Grounded Multi-Hop VideoQA in Long-Form Egocentric Videos

Qirui Chen[1,2], **Shangzhe Di**[1,2], **Weidi Xie**[1*]

[1]School of Artificial Intelligence, Shanghai Jiao Tong University, China
[2]Coop. Medianet Innovation Center, Shanghai Jiao Tong University, China
{chen_qirui, dishangzhe, weidi}@sjtu.edu.cn

Abstract

This paper considers the problem of *Multi-Hop Video Question Answering (MH-VidQA)* in long-form egocentric videos. This task not only requires to answer visual questions, but also to localize multiple relevant time intervals within the video as visual evidences. We develop an automated pipeline to create multi-hop question-answering pairs with associated temporal evidence, enabling to construct a large-scale dataset for instruction-tuning. To monitor the progress of this new task, we further curate a high-quality benchmark, **MULTIHOP-EGOQA**, with careful manual verification and refinement. Experimental results reveal that existing multi-modal systems exhibit inadequate multi-hop grounding and reasoning abilities, resulting in unsatisfactory performance. We then propose a novel architecture, termed as **G**rounding Scattered **E**vidence with Large **L**anguage **M**odel (**GeLM**), that enhances multi-modal large language models by incorporating a grounding module to retrieve temporal evidence from videos using flexible grounding tokens. Trained on our visual instruction-tuning data, **GeLM** demonstrates improved multi-hop grounding and reasoning capabilities, setting a baseline for this new task. Furthermore, when trained on third-person view videos, the same architecture also achieves state-of-the-art performance on the single-hop VidQA benchmark, ActivityNet-RTL, demonstrating its effectiveness.

Code — https://qirui-chen.github.io/MultiHop-EgoQA

1 Introduction

With the rapid development of computer vision, the community has witnessed a significant interest in deploying vision systems within embodied agents, such as AR glasses (Datta et al. 2022) and humanoid robots (Majumdar et al. 2024). In such scenarios, the inputs are typically long, continuous video streams from a first-person perspective, capturing the world through the eyes of an agent, actively interacting with its environment. For the virtual assistants or physical robots to be useful, the ability to perform egocentric video question answering (VidQA) is crucial, stemming from two aspects: *first*, VidQA leverages language as a natural interface for human-machine interaction, thereby enhancing the usability and accessibility for the general public; *second*, it can encompass various vision tasks about the 'who', 'when', 'where', and 'what' of an individual's daily life, *e.g.*, action recognition, object detection, and scene understanding, thus acting as a robust and comprehensive benchmark for video understanding.

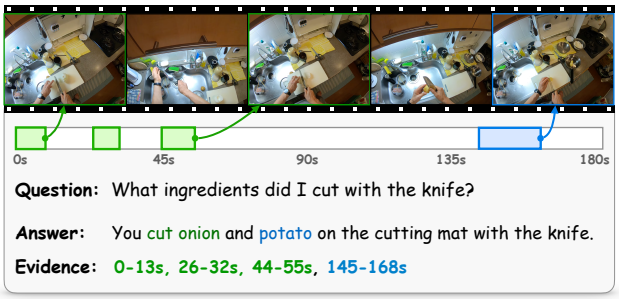

Figure 1: We introduce the problem of *Multi-Hop Video Question Answering* for long-form egocentric video understanding. This task requires the model to answer questions by gathering and reasoning across scattered visual clues, necessitating the grounding of multiple relevant time spans as supporting evidence.

Recently, the introduction of Ego4D dataset (Grauman et al. 2022) has enabled a series of research in visual-language understanding in egocentric videos, *e.g.*, question answering that focuses on summarizing entire video content (Mangalam, Akshulakov, and Malik 2023); natural language query (NLQ) that requires temporal localization based on a given query (Ramakrishnan, Al-Halah, and Grauman 2023); grounded question answering that considers answering the query, while localizing the query-related time span simultaneously (Bärmann and Waibel 2022). However, the above-mentioned settings often fall into an over-simplistic scenario, where questions are typically answerable based on visual cues from a single time point, or only one time span is annotated among multiple relevant spans. For instance, the question *"How many shirts did I pack in my suitcase?"* is deliberately excluded if the *packing* process occurs across multiple, non-contiguous time spans, as described in the annotation process of the NLQ task in Ego4D.

As a consequence, VidQA systems built upon the above-mentioned tasks can hardly be applied in multi-hop scenar-

*Corresponding author.

ios, due to two primary limitations: the scarcity of data supporting multi-hop reasoning, the deficiency in architecture design to support multi-hop temporal perception. *First*, there is a notable insufficiency of training data on questions that require reasoning across multiple temporal spans, especially in long-form egocentric videos; *Second*, existing architectures that treat temporal grounding as a language modeling task (Ren et al. 2024; Huang et al. 2024a,b), *e.g.*, directly incorporating the timestamp as the target of auto-regressive prediction, which is less effective than task-specific models (Lin et al. 2023; Mu, Mo, and Li 2024).

To bridge the gap, this paper introduces the problem of Multi-Hop Video Question-Answering *(MH-VidQA)*. As illustrated in Fig. 1, this task requires the model to simultaneously answer questions that involve visual information from multiple time intervals and localize these time spans as evidence within long, egocentric videos.

To acquire the data necessary for visual instruction tuning, we have developed an automated pipeline to construct la rge-scale question-answer-evidence triplets from the narrations of Ego4D (Grauman et al. 2022). Specifically, we build action scene graphs (Ji et al. 2020) by extracting syntax trees from narrations, allowing to analyze the temporal progression of actions, objects, and their relationships, thereby identifying potential questions that require information from multiple time points to answer. We then utilize large language models (LLMs) to generate eligible triplets across six different question types, that encompass real-world scenarios with an emphasis on the interactions between the individual and the external environment, as well as the long-term temporal relation of events. The categories include repeated activities, multiple actions, multiple objects, multiple locations/people, event composition, and event comparison.

Leveraging our automatically constructed data for visual instruction tuning, despite the existing VideoLLM (Ren et al. 2024) has demonstrated improved multi-hop reasoning abilities, it still struggles to localize the relevant time spans, primarily due to the limitations in predicting timestamps accurately. We further propose a novel architecture, termed as **G**rounding Scattered **E**vidence with Large **L**anguage **M**odel (**GeLM**). This architecture incorporates grounding tokens into the vocabulary of a multi-modal large language model, that are generated within the responses and then fused with visual features in a temporal grounding module to provide corresponding evidences, thereby enhancing the interpretability of the answers.

To track the development progress on *MH-VidQA* task, we have established a new benchmark, termed as **MULTIHOP-EGOQA**, that involves participants for validating and refining the generated triplets. Comprehensive evaluations show that both proprietary and open-source large multi-modal models largely fall behind human performance, highlighting the substantial challenge presented by **MULTIHOP-EGOQA**. Our architecture, trained on the automatically constructed instruction-tuning data, has shown significant improvement in multi-hop reasoning and grounding. We also evaluate our architecture on another public single-hop VidQA benchmark, ActivityNet-RTL (Huang et al. 2024b), outperforming existing approaches by a large margin.

Dataset	Annotation	Avg. Duration (s)	Ego?	Time Labels	Multi-Spans
Conventional VidQA Benchmarks					
MovieQA	Manual	211.4	✗	✓	✗
MSRVTT-QA	Auto	15	✗	✗	✗
MSVD-QA	Auto	10	✗	✗	✗
TVQA	Manual	76.2	✗	✓	✗
How2QA	Manual	60	✗	✓	✗
NeXT-QA	Manual	44	✗	✗	✗
iVQA	Manual	18.6	✗	✗	✗
EgoSchema	Auto + Manual	180	✓	✗	✗
Grounded VidQA Benchmarks					
QAEgo4D	Manual	498	✓	✓	✗
NExT-GQA	Manual	39.5	✗	✓	✓
ActivityNet-RTL	Auto + Manual	180	✗	✓	✗
MULTIHOP-EGOQA	Auto + Manual	180	✓	✓	✓

Table 1: Comparison of existing VidQA benchmarks.

2 Related Work

Video Question Answering Datasets. Video Question Answering (VidQA) is a video understanding task that involves answering natural language queries using visual-only or multi-modal information from videos. MovieQA (Tapaswi et al. 2016) proposes one of the earliest datasets in this area. However, most of its questions can be answered based on subtitles alone, with few relying on visual cues (Jasani, Girdhar, and Ramanan 2019). ActivityNet-QA (Yu et al. 2019) and How2QA (Sanabria et al. 2018) have focused on visual understanding in daily life and instructional videos. More recent datasets like NeXT-QA (Xiao et al. 2021), Perception Test (Patraucean et al. 2024), STAR (Wu et al. 2021), and AGQA (Grunde-McLaughlin, Krishna, and Agrawala 2021) focus on designing questions requiring spatio-temporal reasoning and causal relations. Additionally, EgoSchema (Mangalam, Akshulakov, and Malik 2023) proposes to generate questions through LLMs and manual efforts for long-form egocentric videos.

Multi-Hop QA with Grounding. In Natural Language Processing, multi-hop question-answering involves reasoning across multiple pieces of information, often requiring the retrieval of evidence from various sources (Yang et al. 2018; Ho et al. 2020; Xiong et al. 2021; Trivedi et al. 2022; Zhang et al. 2024a). In video understanding, conventional VidQA benchmarks do not necessitate models to explicitly localize or reason over temporally scattered evidence. However, recent works like EGOTIMEQA (Di and Xie 2024), NExT-GQA (Xiao et al. 2024), and REXTIME (Chen et al. 2024a) emphasize the importance of the grounding evidence in VidQA. These benchmarks, though, assume that evidence is confined to a single time span, overlooking the need for long-term temporal modelling and multi-step reasoning, which can be an oversimplification in video understanding.

Multi-modal Large Language Models. With the recent advancements in Large Language Models (LLMs) (Achiam et al. 2023; Chiang et al. 2023; AI@Meta 2024; Jiang et al. 2024), researchers are endeavouring to develop Multi-modal Large Language Models (MLLMs) by aligning visual and linguistic modalities through visual instruction tuning. For image understanding, several studies (Alayrac et al. 2022; Li et al. 2023a; Zhu et al. 2023; Liu et al. 2023a, 2024)

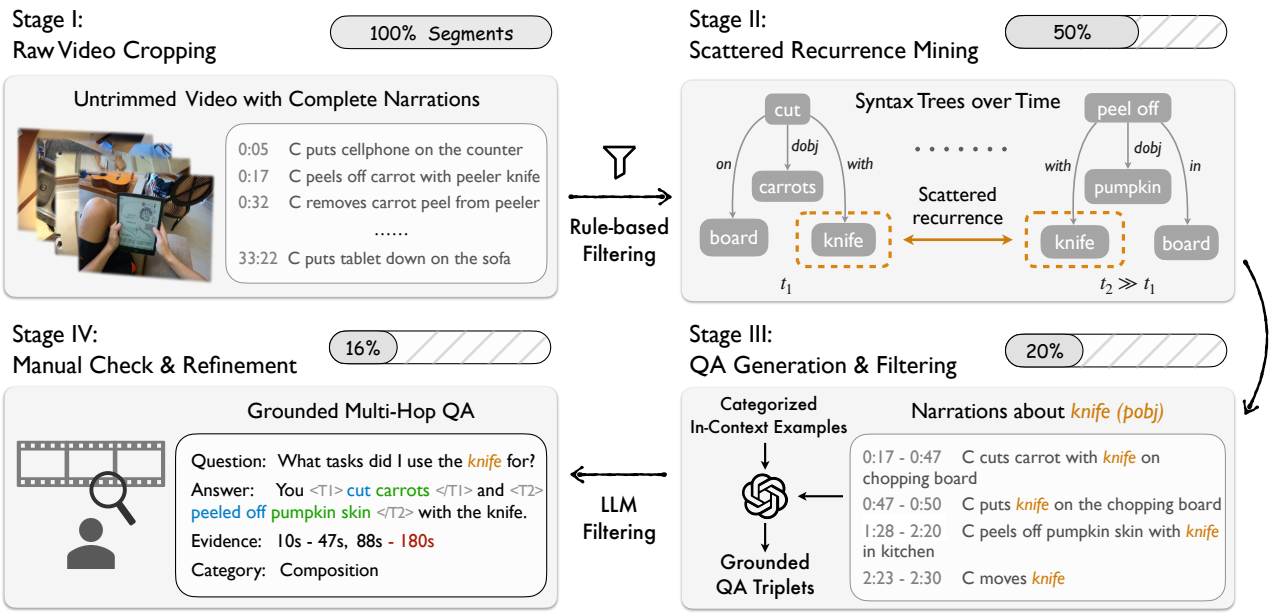

Figure 2: Illustration of our data curation pipeline. To collect large-scale multi-hop VidQA data, we have developed an automated pipeline. We begin by using action scene graphs to identify potential multi-hop reasoning questions based on the syntax trees of annotated narrations. Next, we use GPT-4o to generate data samples that include questions, answers, and relevant time spans. Finally, we perform manual validation and refinement to create the new benchmark.

have shown strong performance across various VQA benchmarks (Lu et al. 2023; Liu et al. 2023b; Yue et al. 2024). In video understanding, while some works (Li et al. 2023b; Ataallah et al. 2024; Zhang et al. 2024b) have made progress on traditional VidQA benchmarks, they are generally designed for short videos. Recent efforts to improve temporal awareness (Ren et al. 2024; Huang et al. 2024a; Qian et al. 2024) still lag behind task-specific models (Lin et al. 2023; Mu, Mo, and Li 2024) in the temporal grounding ability. To address this gap, our proposed dataset supports both instruction tuning and the evaluation of multi-hop reasoning and grounding in long-form egocentric videos, thereby advancing the development of video-language models.

3 Problem Formulation

Given a video stream and a question in the format of free-form text, *e.g.*, $\mathcal{V}$ and $\mathcal{Q}$ respectively, the objective is to generate the answer and localise the temporal evidence :

$$[\hat{\mathcal{A}}, \hat{\mathcal{T}}] = \Phi(\mathcal{V}, \mathcal{Q}), \tag{1}$$

where $\hat{\mathcal{T}} = \{[s_1, e_1], [s_2, e_2], \ldots, [s_n, e_n]\}$ refers to a set of non-overlapping start-end time intervals, in which the video content is necessary for deriving the answer $\hat{\mathcal{A}}$.

To develop the vision systems that address our considered *MH-VidQA* task, it is essential to collect data in triplet form, *i.e.*, $(\mathcal{Q}, \mathcal{A}, \mathcal{T})$, to train the architecture that can simultaneously answer questions, and ground them across multiple time spans. In the following sections, we will detail an automated pipeline for constructing visual instructions and training our proposed architecture.

4 MultiHop-EgoQA: Curation Pipeline

The curation pipeline involves four stages: (i) cropping and selecting video clips from untrimmed Ego4D dataset; (ii) mining potential questions that demand multi-hop reasoning based on narrations; (iii) producing $(\mathcal{Q}, \mathcal{A}, \mathcal{T})$ triplets across various categories using LLMs; (iv) filtering the generated samples with LLMs, then followed by manual review and refinement.

4.1 Raw Video Cropping & Selection

We start with the 9,611 untrimmed egocentric videos (24-minute duration on average), accompanied with a total of 3.85M timestamped narrations from Ego4D (Grauman et al. 2022). Since these timestamps indicate the occurrence of a new action (*i.e.*, the start time), to estimate the duration of each action, we take the timestamp of the subsequent narration as the end time. Specifically, we segment the raw videos into non-overlapping 3-minute clips. Each clip and the corresponding narrations are denoted as $\mathcal{V}$ and $\mathcal{N} = \{N_i\}_{i=1}^{|\mathcal{N}|}$.

4.2 Mining Multi-Hop QA from Narrations

We propose to mine the multi-hop VidQA triplets from action scene graphs for each long-form video, which provide temporally evolving object descriptions, human-object relationships, and the progression of actions over time. (Ji et al. 2020; Yang et al. 2023; Rodin et al. 2024).

To build action scene graphs, we use the syntax tree of each narration to identify the specific nodes, involving actions, objects, locations and people. We then search for structures where a single node recurs over time, but connects

with different neighbouring nodes across various scenes, since these structures are likely to contain the multi-hop reasoning queries. The detailed procedure is outlined below.

Narration Syntax Tree → Action Scene Graph. We use `spaCy` (Honnibal et al. 2020) to parse each narration and extract the basic nodes, including actions *(verb)*, direct objects *(dobj)*, and prepositional objects *(pobj)* along with their modifiers. For instance, "C puts the cooking pot on the counter top" can be parsed into {put, cooking pot, counter top} as three nodes with distinct syntactic attributes.

Searching Scattered Recurrence in Graph. We then focus on a node u that recurs sporadically throughout the entire set of narrations $\mathcal{N}$. The minimum and maximum recurrence times for the selected node u are denoted as t_{min} and t_{max}, respectively, which determines the number of time intervals involved in the question. We extract the narrations related to the specific node u from $\mathcal{N}$, denoted as $\mathcal{N}_u$. These narrations, though focused on node u, describe different action scenes, making them potential candidates for multi-hop triplet generation in Stage III.

4.3 QA Generation & Filtering with LLM

Based on the nodes' syntactic attributes, we use different in-context learning examples to guide the LLM-based QA generation processes. The resulting multi-hop questions are divided into six categories, involving repeated activities, multiple actions, multiple objects, multiple locations/people, event composition, and event comparison. The detailed prompts and question examples are presented in the Supplementary Material. Formally, for the selected narrations $\mathcal{N}_u$, the triplet is generated with prompt ($\mathcal{P}$), denoted as:

$$\{(\mathcal{Q}, \mathcal{A}, \mathcal{T})\} = \text{LLM}(\mathcal{P}; \mathcal{N}_u) \tag{2}$$

After the automated generation, we use an LLM to filter out unreasonable QA pairs, resulting in 4,412 clips with 14,397 triplets. We utilize `GPT-4o` for both generation and filtration processes due to its superior capabilities.

4.4 Manual Check & Refinement

To construct a benchmark, we select 380 clips with 1,208 triplets and hire 12 graduate students majoring in computer vision, to validate the clarity of the data and further refine the temporal annotations. As a result, we obtain 360 clips with 1,080 triplets, which form the final benchmark, termed as **MULTIHOP-EGOQA**. The annotation details and benchmark statistics are provided in the Supplementary Material.

5 GeLM: A Baseline Method for *MH-VidQA*

Existing models for video question answering typically provide answers without supporting temporal evidence, or are restricted to identifying a single time interval. Here, we propose a novel architecture, termed as **GeLM**: **G**rounding Scattered **E**vidence with Large **L**anguage **M**odel for Multi-Hop Video Question-Answering. As depicted in Fig. 3, our model primarily comprises a multi-modal large language model and a grounding module, with special grounding tokens (`<T> </T>`) indicating the time span of the enclosed key information in the response.

5.1 Visual-language Encoding Module

Given the video clip with L frames and the associated question, we first adopt a frozen visual encoder to extract the visual features. The given question is tokenized and transformed into textual embeddings, while the visual features are projected into visual embeddings with the same dimension through a linear projector:

$$\mathbf{x}_v = \phi_{\text{proj}}(\Phi_{\text{v-enc}}(\mathcal{V})), \quad \mathbf{x}_q = \phi_{\text{emb}}(\mathcal{Q}) \tag{3}$$

where $\mathbf{x}_v \in \mathbb{R}^{L\times D}$, $\mathbf{x}_q \in \mathbb{R}^{Q\times D}$ denote the computed embeddings for visual and question respectively.

5.2 Multi-modal Large Language Model

The visual and textual embeddings are then fed into a multi-modal large language model:

$$\{\mathbf{h}_v, \mathbf{h}_q, \mathbf{h}_a\} = \text{MLLM}([\mathbf{x}_v : \mathbf{x}_q]) \tag{4}$$

where $\mathbf{h}_v, \mathbf{h}_q, \mathbf{h}_a$ represent the hidden states of the input frames, question, and the output response respectively. The answer texts $\hat{\mathcal{A}}$ are then decoded using a linear head on $\mathbf{h}_a$.

Grounding Tokens. Inspired by approaches that enable MLLMs to segment visual entities (Lai et al. 2024; Zhang et al. 2024c; Yan et al. 2024), we expand the vocabulary by adding grounding token pairs, *i.e.*, `<T> </T>`, which indicate the start-end time span. As illustrated in Fig. 3, when the MLLM needs to ground the temporal evidence for its response, the relevant part of the response is enclosed by `<T>` and `</T>`. We concatenate the last-layer hidden states of each pair, *i.e.*, `<T1>` and `</T1>`, ..., `<TK>` and `</TK>` along the channel dimension to form a single grounding query vector, resulting in K grounding queries $\mathbf{H}_g \in \mathbb{R}^{K\times 2D}$. Note that, the value of K can vary for different responses. These queries are then processed through the grounding module, which interacts with the visual hidden states $\mathbf{h}_v \in \mathbb{R}^{L\times D}$.

5.3 Evidence Grounding Module

To ground the time spans that support the answer, we design an evidence grounding module that processes a variable number of grounding queries and predicts the corresponding temporal proposals in the video: $\hat{\mathcal{T}} = \{[s_i, e_i]\}_{i=1}^{|\hat{\mathcal{T}}|}$.

We begin by projecting the hidden states of frames $\mathbf{h}_v$ and grounding queries $\mathbf{H}_g$ into the same dimension C:

$$\mathbf{w}_v = \phi_v(\mathbf{h}_v) \in \mathbb{R}^{L\times C}, \quad \mathbf{w}_g = \phi_g(\mathbf{H}_g) \in \mathbb{R}^{K\times C} \tag{5}$$

Following this, two separate branches are used to predict the temporal evidence for the answer: the saliency branch and the similarity branch, as depicted in Fig. 3. The saliency branch utilizes a self-attention mechanism across all visual features and grounding tokens to identify all temporal evidence for the question holistically. The similarity branch calculates the visual-textual similarity between each grounding query and all visual features, to determine the time spans for each part of the response in a fragmented manner.

Saliency branch. As illustrated in Fig. 3, the saliency branch utilizes three Transformer Encoder layers as the temporal aggregator, to fuse information between grounding queries and visual hidden states:

$$\{\mathbf{o}_v, \mathbf{o}_g\} = \Phi_{\text{temp-agg}}([\mathbf{w}_v : \mathbf{w}_g]) \tag{6}$$

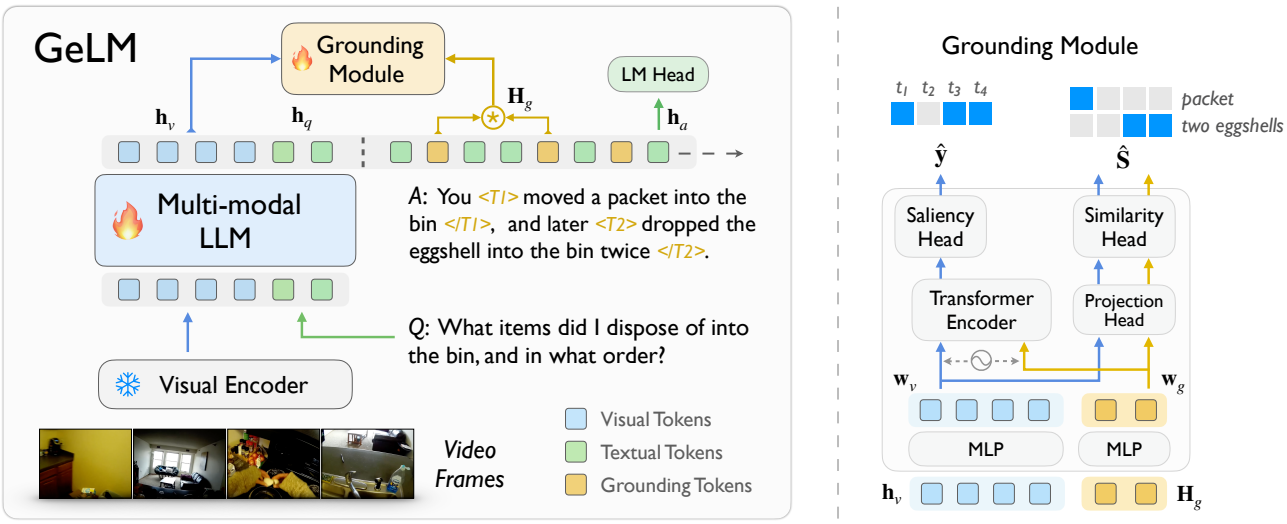

Figure 3: Overview of the proposed architecture. **GeLM** can generate grounding token pairs, *i.e.*, <T> </T>, in the response of a multi-modal large language model, which denote the start and end times of the enclosed statement. These grounding tokens are then processed with visual hidden states to the ground multiple time spans that provide evidence supporting the answer.

The predicted saliency score $\hat{\mathbf{y}} \in \mathbb{R}^L$ is derived through the saliency head $\phi_{\text{saliency}}(\cdot)$ and the sigmoid function $\sigma(\cdot)$:

$$\hat{\mathbf{y}} = \sigma(\phi_{\text{saliency}}(\mathbf{o}_v)) \tag{7}$$

where a higher score indicates a higher probability that each frame serves as visual evidence for the question-answering. The saliency head $\phi_{\text{saliency}}(\cdot)$ consists of two Conv1D layers with ReLU activation.

Similarity branch. Apart from predicting the saliency, we also aim to determine the time spans of key information enclosed by each grounding token pair separately, represented as a similarity matrix $\hat{\mathbf{S}} \in \mathbb{R}^{K \times L}$ between K grounding queries and L frames. Specifically, we project the hidden states with a linear layer,

$$\mathbf{z}^v = \psi_{\text{v-proj}}(\mathbf{w}_v), \quad \mathbf{z}^g = \psi_{\text{g-proj}}(\mathbf{w}_g) \tag{8}$$

and compute the cosine similarity matrix:

$$\hat{\mathbf{S}}_{ij} = \frac{\mathbf{z}_i^g \cdot \mathbf{z}_j^v}{\|\mathbf{z}^g\| \cdot \|\mathbf{z}^v\|} \in [-1, 1] \tag{9}$$

where a higher value of value of $\hat{\mathbf{S}}_{ij}$ indicates that the j-th frame is more relevant to the i-th grounding query.

Proposal generation strategy. During inference, to generate temporal proposals ($\hat{\mathcal{T}}$), we apply the following post-processing. Utilizing the saliency score vector $\hat{\mathbf{y}} \in \mathbb{R}^L$, we set a threshold at 70% of the maximum saliency score. Timestamps with scores above this threshold are merging into time spans.

Leveraging the similarity matrix $\hat{\mathbf{S}} \in \mathbb{R}^{K \times L}$, we apply an average pooling kernel with a size of 3 and a stride of 1 to smooth the values. Then we perform a softmax function along each row to get positive scores. Since each row vector $\hat{\mathbf{S}}_{k,:} \in \mathbb{R}^L$ in the matrix represents the predicted temporal relevance for the k-th grounding query, we apply the same thresholding method to each row and take the union of the results to obtain a set of proposals.

Training objective. For question answering, the cross entropy loss $\mathcal{L}_{\text{CE}}(\hat{\mathcal{A}}, \mathcal{A})$ is utilized for next token prediction. For evidence grounding, given the ground truth binary saliency labels $\mathbf{y} \in \{0,1\}^L$, we use binary cross entropy as loss function: $\mathcal{L}_{\text{BCE}} = \frac{1}{L}\sum_{i=1}^{L} -\mathbf{y}_i \log \hat{\mathbf{y}}_i$. With the ground truth binary similarity matrix $\mathbf{S} \in \{0,1\}^{K \times L}$, we adopt the Multiple Instance Learning NCE (MIL-NCE) loss (Miech et al. 2019) for contrastive learning:

$$\mathcal{L}_{\text{NCE}} = -\frac{1}{K}\sum_{i=1}^{K} \log \frac{\sum_{j=1}^{L} \mathbf{S}_{ij} \exp(\hat{\mathbf{S}}_{ij}/\tau)}{\sum_{j=1}^{L} \exp(\hat{\mathbf{S}}_{ij}/\tau)} \tag{10}$$

where τ denotes temperature. i, j correspond to i-th grounding query and j-th frame, respectively. The final loss is a weighted sum of the above losses: $\mathcal{L} = \mathcal{L}_{\text{CE}} + \lambda_{\text{BCE}}\mathcal{L}_{\text{BCE}} + \lambda_{\text{NCE}}\mathcal{L}_{\text{NCE}}$.

6 Experiments

In this section, we first describe the metrics for our benchmark, **MULTIHOP-EGOQA**, and then evaluate the performance of existing approaches. Next, we employ instruction tuning with the automatically constructed dataset, to establish a strong baseline for the multi-hop VidQA task. Lastly, we show that our method also achieves state-of-the-art performance on the existing public single-hop VidQA task.

6.1 Evaluation Metrics

We evaluate the performance of question answering and evidence grounding separately on **MULTIHOP-EGOQA**.

Question answering. To evaluate open-ended answers, we use GPT-4o as the primary evaluator for scoring, as it more

Methods	Input	Temporal Grounding				Question Answering	
		mIoP	mIoG	**IoU@0.3**	**mIoU**	Sent. Sim.	**Score** (10 ↑)
Human	-	71.8	81.0	87.0	61.8	74.3	7.5
GPT-4o (OpenAI 2024)	60 frames	18.9	24.4	**12.0**	**12.2**	**73.7**	**5.4**
End-to-End MLLMs							
InternVL2-8B (Chen et al. 2024b)	30 frames	11.8	24.0	6.3	6.6	71.9	4.5
LLaVA-NeXT-Video-7B (Zhang et al. 2024b)	32 frames	-	-	-	-	62.1	4.2
TimeChat-7B (Ren et al. 2024)	96 frames	10.2	5.6	3.0	3.6	58.9	3.3
VTimeLLM-7B (Huang et al. 2024a)	100 frames	12.4	**28.2**	8.8	9.2	70.5	4.3
Pipeline: Caption Module → LLM (QA + Grounding)							
LLaVa-NeXT-7B → Llama-3.1-8B	180 frames	**21.4**	22.3	10.1	9.7	63.6	3.5

Table 2: Zero-shot performance of existing multi-modal models on MULTIHOP-EGOQA. Existing approaches of various types fall short of human performance on this challenging task.

closely aligns with human judgment and is widely adopted for assessment purposes (Chiang and Lee 2023; Zheng et al. 2024). We also report the average Sentence Similarity (Sent. Sim.) between the ground truth answers and the predicted answers (Reimers and Gurevych 2019). For time-related questions, we exclude them from metrics of answering, as they can be accurately evaluated with localization metrics.

Evidence localization. Given the m predicted non-overlap time spans: $\hat{\mathcal{T}} = \{\hat{T}_1, \hat{T}_2, \ldots, \hat{T}_m\}$ and the ground truth consisting of n spans: $\mathcal{T} = \{T_1, T_2, \ldots, T_n\}$ for each video, IoU (Intersection over Union) is computed as follows:

$$\text{IoU}(\mathcal{T}, \hat{\mathcal{T}}) = \frac{\sum_{i=1}^{m} \sum_{j=1}^{n} |\hat{T}_i \cap T_j|}{\left| \bigcup_{i=1}^{m} \hat{T}_i \cup \bigcup_{j=1}^{n} T_j \right|} \tag{11}$$

This can be seen as an extension of the IoU between two intervals, measuring the Jaccard Distance between two sets of time spans. Subsequently, the mean IoU (mIoU) is determined by averaging the IoU values across the entire test set. Additionally, we calculate the proportion of videos with an IoU exceeding 0.3, designated as IoU@0.3. Similar to precision and recall, we compute mIoP and mIoG by averaging Intersection over Prediction (IoP) and Intersection over Ground Truth (IoG), replacing the denominator in Eq. (11) with $\left|\bigcup_{i=1}^{m} \hat{T}_i\right|$ and $\left|\bigcup_{j=1}^{n} T_j\right|$, respectively.

6.2 Evaluation on MULTIHOP-EGOQA

In this section, we evaluate several latest multi-modal models on **MULTIHOP-EGOQA**, exploring their abilities of multi-hop reasoning and temporal grounding.

Human and advanced proprietary model. Initially, we invite participants (different from annotators in the curation pipeline) to assess human performance on this task. We randomly sample 10% of the test split and request participants to answer the questions and localise relevant time spans. Additionally, we evaluate the advanced proprietary model, GPT-4o, by leveraging its visual capabilities with uniformly sampled frames from the video clip, to perform answering and grounding.

End-to-end models. We conduct investigations across various popular MLLMs, including Image LLM (InternVL2-8B), Short Video LLM (LLaVA-NeXT-Video-7B), and Long Video LLMs (TimeChat-7B, VTimeLLM-7B).

Multi-stage pipeline. To explore the effectiveness of dense captioning for *MH-VidQA*, we adopt a multi-stage pipeline, consisting of an image caption module, followed by an LLM. The captions of sampled frames with timestamps will be utilized by the LLM for answering and grounding.

Overall Results. From experiments presented in Tab. 2, we can draw the following observations: *1) Both the proprietary model and open-source multi-modal LLMs significantly lag behind human performance*, underscoring the current limitations in multi-hop reasoning and grounding capabilities within multi-modal systems. *2) Reasoning and grounding abilities are disentangled in existing visual systems.* For instance, LLaVA-NeXT-Video is unable to handle requests involving temporal grounding, but can still answer part of questions that do not involve temporal grounding. *3) Instruction-tuning with single-hop data does not guarantee superiority in multi-hop grounding.* For example, despite TimeChat and VTimeLLM have been fine-tuned with temporally aware instructions and multi-turn conversations, the ability to ground multiple intervals for a single query remains limited. *4) Dense captions do indeed help temporal grounding, but errors may cascade.* Although captioning at per second provides explicit temporal information for grounding, errors in the captioning process are difficult to correct through the subsequent stages. **The evaluation details are presented in the Supplementary Material.**

6.3 A Baseline Method for MULTIHOP-EGOQA

In the following section, we propose a new baseline for this challenging task and conduct ablation experiments to evaluate the effectiveness of the automatically constructed training data, and our architectural design for future research.

Implementation Details

Training data. We utilize the triplets generated in our automated pipeline to train the multi-modal LLM and the

Method	Training Data	Temporal Grounding		QA
		IoU@0.3	mIoU	Score ↑
TimeChat	✗	3.0	3.6	3.3
	100%	8.6	8.1	4.4
VTimeLLM	✗	8.8	9.2	4.3
	100%	13.1	12.8	4.6
GeLM (Ours)	25%	13.0	11.3	4.6
	50%	16.7	16.1	4.7
	100%	**18.2**	**16.7**	**4.8**

Table 3: Effect of the instruction-tuning data. We establish a new baseline for the new task after instruction-tuning.

grounding module. These triplets have been filtered by the LLM, but not manually refined in Stage IV, consisting of 3,156 clips with a total of 10,414 samples.

Architecture. The visual features of MULTIHOP-EGOQA are extracted with the InternVideo-MM-L-14 (Wang et al. 2022) from 8 frames per second. The large language model employed is Vicuna-7B v1.3 (Chiang et al. 2023). The dimensions of the hidden states for the LLM and the grounding module are 4096 and 1024, respectively.

Training setup. The experiments are conducted using 4 NVIDIA H800 (80GB) GPUs, with a batch size of 32 per device. The model is trained for 10 epochs with a learning rate of 2×10^{-5}, employing a warmup cosine decay strategy.

Ablation Studies

Effect of the visual instruction-tuning data. To validate the effectiveness of our data curation pipeline for mining large-scale multi-hop VidQA data, we utilize the automatically collected instructions to fine-tune pre-trained Video LLMs (*i.e.*, TimeChat and VTimeLLM), and evaluate on MULTIHOP-EGOQA. As Tab. 3 shows, visual instruction tuning enhances the multi-hop reasoning and grounding abilities of existing models, demonstrating the effectiveness of constructed data. Additionally, we also trained our model by varying percentages of data, the results demonstrates the potential of our automated pipeline to collect scalable data and the effectiveness of our proposed architecture, **GeLM**.

Effect of training objective and inference strategy. We explore the role of each training loss and the effect of the two branches for generating temporal proposals. As shown in Tab. 4, although the saliency branch is unable to distinguish the time interval of each grounding token pair, the binary cross entropy loss tends to benefit the temporal grounding, improving the performance of the similarity branch, with IoU@0.3 increasing from 14.1 to 18.2, and mIoU from 13.4 to 16.7. Correspondingly, the similarity branch also enhances the inference results of the saliency branch, demonstrating the complementarity of both branches. We calculate the Pearson correlation between the QA score and Grounding IoU, obtaining a correlation of 0.65, which highlights a synergy between precise grounding and accurate answering.

Training Loss	Strategy	Temporal Grounding		QA
		IoU@0.3	mIoU	Score ↑
$\mathcal{L}_{\text{CE}}$	-	-	-	4.7
$+\mathcal{L}_{\text{BCE}}$	Saliency	13.8	14.2	4.7
$+\mathcal{L}_{\text{NCE}}$	Similarity	14.1	13.4	4.6
$+\mathcal{L}_{\text{NCE}}+\mathcal{L}_{\text{BCE}}$	Saliency	**19.2**	14.7	4.7
	Similarity	18.2	**16.7**	**4.8**

Table 4: Ablation of the training objective and inference strategy. The gray shading indicates the default setting.

6.4 On Existing Single-Hop VidQA Benchmark

Dataset and metrics. In addition to our multi-hop benchmark, we validate the effectiveness of our method on the public single-hop VidQA benchmark (Huang et al. 2024b), which contains 229 question-answer pairs across 160 videos. For this benchmark, the temporal grounding metrics are mIoU and Precision@0.5 (P@0.5), with the latter measuring the percentage of predictions with an IoU over 0.5. Additionally, the GPT-4 Relative Score (R. Score) is computed for evaluating the predicted explanations.

Comparison. In the existing state-of-the-art method, for example, LITA (Huang et al. 2024b) adds special time tokens into the vocabulary to process temporal grounding as a next-token prediction task on this benchmark. As shown in Tab. 5, our architecture significantly exceeds LITA with both temporal grounding branches after fine-tuning.

Model	Strategy	Temporal Grounding		QA
		mIoU	P@0.5	R. Score ↑
LITA-7B	Time Token	24.1	21.2	44.0
LITA-13B	Time Token	28.6	25.9	46.3
GeLM-7B	Saliency	31.8	28.2	**45.3**
GeLM-7B	Similarity	35.4 ↑11.3	**31.0** ↑9.8	45.1 ↑1.1

Table 5: Comparison with the state-of-the-art method on ActivityNet-RTL, a public single-hop VidQA benchmark.

7 Conclusion

To conclude, we have initiated the *MH-VidQA* task for long-form egocentric video understanding. To acquire the associated dataset, we have devised an automated pipeline to mine large-scale multi-hop QA triplets, a subset of which are subsequently validated and refined manually, resulting in a new benchmark. Existing multi-modal systems demonstrate improvement in multi-hop reasoning abilities after training on the automatically collected data, but they still struggle to ground relevant temporal spans. To bridge this gap, we have proposed a novel model capable of answering multi-hop questions and concurrently grounding scattered visual clues, which establishes a baseline for this challenging task after visual instruction tuning. Our method also achieves state-of-the-art performance on the public single-hop VidQA benchmark, further underscoring its effectiveness.

Acknowledgments

This work is supported by National Key R&D Program of China (No. 2022ZD0161400).

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. GPT-4 technical report. arXiv:2303.08774.

AI@Meta. 2024. Llama 3 Model Card.

Alayrac, J.-B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M.; et al. 2022. Flamingo: a visual language model for few-shot learning. In *NeurIPS*.

Ataallah, K.; Shen, X.; Abdelrahman, E.; Sleiman, E.; Zhu, D.; Ding, J.; and Elhoseiny, M. 2024. MiniGPT4-Video: Advancing Multimodal LLMs for Video Understanding with Interleaved Visual-Textual Tokens. arXiv:2404.03413.

Bärmann, L.; and Waibel, A. 2022. Where did i leave my keys?-episodic-memory-based question answering on egocentric videos. In *CVPRW*.

Chen, J.-J.; Liao, Y.-C.; Lin, H.-C.; Yu, Y.-C.; Chen, Y.-C.; and Wang, Y.-C. F. 2024a. ReXTime: A Benchmark Suite for Reasoning-Across-Time in Videos. arXiv:2406.19392.

Chen, Z.; Wang, W.; Tian, H.; Ye, S.; Gao, Z.; Cui, E.; Tong, W.; Hu, K.; Luo, J.; Ma, Z.; et al. 2024b. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. arXiv:2404.16821.

Chiang, C.-H.; and Lee, H.-y. 2023. Can Large Language Models Be an Alternative to Human Evaluations? In *ACL*.

Chiang, W.-L.; Li, Z.; Lin, Z.; Sheng, Y.; Wu, Z.; Zhang, H.; Zheng, L.; Zhuang, S.; Zhuang, Y.; Gonzalez, J. E.; Stoica, I.; and Xing, E. P. 2023. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality.

Datta, S.; Dharur, S.; Cartillier, V.; Desai, R.; Khanna, M.; Batra, D.; and Parikh, D. 2022. Episodic memory question answering. In *CVPR*.

Di, S.; and Xie, W. 2024. Grounded Question-Answering in Long Egocentric Videos. In *CVPR*.

Grauman, K.; Westbury, A.; Byrne, E.; Chavis, Z.; Furnari, A.; Girdhar, R.; Hamburger, J.; Jiang, H.; Liu, M.; Liu, X.; Martin, M.; Nagarajan, T.; Radosavovic, I.; Ramakrishnan, S. K.; Ryan, F.; Sharma, J.; Wray, M.; Xu, M.; Xu, E. Z.; Zhao, C.; Bansal, S.; Batra, D.; Cartillier, V.; Crane, S.; Do, T.; Doulaty, M.; Erapalli, A.; Feichtenhofer, C.; Fragomeni, A.; Fu, Q.; Gebreselasie, A.; Gonzalez, C.; Hillis, J.; Huang, X.; Huang, Y.; Jia, W.; Khoo, W.; Kolar, J.; Kottur, S.; Kumar, A.; Landini, F.; Li, C.; Li, Y.; Li, Z.; Mangalam, K.; Modhugu, R.; Munro, J.; Murrell, T.; Nishiyasu, T.; Price, W.; Puentes, P. R.; Ramazanova, M.; Sari, L.; Somasundaram, K.; Southerland, A.; Sugano, Y.; Tao, R.; Vo, M.; Wang, Y.; Wu, X.; Yagi, T.; Zhao, Z.; Zhu, Y.; Arbelaez, P.; Crandall, D.; Damen, D.; Farinella, G. M.; Fuegen, C.; Ghanem, B.; Ithapu, V. K.; Jawahar, C. V.; Joo, H.; Kitani, K.; Li, H.; Newcombe, R.; Oliva, A.; Park, H. S.; Rehg, J. M.; Sato, Y.; Shi, J.; Shou, M. Z.; Torralba, A.; Torresani, L.; Yan, M.; and Malik, J. 2022. Ego4D: Around the World in 3,000 Hours of Egocentric Video. In *CVPR*.

Grunde-McLaughlin, M.; Krishna, R.; and Agrawala, M. 2021. AGQA: A Benchmark for Compositional Spatio-Temporal Reasoning. In *CVPR*.

Ho, X.; Nguyen, A.-K. D.; Sugawara, S.; and Aizawa, A. 2020. Constructing A Multi-hop QA Dataset for Comprehensive Evaluation of Reasoning Steps. In *ACL*.

Honnibal, M.; Montani, I.; Van Landeghem, S.; Boyd, A.; et al. 2020. spaCy: Industrial-strength natural language processing in python.

Huang, B.; Wang, X.; Chen, H.; Song, Z.; and Zhu, W. 2024a. Vtimellm: Empower llm to grasp video moments. In *CVPR*.

Huang, D.-A.; Liao, S.; Radhakrishnan, S.; Yin, H.; Molchanov, P.; Yu, Z.; and Kautz, J. 2024b. LITA: Language Instructed Temporal-Localization Assistant. In *ECCV*.

Jasani, B.; Girdhar, R.; and Ramanan, D. 2019. Are we asking the right questions in MovieQA? In *ICCVW*.

Ji, J.; Krishna, R.; Fei-Fei, L.; and Niebles, J. C. 2020. Action genome: Actions as compositions of spatio-temporal scene graphs. In *CVPR*.

Jiang, A. Q.; Sablayrolles, A.; Roux, A.; Mensch, A.; Savary, B.; Bamford, C.; Chaplot, D. S.; Casas, D. d. l.; Hanna, E. B.; Bressand, F.; et al. 2024. Mixtral of experts. arXiv:2401.04088.

Lai, X.; Tian, Z.; Chen, Y.; Li, Y.; Yuan, Y.; Liu, S.; and Jia, J. 2024. Lisa: Reasoning segmentation via large language model. In *CVPR*.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*.

Li, K.; Wang, Y.; He, Y.; Li, Y.; Wang, Y.; Liu, Y.; Wang, Z.; Xu, J.; Chen, G.; Luo, P.; et al. 2023b. Mvbench: A comprehensive multi-modal video understanding benchmark. arXiv:2311.17005.

Lin, K. Q.; Zhang, P.; Chen, J.; Pramanick, S.; Gao, D.; Wang, A. J.; Yan, R.; and Shou, M. Z. 2023. Univtg: Towards unified video-language temporal grounding. In *ICCV*.

Liu, H.; Li, C.; Li, Y.; Li, B.; Zhang, Y.; Shen, S.; and Lee, Y. J. 2024. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023a. Visual Instruction Tuning. In *NeurIPS*.

Liu, Y.; Duan, H.; Zhang, Y.; Li, B.; Zhang, S.; Zhao, W.; Yuan, Y.; Wang, J.; He, C.; Liu, Z.; et al. 2023b. Mmbench: Is your multi-modal model an all-around player? arXiv:2307.06281.

Lu, P.; Bansal, H.; Xia, T.; Liu, J.; Li, C.; Hajishirzi, H.; Cheng, H.; Chang, K.-W.; Galley, M.; and Gao, J. 2023. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. arXiv:2310.02255.

Majumdar, A.; Ajay, A.; Zhang, X.; Putta, P.; Yenamandra, S.; Henaff, M.; Silwal, S.; Mcvay, P.; Maksymets, O.; Arnaud, S.; et al. 2024. Openeqa: Embodied question answering in the era of foundation models. In *CVPR*.

Mangalam, K.; Akshulakov, R.; and Malik, J. 2023. EgoSchema: A Diagnostic Benchmark for Very Long-form Video Language Understanding. In *NeurIPS Datasets and Benchmarks Track*.

Miech, A.; Zhukov, D.; Alayrac, J.-B.; Tapaswi, M.; Laptev, I.; and Sivic, J. 2019. HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips. In *ICCV*.

Mu, F.; Mo, S.; and Li, Y. 2024. SnAG: Scalable and Accurate Video Grounding. In *CVPR*.

OpenAI. 2024. GPT-4o. https://openai.com/index/hello-gpt-4o/.

Patraucean, V.; Smaira, L.; Gupta, A.; Recasens, A.; Markeeva, L.; Banarse, D.; Koppula, S.; Malinowski, M.; Yang, Y.; Doersch, C.; et al. 2024. Perception test: A diagnostic benchmark for multimodal video models. In *NeurIPS*.

Qian, L.; Li, J.; Wu, Y.; Ye, Y.; Fei, H.; Chua, T.-S.; Zhuang, Y.; and Tang, S. 2024. Momentor: Advancing video large language model with fine-grained temporal reasoning. In *ICML*.

Ramakrishnan, S. K.; Al-Halah, Z.; and Grauman, K. 2023. Naq: Leveraging narrations as queries to supervise episodic memory. In *CVPR*.

Reimers, N.; and Gurevych, I. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *EMNLP*.

Ren, S.; Yao, L.; Li, S.; Sun, X.; and Hou, L. 2024. Timechat: A time-sensitive multimodal large language model for long video understanding. In *CVPR*.

Rodin, I.; Furnari, A.; Min, K.; Tripathi, S.; and Farinella, G. M. 2024. Action Scene Graphs for Long-Form Understanding of Egocentric Videos. In *CVPR*.

Sanabria, R.; Caglayan, O.; Palaskar, S.; Elliott, D.; Barrault, L.; Specia, L.; and Metze, F. 2018. How2: A Large-scale Dataset for Multimodal Language Understanding. In *NeurIPS*.

Tapaswi, M.; Zhu, Y.; Stiefelhagen, R.; Torralba, A.; Urtasun, R.; and Fidler, S. 2016. MovieQA: Understanding Stories in Movies through Question-Answering. In *CVPR*.

Trivedi, H.; Balasubramanian, N.; Khot, T.; and Sabharwal, A. 2022. MuSiQue: Multihop Questions via Single-hop Question Composition. In *TACL*.

Wang, Y.; Li, K.; Li, Y.; He, Y.; Huang, B.; Zhao, Z.; Zhang, H.; Xu, J.; Liu, Y.; Wang, Z.; et al. 2022. Internvideo: General video foundation models via generative and discriminative learning. arXiv:2212.03191.

Wu, B.; Yu, S.; Chen, Z.; Tenenbaum, J. B.; and Gan, C. 2021. STAR: A Benchmark for Situated Reasoning in Real-World Videos. In *NeurIPS Datasets and Benchmarks Track*.

Xiao, J.; Shang, X.; Yao, A.; and Chua, T.-S. 2021. NExT-QA:Next Phase of Question-Answering to Explaining Temporal Actions. In *CVPR*.

Xiao, J.; Yao, A.; Li, Y.; and Chua, T.-S. 2024. Can i trust your answer? visually grounded video question answering. In *CVPR*.

Xiong, W.; Li, X.; Iyer, S.; Du, J.; Lewis, P.; Wang, W. Y.; Mehdad, Y.; Yih, S.; Riedel, S.; Kiela, D.; et al. 2021. Answering Complex Open-Domain Questions with Multi-Hop Dense Retrieval. In *ICLR*.

Yan, C.; Wang, H.; Yan, S.; Jiang, X.; Hu, Y.; Kang, G.; Xie, W.; and Gavves, E. 2024. VISA: Reasoning Video Object Segmentation via Large Language Models. In *ECCV*.

Yang, J.; Peng, W.; Li, X.; Guo, Z.; Chen, L.; Li, B.; Ma, Z.; Zhou, K.; Zhang, W.; Loy, C. C.; and Liu, Z. 2023. Panoptic Video Scene Graph Generation. In *CVPR*.

Yang, Z.; Qi, P.; Zhang, S.; Bengio, Y.; Cohen, W. W.; Salakhutdinov, R.; and Manning, C. D. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. arXiv:1809.09600.

Yu, Z.; Xu, D.; Yu, J.; Yu, T.; Zhao, Z.; Zhuang, Y.; and Tao, D. 2019. ActivityNet-QA: A Dataset for Understanding Complex Web Videos via Question Answering. In *AAAI*.

Yue, X.; Ni, Y.; Zhang, K.; Zheng, T.; Liu, R.; Zhang, G.; Stevens, S.; Jiang, D.; Ren, W.; Sun, Y.; et al. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *CVPR*.

Zhang, J.; Zhang, H.; Zhang, D.; Yong, L.; and Huang, S. 2024a. End-to-End Beam Retrieval for Multi-Hop Question Answering. In *NAACL*.

Zhang, Y.; Li, B.; Liu, h.; Lee, Y. j.; Gui, L.; Fu, D.; Feng, J.; Liu, Z.; and Li, C. 2024b. LLaVA-NeXT: A Strong Zero-shot Video Understanding Model.

Zhang, Y.; Ma, Z.; Gao, X.; Shakiah, S.; Gao, Q.; and Chai, J. 2024c. Groundhog: Grounding large language models to holistic segmentation. In *CVPR*.

Zheng, L.; Chiang, W.-L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E.; et al. 2024. Judging llm-as-a-judge with mt-bench and chatbot arena. In *NeurIPS*.

Zhu, D.; Chen, J.; Shen, X.; Li, X.; and Elhoseiny, M. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. arXiv:2304.10592.