoning failure	24	6
Factual Inaccuracy	12	7
Semantic Drift	11	5
Grammatical Errors	5	3
Redundancy	9	4
Ambiguity or Vagueness	14	6
External Knowledge Leakage	32	8

Table 10: Summary of Explored Video Question Answering Datasets

No.	Dataset Name	Primary Focus	Short Description
1	MSR-VTT (Xu et al., 2016a)	Open Domain Captioning	A large-scale dataset primarily for video captioning, containing 10,000 videos with 20 human-annotated captions per video. It's also used as a base for VQA tasks.
2	HowTo100M (Miech et al., 2019a)	Instructional Videos	A massive dataset of over 1 million narrated instructional videos, focusing on explaining how to perform various tasks.
3	TVQA (Lei et al., 2018)	TV Shows (6)	A large-scale VideoQA dataset built upon 6 popular English-language TV shows, featuring multiple-choice questions and answers, along with subtitles and video frames.
4	ActivityNet-QA (Yu et al., 2019a)	Activity Understanding	Contains human-annotated question-answer pairs on videos from the ActivityNet dataset, designed to test models' long-term spatio-temporal reasoning abilities.
5	NExT-QA (Xiao et al., 2021)	Explanation of Video	A VideoQA benchmark specifically created to evaluate the explanation of video content, requiring models to reason about causal and temporal relationships between actions and objects.
6	TGIF-QA (Jang et al., 2017)	Animated GIFs	Features question-answer pairs for animated GIFs from the TGIF dataset, suitable for evaluating video-based Visual Question Answering techniques on short, dynamic visual content.
7	MovieQA (Tapaswi et al., 2016)	Movies	A dataset for question answering about movies, evaluating story comprehension from both video and textual sources like plot synopses and subtitles, with multiple-choice answers.
8	MVBench (Li et al., 2024)	Temporal Understanding	A comprehensive benchmark designed to evaluate the temporal understanding capabilities of multimodal large language models (MLLMs) across 20 diverse dynamic video tasks.
9	MSRVTT-QA (Xu et al., 2017)	VQA on MSR-VTT	A benchmark for Visual Question Answering created based on the MSR-VTT video captioning dataset, evaluating the ability to answer questions grounded in video content described by captions.
10	MSVD-QA (Xu et al., 2017)	VQA from MSVD	A VideoQA dataset generated from the descriptive sentences in the MSVD dataset, providing a large set of question-answer pairs based on short video snippets and their textual descriptions.
11	TVQA+ (Lei et al., 2021)	Visual Grounding in TVQA	An extension of the TVQA dataset that includes detailed bounding box annotations, explicitly linking depicted objects to visual concepts mentioned in the questions and answers for enhanced visual grounding.
12	TGIF (Tumblr GIF) (Li et al., 2016)	GIF Descriptions	A dataset of 100,000 animated GIFs collected from Tumblr, each accompanied by several descriptive sentences provided by humans.
13	VideoInstruct (Maaz et al., 2024b)	Video Instruction Following	A dataset comprising high-quality video and instruction pairs, used to train models like Video-ChatGPT to follow instructions presented in video format.
14	AGQA (Grunde-McLaughlin et al., 2021)	Spatio-Temporal Reasoning	A benchmark designed to evaluate compositional spatio-temporal reasoning in videos, focusing on understanding actions, their attributes, and their relationships within a scene.
15	VALUE (Li et al., 2021)	General V&L Understanding	A benchmark created to test the generalizability of video-and-language understanding models across a wide range of tasks, domains, and existing datasets.
16	How2QA (Li et al., 2020)	QA on HowTo100M	A VideoQA dataset collected on the same videos as the HowTo100M dataset, featuring multiple-choice questions and answers related to the instructional content of the videos.
17	iVQA (Liu et al., 2018)	Instructional Video QA	An open-ended VideoQA benchmark specifically for instructional videos, featuring multiple correct answer annotations for each question and requiring detailed video understanding.
18	IntentQA (Li et al., 2023a)	Social Activity Intents	A dataset focusing on the diverse intents behind actions observed in daily social activities, designed to evaluate models' ability to understand the underlying motivations in videos.
19	SUTD-TrafficQA (Xu et al., 2021)	Traffic Video QA	A dataset specifically focused on question answering related to traffic scenarios, requiring understanding of various events, objects, and their interactions within traffic videos.
20	STAR Benchmark (Wu et al., 2024)	Situated Reasoning	A benchmark aimed at evaluating how well models can capture and utilize present knowledge directly from the surrounding visual situations depicted in videos.
21	MSRVTT-MC (Yu et al., 2018)	Multiple Choice VQA on MSR-VTT	A multiple-choice video question-answering dataset created based on the MSR-VTT dataset, offering a different evaluation format compared to open-ended QA.
22	TVBench (Cores et al., 2024)	Temporal Understanding in VQA	A benchmark specifically created to evaluate the temporal understanding capabilities of VideoQA models, focusing on questions that require reasoning over time within the video.
23	Neptune (Nagrani et al., 2024)	Long Video QA	A dataset consisting of challenging question-answer-decoy sets for long-form videos (up to 15 minutes), pushing the limits of models' ability to understand and reason over extended video durations.
24	DramaQA (Choi et al., 2021)	Dialogue and Narrative Understanding	A dataset focused on two key aspects of video understanding: comprehending character dialogue and understanding the broader narrative flow in movies and TV series.
25	EgoTaskQA (Jia et al., 2022)	Egocentric Video QA	A benchmark containing balanced question-answer pairs for egocentric (first-person perspective) videos, primarily focusing on understanding the actions performed by the camera wearer.
26	Perception Test (Patraucean et al., 2023)	Perception and Reasoning	A benchmark designed to evaluate the fundamental perception and reasoning skills of multimodal models when processing video content.
27	VLEP (Lei et al., 2020b)	Video-and-Language Event Prediction	Contains examples of future event prediction in videos along with their textual rationales, testing the ability to anticipate what will happen next based on the observed context.
28	KnowIT VQA (Garcia et al., 2020)	QA on The Big Bang Theory	A video dataset with human-generated question-answer pairs specifically centered around the content and characters of the popular TV show "The Big Bang Theory," focusing on domain-specific knowledge.
29	MovieFIB (Maharaj et al., 2017)	Movie Fill-in-the-Blank	A benchmark featuring fill-in-the-blank style questions based on detailed descriptive video annotations created for the visually impaired, testing fine-grained visual understanding.

Table 11: Some More Video Question Answering Datasets

No.	Dataset Name	Primary Focus	Short Description
1	HowToVQA69M (Yang et al., 2021)	Large-Scale HowTo VQA	A very large-scale VideoQA dataset built upon the HowTo100M video dataset, containing approximately 69 million question-answer pairs.
2	TutorialVQA (Colas et al., 2019)	Answer Spans in Tutorials	A dataset designed for the task of finding specific answer spans within the transcripts of tutorial videos, with questions and manually collected answer spans.
32	Video Localized Narratives (Voigtlaender et al., 2023)	Vision and Language Connection	A dataset that explicitly connects vision and language by providing detailed, localized narratives describing objects and actions within specific regions of video frames.
3	CRIPP-VQA (Patel et al., 2022)	Counterfactual Reasoning	A dataset designed to evaluate counterfactual reasoning about implicit physical properties through video question answering, requiring models to understand "what if" scenarios.
4	CausalChaos! (Parmar et al., 2024)	Causal Video QA	A dataset specifically created for evaluating causal reasoning in video question answering, using animated content from Tom and Jerry cartoons to focus on cause-and-effect relationships.
5	CinePile (Rawal et al., 2024)	Long Video QA (Movies)	A question-answering-based dataset focused on the understanding of long-form video content, specifically utilizing movie data.
6	RoadTextVQA (Tom et al., 2023)	Text Understanding in Driving Videos	A dataset focused on the task of understanding text and signs present in videos captured from driving scenarios, crucial for applications in autonomous driving.
7	Social-IQ 2.0 (Zadeh et al., 2019)	Social Intelligence in Video	A dataset designed to evaluate the social intelligence of AI models in understanding videos, focusing on social interactions, nonverbal cues, and human behavior.
8	VCG+112K (Maaz et al., 2024a)	Video Instruction Following	Another large-scale dataset for video instruction following, containing over 112,000 video-instruction pairs for training models to execute tasks based on video instructions.
9	WildQA (Castro et al., 2022)	VQA in Outside Settings	A video understanding dataset comprising videos recorded in unconstrained, real-world outside environments, designed to evaluate the robustness of models in more natural settings.
10	Vript (Yang et al., 2024)	Fine-Grained Video-Text	A fine-grained video-text dataset featuring high-resolution videos and detailed, rich annotations, aiming for a deeper understanding of the relationship between visual and textual information in videos.