Preprint – IJCAI 2025: This is the accepted version made available for conference attendees.
Do not cite. The final version will appear in the IJCAI 2025 proceedings.

Can Retelling Have Adequate Information for Reasoning? An Enhancement Method for Imperfect Video Understanding with Large Language Model

Mingxin Li[†] , **Wenhao Wang**[†] , **Hongru Ji** , **Xianghua Li**[*] and **Chao Gao**
Northwestern Polytechnical University
{mingxinli, w.wenhao, jihongru}@mail.nwpu.edu.cn, {li_xianghua, cgao}@nwpu.edu.cn

Abstract

Large Language Models (LLMs) demonstrate strong capabilities in video understanding. However, it exhibits hallucinations and factual errors in video description. On the one hand, existing Multimodal Large Language Models (MLLMs) are primarily trained by combining language models and vision models, with their visual understanding capabilities depending on the performance of the backbone. Moreover, video descriptions often suffer from incomplete content and the possibility of errors. Given the proven assessment of the strong reasoning capabilities of LLMs, this paper proposes **ERSR**, a novel **E**ntity and **R**elationship based **S**elf-Enhanced **R**easoning method for imperfect video understanding. Specifically, an entities and relationships strategy is designed to perform scene graphs based on the limited observed entity relationships, thereby enhancing video descriptions. Furthermore, by providing question feedbacks, a self-enhanced forward and feedback reasoning strategy is provided to enhance reasoning logic. Finally, the prediction question answering results are re-validated through rethinking and verifying using the LLMs. Extensive experiments show that the proposed method achieves competitive results on real-world video understanding datasets, with an overall improvement of no less than 1.4%.

1 Introduction

Nowadays, Large Language Models (LLMs) have achieved tremendous success in Natural Language Processing (NLP) [Touvron *et al.*, 2023]. Multimodal Large Language Models (MLLMs), which can simultaneously address tasks such as object detection and commonsense reasoning, offer greater potential for development and have garnered even more attention from researchers [Ataallah *et al.*, 2024]. Compared to text and images, videos have more complex and heterogeneous modalities in comparison to the common tasks in Natural Language Processing and Computer Vision. A significant amount of work has been conducted to address video understanding, achieving effective results. However, for video understanding and video question answering, especially reasoning based on LLM-generated video descriptions, is an urgent and pressing issue that needs to be addressed.

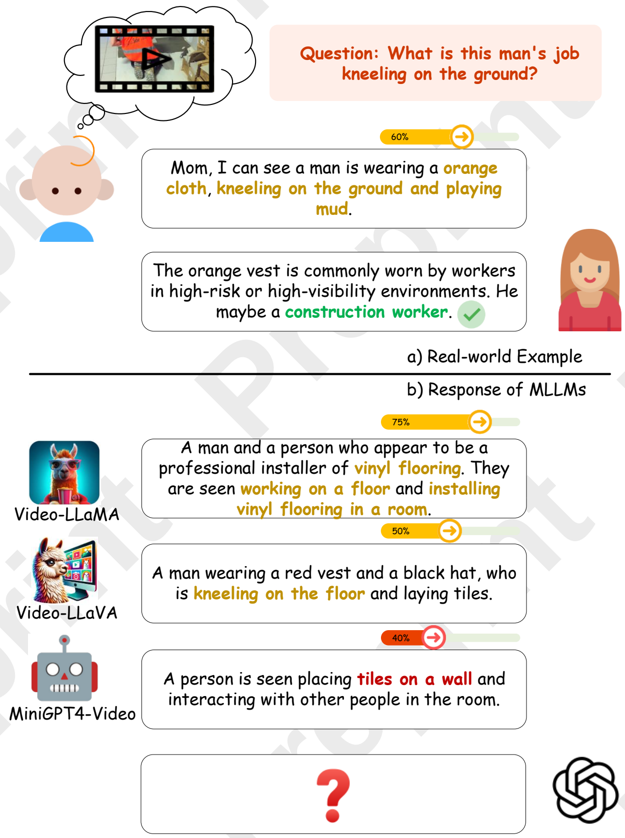


Figure 1: Illustration of the reasoning processes for both humans and LLMs when faced with an incomplete description. a) In the case of human reasoning, when a child provides an incomplete description of visual information, an adult can infer the result through the available information and background knowledge. b) As for the reasoning process of LLMs, the derivation from incomplete descriptions is still a subject of exploration.

Existing MLLMs primarily rely on vision-language generation methods, such as CLIP, and leverage the powerful reasoning capabilities of LLMs to understand visual content [Radford *et al.*, 2021]. For video understanding, existing methods primarily involve extracting frames from videos and then using LLMs to generate contents. To understand the increasingly complex and varied videos, some studies have been proposed. The main methods include BLIP-2 [Li *et al.*, 2023c], Instruct-BLIP [Dai *et al.*, 2023], Video-LLaMA [Zhang *et al.*, 2023], Video-LLaVA [Lin *et al.*, 2024], Video-ChatGPT [Maaz *et al.*, 2024], and MiniGPT4-Video [Ataallah *et al.*, 2024], etc. These methods have demonstrated powerful understanding capabilities in video understanding, enabling to integrate and response to rapidly changing visual dynamics.

However, in real-life scenarios, there are still many cases of incomplete video descriptions. As shown in Figure 1 a), for instance, a child may be unable to provide a complete description of visual content due to a lack of understanding of the scene. This requires adults to infer and respond based on limited incomplete information. Similarly, for vision-language models, due to limitations in model capability and computational resources, it is often impossible to fully learn and provide a complete description of a video. Furthermore, as shown in Figure 1 b), due to limitations in their knowledge boundaries and reasoning capabilities, MLLMs retelling exhibits certain hallucinations and factual reasoning errors. Although current proprietary LLMs, such as ChatGPT, GPT-4*, and Doubao†, can enhance their knowledge retrieval capabilities through web searches, existing open-source LLMs still lag behind them to some extent. Overall, let us think a question. *Can retelling have adequate information for reasoning using MLLMs ?*

To address the above issues, this paper proposes **ERSR**, a novel **E**ntity and **R**elationship based **S**elf-Enhancing **R**easoning framework for imperfect video understanding. Specifically, first, an entity and relationships generation pipeline is proposed, which progressively performs rewriting, augmentation, entity recognition, and scene graph prediction and generation on the obtained video description (or retelling). Furthermore, a novel self-enhanced forward and feedback reasoning strategy is developed. Due to the existence of multi-turn questions for the same video, a forward step-by-step chain is designed, which self-enhances and expands the knowledge and video description through different questions. In addition, a feedback chain is proposed to allow LLMs to reassess the correctness of previous answers based on the expanded description. Pruning is also used to reduce computational costs. Finally, to verify the correctness of the generated content, a rethinking verification strategy is designed to recheck, ultimately achieving convergence. In general, the above reasoning method, enhanced by the reasoning capabilities of Qwen2.5-7B, achieved an overall improvement of more than 1.4%.

For these reasons, the main contributions of this paper are as following 3 aspects:

- **Entity and Relationships Generation.** Based on the powerful capabilities of large language models in entity extraction and relationship reasoning, an entity and relationships generation pipeline is designed to obtain the enhanced entity and relationships of the video descriptions.
- **Self-Enhanced Reasoning.** A self-enhanced reasoning method based on a forward step-by-step chain and feedback chain is proposed, which enhances the imperfect video descriptions.
- **Rethinking and Better Performance.** By using rethinking to reduce hallucinations of LLMs. Extensive experimental results show that the proposed method achieves competitive results in video understanding and question answering datasets.

*https://chat.openai.com
†https://www.doubao.com

2 Related Works

2.1 Video Understanding and Question Answering

Video understanding has developed over many years [Soomro, 2012]. Current video understanding methods mainly rely on transformer [Vaswani, 2017] and pre-training [Radford *et al.*, 2021]. Several methods, such as cross-modal attention, motion-appearance memory, and others, have been applied in video understanding and question answering [Jiang *et al.*, 2020; Liu *et al.*, 2021].

With the development of LLMs, video understanding and question answering, particularly for complex tasks, have gradually become dominant due to their exceptional inferential capabilities. Amount of benchmarks have been proposed to evaluate these advances [Xiao *et al.*, 2021; Grauman *et al.*, 2022; Mangalam *et al.*, 2023]. To address the issue of complex reasoning in videos, researchers have conducted extensive works [Li *et al.*, 2023a; Chen *et al.*, 2023b]. Some studies combine subtitles, visual information, and other modalities for video understanding [Wang *et al.*, 2022]. Other works focus on using ChatGPT to pose questions to visual-language models [Wang *et al.*, 2024b; Yang *et al.*, 2024]. Furthermore, some works have proposed question-guided visual description video question answering methods [Mogrovejo and Solorio, 2024]. However, most existing methods rely on proprietary LLMs such as GPT-3.5, GPT-4, etc. Although they have achieved SOTA on real-world datasets, the models are complex and overly dependent on the knowledge boundary of the LLM.

2.2 Large Language Model Based Enhancement Reasoning

Large Language Models exhibit their powerful capabilities through prompt enhancement. Existing research indicates that by applying appropriate prompts and step-by-step reasoning, such as Chain-of-Thought [Kojima *et al.*, 2022]. In recent years, many LLM enhancement methods have emerged for reasoning, such as training and calling external reasoning [Creswell and Shanahan, 2022], program interpreters [Schick *et al.*, 2023], RAG [Lewis *et al.*, 2020; Gao *et al.*, 2023], rethinking [He *et al.*, 2022; Ma *et al.*, 2024; He *et al.*, 2024], and knowledge-based question answering (KBQA) [Wang *et al.*, 2024c]. However, these approaches also introduce hallucinations and cumulative errors during the reasoning process, which limits the model's ability to perform multi-step reasoning.

For video reasoning based on LLMs, Video-of-Thought was proposed, which decomposes the original question into multiple sub-solutions [Fei *et al.*, 2024]. Additionally, VideoAgent applies methods such as object detection and object tracking, and designs a database retrieval approach to reduce the hallucination problem in MLLMs [Fan *et al.*, 2024]. However, the above solutions make full use of video but do not effectively address the incomplete visual detection generated by LLMs. Therefore, our method aims to design a zero-shot solution for incomplete video description reasoning and generation.

3 Methodology

This section will introduce the method ERSR in detail, the illustration and the system prompt are shown in Figure 2.

3.1 Preliminaries

Definition 1 (Video Understanding)**.** *It refers to the process of extracting and analyzing semantic information from video. The input is video* $V = \{v_1, v_2, \ldots, v_T\}$*, and the output is semantic labels, events, or action inferences* $Y = \{y_1, y_2, \ldots, y_M\}$*, where* y_i *represents an event or action class at a specific moment or region in the video. It involves various tasks such as object detection and tracking, action recognition, scene understanding, video captioning, video question answering, et al.*

Definition 2 (Scene Graph)**.** *A scene graph is defined as a graph* $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{A})$ *with objects or entities in the scene* $\mathcal{V}$ *and the set of edges representing the semantic relationships between the objects* $\mathcal{E}$*.* $\mathcal{A}$ *contains textual descriptions of the relationship of objects (e.g. "on", "take care of").*

Task 1 Video Question Answering (VQA) based Retelling is the task of generating answers Y by reasoning over video description V and natural language question $Q = \{q_1, \ldots q_L\}$, which can be expressed as:

$$Y = g(V, Q) \quad \text{where} \quad g : \mathbb{R}^N \times \mathbb{R}^L \to \mathbb{R}^{N \times L},$$

where N is the number of video descriptions, L is the number of the question of the same video.

3.2 Step 1: Entity and Relationships Generation

Based on the description incompleteness of LLMs, this section designs a video description enhancement method based on entity and relationship generation. The strategy rewrites and enhances the video description by further extracting entities and scene graphs, transforming the video scene relationships from natural language into Euclidean space, thereby improving the reasoning logic of the LLM.

Incomplete Description Rewriting and Augment

The video description mainly involves entities, scenes, and others from the video. These elements may not fully align with the question due to synonyms, similar relationships, or other factors. Therefore, the description and the question need to be aligned first. The description should be rewritten using the LLM, and the few-shot prompt is as follows:

> **Task**: There may be the video description and the question contains synonyms or ambiguous entity relationships. Please align the description and the question, and rewrite the description.
> **Assistent prompt:** For example, "young boy" and "baby" are synonyms. Give your rewritten description:

After this step, all synonyms and entity relationships will be aligned.

Furthermore, since the description provides an incomplete description of the video, it is necessary to pre-augment the description. The prompt format is as follows:

> **Task**: There is incompleteness in the description. Please expand the description and provide 5 complete descriptions that you consider comprehensive.

Compare the relevance of the 5 descriptions above and select the highest relevance to the original description as the candidate. The equation is (1).

$$sim(\mathbf{c}_i, \mathbf{c}_j) = (\sum_{k=1}^{D} |\mathbf{x}_{i,k} - \mathbf{x}_{j,k}|^D)^{1/D}, \tag{1}$$

where $\mathbf{c}_i, \mathbf{c}_j$ are the i-th and the j-th candidate description. $\mathbf{x}_i, \mathbf{x}_j$ are attributes of different generated descriptions, k is an iterator, and D here is the description feature dimension.

Entity Generation

The entity is the most fundamental aspect of natural language. Therefore, entity recognition plays a crucial role in language understanding. In this task, the entity is one of the objectives of the question answering process. So entities need to be extracted from both the description and the question. The prompt is as follows:

> **Task**: Based on the question, identify the entities in the scene, match them with the entities in the question and description.
> **Assistent prompt:** Provide the JSON format. You should only include the entities, and the format is as follows: [See Figure 2].

Furthermore, construct the scene graph based on the extracted entities.

Scene Graph Prediction & Generation

This task focuses on visual reasoning representations in natural language. Scene graphs play an important role in vision, as they help understand high-level semantic relationships in images through objects. Furthermore, providing an accurate scene graph can help avoid hallucinations caused by unclear task instructions in LLMs [Fei *et al.*, 2023]. Therefore, the prompt for constructing a scene graph is as follows:

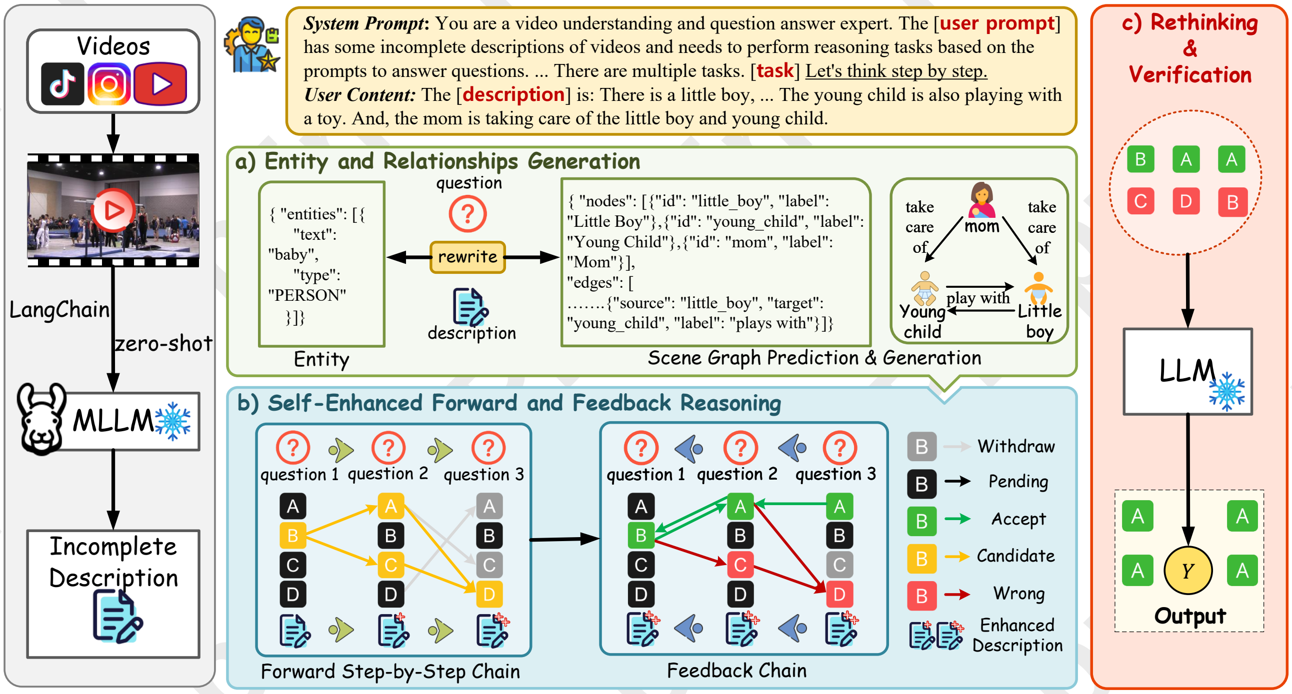


Figure 2: Illustration of the framework of the proposed method, ERSR. The whole framework is divided into 3 modules. After the zero-shot by the MLLM, the incomplete description is obtained. Through a) Entity and Relationships Generation, b) Self-Enhanced Forward and Feedback Reasoning, and c) Rethinking and Verification. **Note** that in b), the different colored options represent different states of reasoning.

> **Task**: Provide an inference process based on the scene and description, along with a scene relation diagram, and then extend the scene diagram through reasoning.
> **Assistent prompt:** Provide the JSON format.

In this step, the scene graph is constructed, which will be used for the next step of description reasoning.

3.3 Step 2: Self-Enhanced Forward and Feedback Reasoning

Since the same video contains multiple scenes, multiple questions are posed, each corresponding to different scenes. Through question history and the memory of the LLM, it can learn from the already completed reasoning and further understand and expand the incomplete description. Therefore, this section proposes a self-enhanced question answering reasoning method based on forward and feedback reasoning. By leveraging multiple sets of questions and answers for the same video, the subsequent questions are enhanced. Additionally, a feedback reasoning chain is constructed, which updates the understanding of the video description by reasoning feedback from existing facts.

Multi-QA Forward Step-by-Step Chain

Firstly, in the case where a video has multiple different perspectives, this paper designs a content selection scoring mechanism based on multiple options. For each option, output a score s_i, which represents the probability of the given question and option matching the description. The scoring prompt is as follows:

> **Task**: Based on the option provided, You need to analyze whether to choose that option. You should give the option a score from 1 to 10, where 1 means would not choose it, and 10 means would definitely choose it.
> **Assistent prompt:** Output the scoring in the following JSON format. [json example]

$$\mathcal{S}_{t,i} = f(q_t, d_t, m, \mathcal{O}_{t,i}), \tag{2}$$

where q_t is the t-th question of the same video, d_t is the t-th generated description, m is the LLM selected, and $\mathcal{O}_{t,i}$ is the i-th option of q_t. Moreover, $\mathcal{S}_{t,i}$ is the i-th predict score of q_t, i is the number of options, $\mathcal{S}$ is the matrix of all answers score of Q.

Due to the severe hallucination problem in smaller language models, it is necessary to specify the output format when calculating scores. Therefore, the score output will be in JSON format.

$$d_t = Prompt(d_{t-1}, q_{t-1}, \mathcal{S}_{t-1}, m), \tag{3}$$

where $q_{t-1}, \mathcal{S}_{t-1}$ is the last question and the last answer scores of the video question answering chain of the same video. Similarly, d_{t-1} is the last generated description. It needs to note that $t > 2$.

Furthermore, a step-by-step chain reasoning method is designed to progressively answer multiple questions about the same video while saving previous answers. The description is also enhanced using prompt engineering, with the equation 3 and the prompt as follows:

> **Task**: Rewrite the description. Provide a more detailed description based on [question] and [answer].
> **Assistant prompt**: The rewrite format should be: [new description]

Store the option scores calculated for each question in matrix $\mathcal{S}$. It is important to note that there may be cases where different questions have the same score for the options. The strategy is to accept all these options and perform greedy selection and pruning on the selected options.

$$\hat{s} = \arg\max\left(\mathcal{S}_k\right), \tag{4}$$

where $\hat{s}_k$ is the highest score of the k-th option across all the questions.

$$\hat{y_i} = \mathcal{S} \setminus \{\hat{s}_k\}, \tag{5}$$

where $\hat{y_i}$ is the predict option, and$\setminus$ represents the set difference, meaning all paths except for $\hat{s}_k$ are pruned.

Multi-QA Feedback Chain

After the forward step-by-step reasoning chain, a feedback reasoning chain is constructed. This involves revisiting previous questions and answers to refine the understanding of the video. The model updates its interpretation based on the existing facts.

$$d^U_{t-1} = h(q_t, y_t, d_t) \tag{6}$$

Furthermore, recalculate the new scores and re-evaluate the previously Withdrawn candidate options, updating matrix $\mathcal{A}$. Additionally, determine whether each option is predicted correct (Accept) or incorrect (Wrong), perform feedback calculation, and further prune the options. This process will yield the preliminary answers for all questions related to the video description.

3.4 Step 3: Rethinking

During the reasoning process, due to the complexity of the task and the length of the entire reasoning chain, the LLM may generate wrong answers. Therefore, it is necessary to validate the answers provided by the LLM. In this section, Rethinking is used to verify the obtained results. The specific process is as follows:

1. Determine whether the answer has factual hallucinations by matching the chosen results with the question and checking whether the answer aligns with common-sense knowledge.
2. Determine whether the answer has fidelity hallucinations by checking if the chosen result contradicts the description.

> **Task:** Complete the answer verification task. The question is: [question], and the answer is: [answer], with alternative options: [options]. Complete the following two tasks [see above]:

During the verification process, if the LLM analyzes the given result and determines that the verification score is below the threshold, it will return to Step 2 for rethinking. After multiple rounds of rethinking and verification, the judgment will converge, leading to a more confident answer. The maximum round number is often set to 3 or 5.

4 Experiments

4.1 Benchmark Datasets

This paper uses 4 widely adopted datasets: IntentQA [Li *et al.*, 2023b], NExT-QA [Xiao *et al.*, 2021], Egoschema [Mangalam *et al.*, 2023], and ActivityNet-QA [Yu *et al.*, 2019]. Among them, NExT-QA and IntentQA involve multi-turn question answering, where videos are divided into different categories for directed multiple-choice questions. Egoschema and ActivityNet-QA each correspond to a single question per video, with Egoschema being a directed multiple-choice question and ActivityNet-QA being an undirected question.

4.2 Baselines

Numerous works have been proposed to understand video and answer questions. This section categorize them into fine-tuning methods, zero-shot methods, and MLLM methods.

Fine-tuning methods (FT)

There are several approaches such as the Small Language Models (SLMs) such as HQGA [Xiao *et al.*, 2022a], VGT [Xiao *et al.*, 2022b], CoVGT [Xiao *et al.*, 2023], HiTeA [Ye *et al.*, 2023] and MC-ViT-L [Balažević *et al.*, 2024]. For LLMs, there are BLIP-2 [Li *et al.*, 2023c], LLama-VQA [Ko *et al.*, 2023], Vamos [Wang *et al.*, 2024a] and CaVIR [Li *et al.*, 2023b].

Zero-shot methods (ZS)

There are some proprietary LLMs methods like LLoVi [Zhang *et al.*, 2024], MoReVQA [Min *et al.*, 2024],LVNet [Park *et al.*, 2024], IG-VLM [Kim *et al.*, 2023] and VideoAgent [Wang *et al.*, 2024b]. Additionally, open-source LLMs like VFC [Momeni *et al.*, 2023], SeViLA [Yu *et al.*, 2024] and Mistral [Jiang *et al.*, 2023] are also under consideration.

MLLM methods (MM)

In recent years, MLLMs for video have gained widespread attention. This paper primarily compares SOTA methods such as MiniGPT-v2 [Chen *et al.*, 2023a], MiniGPT4-Video [Ataallah *et al.*, 2024], Video-LLaMA [Zhang *et al.*, 2023], Video-LLaVA, LLaVA-NEXT-Video [Lin *et al.*, 2024], LangRepo [Kahatapitiya *et al.*, 2024], and LLoVi [Ge *et al.*, 2024]. It is important to note that the above methods only use a zero-shot prompt approach and do not employ chain-of-thought.

4.3 Experimental Setup

The experiments use AMD Ryzen Threadripper PRO 5995WX 64-Cores CPU and 4 × A6000, and the maximum GPU memory usage for LLM is 16G. Due to the strong hallucinations in LLMs, all experiments are repeated 5 times and take the average. Furthermore, for the video question answering task, accuracy is used for evaluation. Specifically, for ActivityNetQA, the open-ended answering dataset, the score [Wu *et al.*, 2025] is also used.

	Method	Pretrain	Params	NExT-QA C	NExT-QA T	NExT-QA D	NExT-QA All	IntentQA W	IntentQA H	IntentQA B	IntentQA All	Egoschema Subset	ActivityNet-QA Acc	ActivityNet-QA Score
	HQGA	✓	46M	-	-	-	-	48.2	54.3	41.7	47.7	-	-	-
	CoVGT	✓	149M	58.8	57.4	69.3	50.0	-	-	-	-	-	-	-
	HiTeA	✓	297M	62.4	58.3	75.6	63.1	-	-	-	-	-	-	-
	MC-ViT-L	✓	424M	-	-	-	65.0	-	-	-	-	62.6	-	-
FT	VGT	✓	511M	-	-	-	-	51.4	56.0	47.6	51.3	-	-	-
	BLIP-2	✓	4B	70.1	65.2	80.1	70.1	-	-	-	-	-	-	-
	LLama-VQA	✓	7B	72.7	69.2	75.8	72.0	-	-	-	-	-	-	-
	Vamos	✓	7B	72.6	69.6	78.0	72.5	69.5	70.2	65.0	68.5	-	-	-
	CaVIR	✓	175B	-	-	-	-	58.4	65.5	50.5	57.6	-	-	-
	MoReVQA	×	340B	70.2	64.6	-	69.2	-	-	-	-	-	-	-
	LVNet	×	1.8T	75.0	65.5	81.5	72.9	75.0	74.4	62.1	71.7	68.2	-	-
	IG-VLM	×	1.8T	69.8	63.6	74.7	68.6	-	-	-	64.2	-	-	-
ZS	VideoAgent	×	1.8T	72.7	64.5	81.1	71.3	-	-	-	-	60.2	-	-
	SeviLA	×	4B	-	-	-	-	-	-	-	60.9	25.7	-	-
	Mistral	×	478M	51.0	48.1	57.4	51.1	52.7	55.4	41.5	50.4	-	-	-
	LLoVi	×	1.8T	69.5	61.0	75.6	67.7	68.4	67.4	51.1	64.0	-	-	-
	MiniGPT2	×	7B	52.0	51.1	52.8	51.9	51.7	53.1	39.2	48.4	38.9	22.4	2.4
	MiniGPT4-Video	×	7B	50.2	47.2	49.4	49.3	49.8	56.6	45.2	48.9	30.0	21.0	2.3
	Video-LLaMA	×	7B	38.2	50.1	47.7	48.8	53.7	55.8	33.7	51.3	34.4	17.4	2.0
MM	Video-LLaVA	×	7B	58.7	53.9	60.1	57.0	54.9	50.5	36.2	50.6	36.8	22.0	2.2
	LLaVA-NEXT-Video	×	7B	57.9	55.0	60.6	58.1	57.2	59.7	43.1	55.2	44.1	27.7	**2.8**
	LangRepo	×	7B	57.8	45.7	61.9	54.6	56.9	60.2	42.1	53.8	**60.8**	-	-
	LLoVi	×	12B	60.2	51.2	66.0	58.2	59.7	**62.7**	45.1	53.6	-	-	-
	ERSR(Ours)*	×	7B	60.4	56.6	69.9	60.6	59.8	58.7	**51.6**	56.6	47.6	**28.6**	2.6
	ERSR(Ours)	×	7B	**66.8**	**63.5**	**72.7**	**66.0**	**65.1**	**62.7**	43.9	**59.5**	-	-	-

Table 1: Performace of all the 4 datasets. The **bolded** represents the best zero-shot performance among open-source LLMs, while the underlined is the second-best. For the proposed method, ERSR* refers to the version without applying the verification strategy, whereas ERSR refers to the full pipeline. "-" represents poor performance or no mention.

4.4 Overall Performance

The experimental results are shown in Table 1. Overall, the proposed method achieves competitive performance on all datasets. The four datasets, due to their different features, lead to distinct application scenarios. In the fine-tuning methods, both NExT-QA and IntentQA achieve close to 70%. However, with the application of large models, the training costs for multimodal models also increase. For the zero-shot results of proprietary large models, it can be observed that at parameter scales of hundreds of billions (e.g., GPT-3.5) and 1.8T (e.g., GPT-4), most methods achieve optimal performance. However, the API call costs for proprietary large models also increase with their accuracy. This paper primarily explores the application scenarios of open-source LLMs. The experimental results show that the proposed ERSR method achieves competitive results, regardless of whether the verification is applied. The method improves performance by 1.4% to 7.8% on the above two datasets.

Furthermore, through experiments on the verify effect, it can be concluded that performing 3-5 rounds of verification, until the results converge, can effectively reduce hallucinations and factual errors in LLMs.

However, for the Egoschema and ActivityNet-QA datasets, the experimental results are not ideal, and there is still room for improvement compared to methods like LangRepo. The main reason is that for incomplete video descriptions, the content is limited, and the model is unable to extract useful information from such sparse data.

4.5 Ablation Study

Method	NExT-QA C	NExT-QA T	NExT-QA D	NExT-QA All	IntentQA W	IntentQA H	IntentQA B	IntentQA All
ERSR(Ours)*	60.4	56.6	69.9	60.6	59.8	58.7	51.6	56.6
w/o Step1	62.0	47.2	63.5	58.0	58.8	65.6	44.4	56.4
w/o Step2	57.5	52.8	63.2	57.3	56.3	63.5	44.1	55.2

Table 2: Ablation study of ERSR.

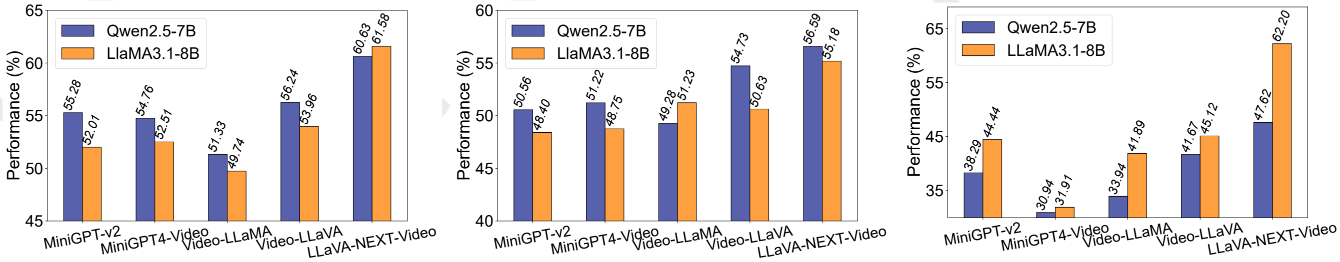
To verify the effectiveness and robustness of the proposed method, two ablation experiments were designed in this paper: Different Strategies and Different Backbones.

Different Strategies
Since the verify method has been tested and applied in Section 4.4, as shown in Table 2, this section only examines the effects of Step 1 and Step 2.

- **w/o Step 1**. Without entity and scene graph extraction, directly using a simple prompt combined with self-entity augmentation.
- **w/o Step 2** Using Chain-of-Thought to extract entities and scene graphs from incomplete descriptions, without considering the enhancement of the same video question answering.

Experiments show that both Step 1 and Step 2 enhance the VQA task on both datasets. When Step 1 is removed, performance drops by 2.6% and 0.2% on the NExT-QA and Inten-

Preprint – IJCAI 2025: This is the accepted version made available for conference attendees.
Do not cite. The final version will appear in the IJCAI 2025 proceedings.



(a) NExT-QA (b) IntentQA (c) Egoschema

Figure 3: Comparison of reasoning results between LLaMA-3.1-8B and Qwen2.5-7B on the NExT-QA, IntentQA, and Egoschema datasets.

tQA datasets, respectively. The main reason is that Step 1 extracts key descriptive information from natural language, reducing LLM's hallucination. As for Step 2, the performance decreases by 3.3% and 1.4%, respectively. This strategy can perform self-enhancement on the previous question history, effectively feeding back valuable information.

Different Backbones

To validate the relationship between the proposed method and the model backbone, experiments were conducted on LLaMA3.1-8B and Qwen2.5-7B. The experimental results are shown in Figure 3.

Overall, both models achieve similar performance across the three datasets. However, Qwen2.5-7B performs better on NExT-QA and IntentQA, while LLaMA3.1-8B shows better performance on Egoschema.

Therefore, it can be concluded that the performance of LLM-based question answering is related to the model's reasoning ability as well as the distribution of the dataset.

4.6 Case Study

Figure 4 shows the application results of the proposed method on a video from the IntentQA dataset. Through the entire pipeline, it can be observed that the method logically infers incomplete descriptions (retelling). By extracting entities, scene graphs, and considering the relationships between different questions, it is ultimately able to complete the question answering task.

5 Conclusion and Future Work

This paper proposes a novel Entity and Relationship-based Self-Enhanced Reasoning method for imperfect video understanding. Specifically, three modules are designed: a) Entity and Relationships Generation, b) Self-Enhanced Forward and Feedback Reasoning, and c) Rethinking and Verification. Through these three modules, the method enhances the ability of open-source Large Language Models to process incomplete video descriptions. Experiments on four datasets demonstrate the effectiveness of the proposed method.

However, it should be noted that due to the limitations in the reasoning abilities of some open-source LLMs, hallucination issues may arise during generation, leading to incorrect reasoning results. Therefore, the next step is to reduce hallucinations in LLMs, which plays an important role in reasoning with incomplete information.

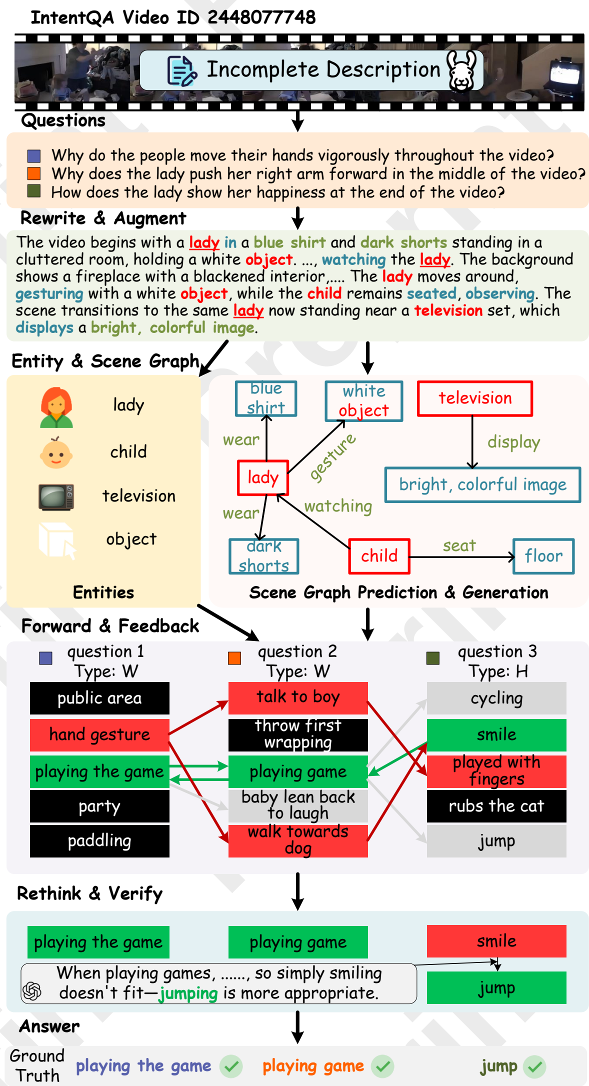


Figure 4: Illustration of the case study in IntentQA dataset.

Contribution Statement

Dr. Xianghua Li* is the corresponding author of this paper. Mingxin Li† and Wenhao Wang† contribute equally.

Acknowledgements

This research was supported by the National Natural Science Foundation of China (Nos. 62271411, U22A2098, 62471403, 62261136549), the Technological Innovation Team of Shaanxi Province (No. 2025RS-CXTD-009), the International Cooperation Project of Shaanxi Province (No. 2025GH-YBXM-017), the Fundamental Research Funds for the Central Universities (Nos. G2024WD0151, D5000240309).

References

[Ataallah *et al.*, 2024] Kirolos Ataallah, Xiaoqian Shen, Eslam Abdelrahman, et al. MiniGPT4-video: Advancing multimodal llms for video understanding with interleaved visual-textual tokens. *arXiv preprint arXiv:2404.03413*, 2024.

[Balažević *et al.*, 2024] Ivana Balažević, Yuge Shi, Pinelopi Papalampidi, Rahma Chaabouni, Skanda Koppula, and Olivier J Hénaff. Memory consolidation enables long-context video understanding. *arXiv preprint arXiv:2402.05861*, 2024.

[Chen *et al.*, 2023a] Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, et al. MiniGPT-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*, 2023.

[Chen *et al.*, 2023b] Liangyu Chen, Bo Li, Sheng Shen, et al. Language models are visual reasoning coordinators. In *ICLR 2023 Workshop on Mathematical and Empirical Understanding of Foundation Models*, 2023.

[Creswell and Shanahan, 2022] Antonia Creswell and Murray Shanahan. Faithful reasoning using large language models. *arXiv preprint arXiv:2208.14271*, 2022.

[Dai *et al.*, 2023] Wenliang Dai, Junnan Li, D Li, et al. Instructblip: Towards general-purpose vision-language models with instruction tuning. arxiv 2023. *arXiv preprint arXiv:2305.06500*, 2, 2023.

[Fan *et al.*, 2024] Yue Fan, Xiaojian Ma, Rujie Wu, et al. Videoagent: A memory-augmented multimodal agent for video understanding. In *ECCV*, pages 75–92, 2024.

[Fei *et al.*, 2023] Hao Fei, Qian Liu, Meishan Zhang, et al. Scene graph as pivoting: Inference-time image-free unsupervised multimodal machine translation with visual scene hallucination. In *ACL*, pages 5980–5994, 2023.

[Fei *et al.*, 2024] Hao Fei, Shengqiong Wu, Wei Ji, et al. Video-of-thought: Step-by-step video reasoning from perception to cognition. In *ICML*, 2024.

[Gao *et al.*, 2023] Yunfan Gao, Yun Xiong, Xinyu Gao, et al. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2023.

[Ge *et al.*, 2024] Suyu Ge, Yunan Zhang, Liyuan Liu, Minjia Zhang, Jiawei Han, and Jianfeng Gao. Model tells you what to discard: Adaptive KV cache compression for LLMs. In *ICLR*, 2024.

[Grauman *et al.*, 2022] Kristen Grauman, Andrew Westbury, Eugene Byrne, et al. Ego4D: Around the world in 3,000 hours of egocentric video. In *CVPR*, pages 18995–19012, 2022.

[He *et al.*, 2022] Hangfeng He, Hongming Zhang, and Dan Roth. Rethinking with retrieval: Faithful large language model inference. *arXiv preprint arXiv:2301.00303*, 2022.

[He *et al.*, 2024] Bolei He, Nuo Chen, Xinran He, Lingyong Yan, Zhenkai Wei, et al. Retrieving, rethinking and revising: The chain-of-verification can improve retrieval augmented generation. In *Findings of EMNLP*, pages 10371–10393, 2024.

[Jiang *et al.*, 2020] Jianwen Jiang, Ziqiang Chen, Haojie Lin, et al. Divide and conquer: Question-guided spatio-temporal contextual attention for video question answering. In *AAAI*, volume 34, pages 11101–11108, 2020.

[Jiang *et al.*, 2023] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, et al. Mistral 7B. *arXiv preprint arXiv:2310.06825*, 2023.

[Kahatapitiya *et al.*, 2024] Kumara Kahatapitiya, Kanchana Ranasinghe, Jongwoo Park, and Michael S Ryoo. Language repository for long video understanding. *arXiv preprint arXiv:2403.14622*, 2024.

[Kim *et al.*, 2023] Sungdong Kim, Jin-Hwa Kim, Jiyoung Lee, and Minjoon Seo. Semi-parametric video-grounded text generation. *arXiv preprint arXiv:2301.11507*, 2023.

[Ko *et al.*, 2023] Dohwan Ko, Ji Lee, Woo-Young Kang, et al. Large language models are temporal and causal reasoners for video question answering. In *EMNLP*, pages 4300–4316, 2023.

[Kojima *et al.*, 2022] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, et al. Large language models are zero-shot reasoners. *NeurIPS*, 35:22199–22213, 2022.

[Lewis *et al.*, 2020] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. *NeurIPS*, 33:9459–9474, 2020.

[Li *et al.*, 2023a] Dongxu Li, Junnan Li, Hung Le, et al. Lavis: A one-stop library for language-vision intelligence. In *ACL*, pages 31–41, 2023.

[Li *et al.*, 2023b] Jiapeng Li, Ping Wei, Wenjuan Han, and Lifeng Fan. Intentqa: Context-aware video intent reasoning. In *ICCV*, pages 11963–11974, 2023.

[Li *et al.*, 2023c] Junnan Li, Dongxu Li, Silvio Savarese, et al. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, pages 19730–19742, 2023.

[Lin *et al.*, 2024] Bin Lin, Yang Ye, Bin Zhu, , et al. Video-LLaVA: Learning united visual representation by alignment before projection. In *EMNLP*, pages 5971–5984, 2024.

[Liu *et al.*, 2021] Fei Liu, Jing Liu, Weining Wang, and Hanqing Lu. HAIR: Hierarchical visual-semantic relational reasoning for video question answering. In *ICCV*, pages 1698–1707, 2021.

[Ma *et al.*, 2024] Wufei Ma, Kai Li, Zhongshi Jiang, Moustafa Meshry, Qihao Liu, et al. Rethinking video-text understanding: Retrieval from counterfactually augmented data. In *ECCV*, pages 254–269, 2024.

[Maaz *et al.*, 2024] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. Video-ChatGPT: Towards detailed video understanding via large vision and language models. In *ACL*, pages 12585–12602, 2024.

[Mangalam *et al.*, 2023] Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. *NeurIPS*, 36:46212–46244, 2023.

[Min *et al.*, 2024] Juhong Min, Shyamal Buch, Arsha Nagrani, et al. MoReVQA: Exploring modular reasoning models for video question answering. In *CVPR*, pages 13235–13245, 2024.

[Mogrovejo and Solorio, 2024] David Mogrovejo and Thamar Solorio. Question-instructed visual descriptions for zero-shot video answering. In *Findings of the ACL*, pages 9329–9339, 2024.

[Momeni *et al.*, 2023] Liliane Momeni, Mathilde Caron, Arsha Nagrani, Andrew Zisserman, and Cordelia Schmid. Verbs in action: Improving verb understanding in video-language models. In *ICCV*, pages 15579–15591, 2023.

[Park *et al.*, 2024] Jongwoo Park, Kanchana Ranasinghe, Kumara Kahatapitiya, et al. Too many frames, not all useful: Efficient strategies for long-form video QA. *arXiv preprint arXiv:2406.09396*, 2024.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Hallacy, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763, 2021.

[Schick *et al.*, 2023] Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, et al. Toolformer: Language models can teach themselves to use tools. *NeurIPS*, 36:68539–68551, 2023.

[Soomro, 2012] K Soomro. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

[Touvron *et al.*, 2023] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, et al. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[Vaswani, 2017] A Vaswani. Attention is all you need. *NeurIPS*, 2017.

[Wang *et al.*, 2022] Yi Wang, Kunchang Li, Yizhuo Li, et al. Internvideo: General video foundation models via generative and discriminative learning. *arXiv preprint arXiv:2212.03191*, 2022.

[Wang *et al.*, 2024a] Shijie Wang, Qi Zhao, Minh Quan Do, Nakul Agarwal, Kwonjoon Lee, and Chen Sun. Vamos: Versatile action models for video understanding. In *ECCV*, pages 142–160, 2024.

[Wang *et al.*, 2024b] Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. Videoagent: Long-form video understanding with large language model as agent. In *ECCV*, pages 58–76, 2024.

[Wang *et al.*, 2024c] Yu Wang, Nedim Lipka, Ryan A Rossi, et al. Knowledge graph prompting for multi-document question answering. In *AAAI*, volume 38, pages 19206–19214, 2024.

[Wu *et al.*, 2025] Zhixuan Wu, Bo Cheng, Jiale Han, Jiabao Ma, Shuhao Zhang, Yuli Chen, and Changbo Li. VideoQA-TA: Temporal-aware multi-modal video question answering. In *COLING*, pages 7239–7252, 2025.

[Xiao *et al.*, 2021] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. NExT-QA: Next phase of question-answering to explaining temporal actions. In *CVPR*, pages 9777–9786, 2021.

[Xiao *et al.*, 2022a] Junbin Xiao, Angela Yao, Zhiyuan Liu, et al. Video as conditional graph hierarchy for multi-granular question answering. In *AAAI*, volume 36, pages 2804–2812, 2022.

[Xiao *et al.*, 2022b] Junbin Xiao, Pan Zhou, Tat-Seng Chua, and Shuicheng Yan. Video graph transformer for video question answering. In *ECCV*, pages 39–58, 2022.

[Xiao *et al.*, 2023] Junbin Xiao, Pan Zhou, Angela Yao, Yicong Li, et al. Contrastive video question answering via video graph transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[Yang *et al.*, 2024] Zongxin Yang, Guikun Chen, Xiaodi Li, Wenguan Wang, and Yi Yang. DoraemonGPT: Toward understanding dynamic scenes with large language models (exemplified as a video agent). In *ICML*, 2024.

[Ye *et al.*, 2023] Qinghao Ye, Guohai Xu, Ming Yan, Haiyang Xu, Qi Qian, Ji Zhang, and Fei Huang. HiTeA: Hierarchical temporal-aware video-language pre-training. In *ICCV*, pages 15405–15416, 2023.

[Yu *et al.*, 2019] Zhou Yu, Dejing Xu, Jun Yu, et al. ActivityNet-QA: A dataset for understanding complex web videos via question answering. In *AAAI*, volume 33, pages 9127–9134, 2019.

[Yu *et al.*, 2024] Shoubin Yu, Jaemin Cho, Prateek Yadav, and Mohit Bansal. Self-chained image-language model for video localization and question answering. *NeurIPS*, 36, 2024.

[Zhang *et al.*, 2023] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. In *EMNLP*, pages 543–553, 2023.

[Zhang *et al.*, 2024] Ce Zhang, Taixi Lu, Md Mohaiminul Islam, Ziyang Wang, et al. A simple LLM framework for long-range video question-answering. In *EMNLP*, pages 21715–21737, 2024.