ReasVQA: Advancing VideoQA with Imperfect Reasoning Process

Jianxin Liang[1] , **Xiaojun Meng**[2] , **Huishuai Zhang**[1,3*] ,
Yueqian Wang[1] , **Jiansheng Wei**[2] , **Dongyan Zhao**[1,3*]
[1] Wangxuan Institute of Computer Technology, Peking University
[2] Huawei Noah's Ark Lab
[3] National Key Laboratory of General Artificial Intelligence, Peking University
{liangjx, wangyueqian, zhanghuishuai, zhaody}@pku.edu.cn,
{xiaojun.meng, weijiansheng}@huawei.com

Abstract

Video Question Answering (VideoQA) is a challenging task that requires understanding complex visual and temporal relationships within videos to answer questions accurately. In this work, we introduce **ReasVQA** (Reasoning-enhanced Video Question Answering), a novel approach that leverages reasoning processes generated by Multimodal Large Language Models (MLLMs) to improve the performance of VideoQA models. Our approach consists of three phases: reasoning generation, reasoning refinement, and learning from reasoning. First, we generate detailed reasoning processes using additional MLLMs, and second refine them via a filtering step to ensure data quality. Finally, we use the reasoning data, which might be in an imperfect form, to guide the VideoQA model via multi-task learning, on how to interpret and answer questions based on a given video. We evaluate ReasVQA on three popular benchmarks, and our results establish new state-of-the-art performance with significant improvements of +2.9 on NExT-QA, +7.3 on STAR, and +5.9 on IntentQA. Our findings demonstrate the supervising benefits of integrating reasoning processes into VideoQA. Further studies validate each component of our method, also with different backbones and MLLMs, and again highlight the advantages of this simple but effective method. We offer a new perspective on enhancing VideoQA performance by utilizing advanced reasoning techniques, setting a new benchmark in this research field.

1 Introduction

Video Question Answering (VideoQA) (Patel et al., 2021; Zhong et al., 2022) is an increasingly important task within the fields of artificial intelligence and computer vision, aiming to enable machines to understand and answer questions about video content. It poses unique challenges due to the complex nature of video data (Zhong et al., 2022), which combines temporal and spatial information, often requiring deep contextual understanding and reasoning over sequences of frames.

Existing approaches to VideoQA (Pan et al., 2023; Liang et al., 2024; Wang et al., 2024; Yu et al., 2023) typically involve models that attempt to map video frames and questions directly to answers. While these methods have shown some success, they often fall short when complex reasoning or temporal relationships are involved (Mangalam et al., 2023; Kahatapitiya et al., 2024). These methods lack the ability to explicitly break down complex visual content into manageable components, which hinders the model's capacity for deeper comprehension and accurate responses. Meanwhile, process supervision (Lightman et al., 2023; Uesato et al., 2022) has been shown to be comparable to, or even outperform, outcome supervision in certain mathematical scenarios. Building on this insight, we try to enhance the model's performance in VideoQA by focusing on improving its advanced reasoning capabilities in this work. However, obtaining accurate and well-annotated reasoning processes, especially through human annotation, is both costly and time-consuming, making it challenging to scale this approach effectively.

On the other hand, Multimodal Large Language Models (MLLMs) (Team et al., 2023; Achiam et al., 2023; Chen et al., 2024a) have demonstrated impressive capabilities in generating detailed explanations, captions, and even reasoning processes for both images and videos, achieving SOTA across many tasks. These models also exhibit strong multimodal understanding (Lin et al., 2023; Li et al., 2024; Zhang et al., 2023b), capable of answering questions about unseen content, thus opening up new avenues for improving VideoQA. Leveraging their powerful multimodal chat abilities, MLLMs have shown the potential to automatically generate synthetic reasoning data, offering a way to bypass

[*] Corresponding authors: Huishuai Zhang and Dongyan Zhao.

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 1696–1709
April 29 - May 4, 2025 ©2025 Association for Computational Linguistics

the need for costly human annotations.

However, directly leveraging these synthetic reasoning processes presents significant challenges. While MLLMs can generate detailed and often insightful reasoning, their outputs are not always dependable. Due to inherent biases or occasional misunderstandings, these models may produce incorrect or irrelevant reasoning. For example, as shown in Table 4, a SOTA MLLM, InternVL (Chen et al., 2024b,a), demonstrates an accuracy of less than 70% across three VideoQA datasets when generating reasoning processes and final answers—leaving over 30% of outputs erroneous. Naturally, such reasoning processes may contain critical mistakes. Directly feeding this flawed synthetic reasoning into VideoQA models can hinder performance, as the inaccuracies may propagate throughout the learning process. This raises a crucial question: How to address the inconsistency between these imperfect reasoning steps and the true answers, and leverage them effectively during training?

We introduce **ReasVQA** (Reasoning-enhanced VideoQA), a novel approach that leverages the reasoning processes generated by MLLMs to improve the performance of VideoQA models. Our approach comprises three phases: Reasoning Generation (**RG**), Reasoning Refinement (**RR**), and Learning from Reasoning (**LR**), as shown in Figure 1. In the **RG** phase, we utilize SOTA MLLMs to produce reasoning processes for a set of VideoQA tasks, which are then utilized in the **LR** phase to train the actual model. Based on the outputs of the MLLMs, we evaluate the correctness of these reasoning processes by examining the associated answers, as illustrated in Figure 2. To mitigate the impact of potential errors within these reasoning processes, we apply data filtering to clean and refine the generated outputs during **RR** phase. Specifically, we remove the sentences containing answers, shifting the focus towards the reasoning steps rather than the conclusions. When the generated reasoning steps contains imperfections, these filtered reasoning steps still offer valuable insights for the model. The third phase, **LR**, employs a multi-task learning framework, where the VideoQA model is trained to both answer questions and generate reasoning processes simultaneously. The training supervision utilizes the dataset's original true answer annotations alongside the refined reasoning processes generated during the RG phase, ensuring that the model learns from both correct answers and the cleaned, structured reasoning paths. This phase allows the model not only to improve its question-answering accuracy but also to develop the ability to articulate the reasoning behind its answers. By integrating these curated reasoning processes into the training regimen, ReasVQA can inherit and refine the reasoning capabilities seen in larger MLLMs, leading to a more robust understanding of video.

We validate the effectiveness of our approach through extensive experiments across multiple model architectures and datasets. Experiments demonstrate significant improvements, achieving new state-of-the-art results with increases of +2.9 on NExT-QA, +7.3 on STAR, and +5.9 on IntentQA (Xiao et al., 2021; Wu et al., 2021; Li et al., 2023a). These findings validate the integration of generated reasoning processes into VideoQA models. Moreover, our detailed analyses confirm the effectiveness of each phase of the approach, offering insights into how reasoning refinement and multi-task learning contribute to the overall performance improvements.

In summary, the contributions of this paper are:

1. we propose ReasVQA to demonstrate the potential of using even imperfect reasoning processes from additional MLLMs to guide VideoQA models and also show how refined reasoning data can lead to significant performance gains.

2. we introduce a multi-task training method that incorporates reasoning into VideoQA tasks, offering a new perspective on learning from reasoning and knowledge integration.

3. we provide empirical evidence of our approach's effectiveness via rigorous experiments and analyses, setting new SOTAs in the VideoQA field.

2 Related Work

Video Question Answering VideoQA typically requires models to comprehend dynamic scenes, temporal information, and multimodal cues (such as visual, audio, and text) from videos. To better capture these features and understand their interactions across modalities, VideoQA methods have evolved significantly from earlier approaches like attention mechanisms, memory modules, and graph neural networks to leveraging more advanced pretrained models (Xu et al., 2017; Jang et al., 2017; Khan et al., 2020; Yang et al., 2020, 2021; Lei et al., 2021; Wang et al., 2022b; Ye et al., 2023; Gao et al., 2023; Wang et al., 2023a; Xu et al., 2023). For example, InternVideo (Wang et al., 2022b) extends a vision transformer pre-trained on images for video representation learning. Recently, with the growing capabilities of foundation models, some studies (Yu et al., 2023; Liang et al., 2024; Wang et al., 2024; Fei et al., 2024) have aimed to tackle VideoQA tasks by using large language models (LLMs)(Ko et al., 2023; Liu et al., 2024; Li et al., 2024). For instance, Wang et al. (2023b) and Ko et al. (2023) try to leverage LLaMA's (Touvron et al., 2023) knowledge of temporal and causal reasoning to address the complexities of VideoQA. However, these approaches typically rely solely on supervising the model using the final result, which often falls short or leads to overfitting, especially when complex reasoning or temporal relationships are involved.

Learning by using LLMs/MLLMs Recently, LLMs have demonstrated strong reasoning abilities (Wei et al., 2022; Kojima et al., 2022; Achiam et al., 2023; Chowdhery et al., 2023; Brown, 2020; Wang et al., 2022a), effectively solving multi-step reasoning tasks by explicitly performing intermediate reasoning steps using Chain-of-Thought or few-shot prompts. Furthermore, Uesato et al. (2022) and Lightman et al. (2023) find that process supervision, where intermediate reasoning steps are supervised, can yield performance that is comparable to or even better than final result supervision in certain mathematical scenarios. This provides a promising direction for integrating reasoning processes into VideoQA models in our work.

Inspired by the successes of LLMs, MLLMs have also explored similar strategies (Guo et al., 2023; Zeng et al., 2022; Li et al., 2023c; Zhang et al., 2023a; Romero and Solorio, 2024; Fei et al., 2024). For instance, models like LLoVi (Zhang et al., 2023a), utilizes GPTs to generate visual descriptions or summaries from captions. Q-ViD (Romero and Solorio, 2024) enhances zero-shot video understanding by incorporating captions relevant to the questions into the model's input. MotionEpic(Fei et al., 2024) breaks down the raw intricate video reasoning problem into a chain of simpler sub-problems and solves them one by one.

Unlike the aforementioned MLLM-based works that primarily rely on final result supervision, in this paper, we focus on mitigating the negative impact of errors in reasoning processes and explore different forms of supervision that can be used for training models, particularly process supervision, which leverages reasoning as a supervision signal. While this approach has been discussed and utilized in the context of LLMs, to the best of our knowledge, this is the first work to introduce process supervision into VideoQA. To achieve this, we decouple the reasoning steps from their predicted answers or conclusions through a simple refinement process, which allows the model to focus more on the reasoning steps during training, rather than the given predictions no matter correct or not. In the subsequent training phase, we employ a multi-task learning approach, where the model learns to both perform the VideoQA task and regenerate the reasoning process. We believe this enables the model to perform videoQA tasks following a logical manner and thus get better performance.

3 Methodology

To harness imperfect synthetic reasoning data to guide VideoQA, we propose the three-phase **ReasVQA** involving Reasoning Generation, Reasoning Refinement, and Learning from Reasoning.

3.1 Reasoning Generation

We utilize existing multimodal large language models (MLLMs) as reasoning process generators to produce a reasoning process r for a given video v and question q, as illustrated in Figure 2.

However, even SOTA models still regularly produce logical errors, especially in zero-shot settings, i.e., the predicted answer $\hat{a}$ included in r is sometimes incorrect. To assess the quality of the generated reasoning processes, we evaluate them based solely on the correctness of their predicted answers. Specifically, we compare $\hat{a}$ with the ground-truth labels a. If $\hat{a}$ does not equal a, we consider the reasoning process r to be imperfect, denoting it as **Incorrect Reasoning**; conversely, if they match, r is classified as **Correct Reasoning**.

For example, as illustrated in Figure 2, for the question "*Which object was tidied up by the person?*", the answer $\hat{a}$ is "*The closet/cabinet*" while "*The blanket*" is the correct answer. Therefore, this reasoning is deemed imperfect and classified as Incorrect Reasoning. This evaluation helps to assess the quality of reasoning processes generated by MLLMs and provides guidance for the next steps to better refine these imperfect reasoning processes.

3.2 Reasoning Refinement

As shown in Table 4, the accuracy of MLLMs is significantly lower than the current SOTAs in VideoQA tasks. This suggests that over 30% of the generated reasoning is unreliable. On the other

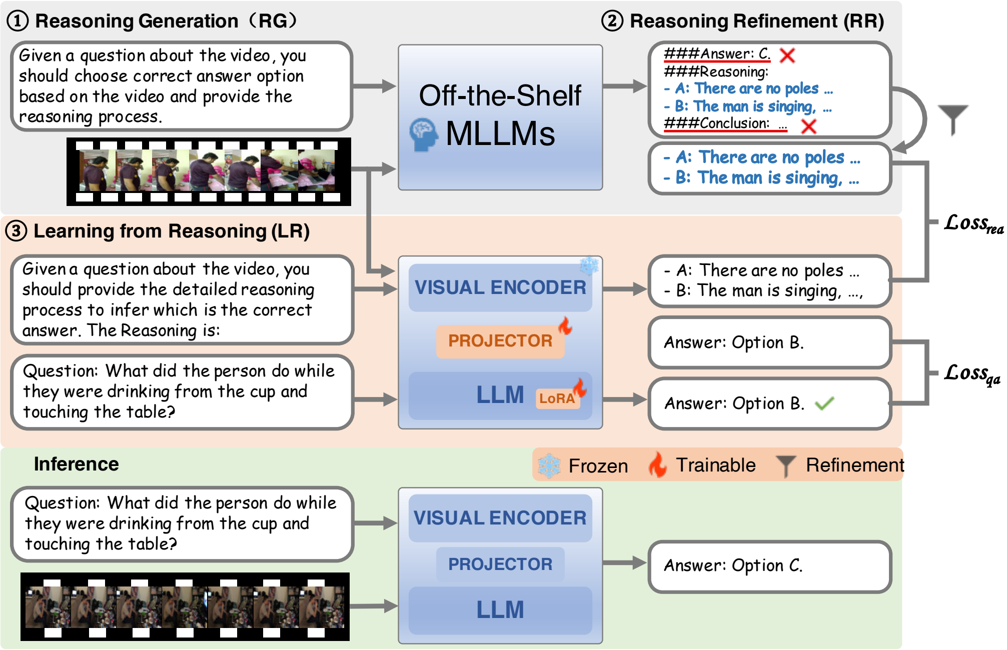

① Reasoning Generation（RG）
Given a question about the video, you should choose correct answer option based on the video and provide the reasoning process.
Off-the-Shelf MLLMs
② Reasoning Refinement (RR)
###Answer: C. ✗
###Reasoning:
- A: There are no poles ...
- B: The man is singing, ...
###Conclusion: ... ✗
- A: There are no poles ...
- B: The man is singing, ...
$Loss_{rea}$
③ Learning from Reasoning (LR)
Given a question about the video, you should provide the detailed reasoning process to infer which is the correct answer. The Reasoning is:
Question: What did the person do while they were drinking from the cup and touching the table?
VISUAL ENCODER
PROJECTOR
LLM
LoRA
- A: There are no poles ...
- B: The man is singing, ...,
Answer: Option B.
Answer: Option B. ✓
$Loss_{qa}$
Inference
Frozen
Trainable
Refinement
Question: What did the person do while they were drinking from the cup and touching the table?
VISUAL ENCODER
PROJECTOR
LLM
Answer: Option C.

Figure 1: Overview of our method ReasVQA. ① Reasoning Generation: a SOTA MLLM is prompted to solve complex questions by generating detailed reasoning explanations. ② Reasoning Refinement: we process and refine the reasoning steps to alleviate the conflict with true answers. ③ Learning from Reasoning (Multi-task Training): the refined reasoning steps are used to guide a model to improve its performance on the VideoQA tasks.

hand, with a close examination of the incorrect answers generated by MLLMs, we find that the reasoning steps are often valuable although the final answers are wrong. Therefore, we try to decouple the reasoning from the final predicted answer. This leads to the following refinement process: we retain only the essential steps that do not include any conclusion through keyword matching, regardless of whether they are classified as *Correct Reasoning* or *Incorrect Reasoning*. For *Incorrect Reasoning*, we further remove any words that contain the ground truth, which allows the model to focus on the logical flow of the reasoning.

This refinement is simple yet effective. We take a refined example in Figure 2, "*The individual is standing in front of a wooden cabinet with slatted doors*" provides relevant context for the video, while "*The other objects mentioned in the hints, such as the table and clothes, are not visible or being interacted with*" uses the logical process of elimination to exclude incorrect options. Both reasoning steps remain valuable, even if the final predicted answer is wrong and the sentences might be incomplete after refinement. This is also the reason we call it an imperfect reasoning process.

Overall, this refinement enables the VideoQA model to particularly focus on the structure of reasoning, filtering out erroneous information or shortcuts. It allows the model to strengthen the ability to generate coherent reasoning in response to video-based questions, improving performance across various VideoQA tasks. Importantly, this entire process occurs solely in the training phase, ensuring that no information leakage occurs in evaluation.

3.3 Learning from Reasoning

The third phase of our approach aims to effectively transfer the valuable information from refined reasoning processes $\hat{r}$ to a VideoQA model $f(\cdot)$. To accomplish this, we explore different training approaches, such as single-task learning (STL) and multi-task learning (MTL).

In the STL approach, we concatenate the refined reasoning $\hat{r}$ with the ground-truth answer a, denoted as $\hat{r}_a$, and use this joint text as the supervision signal for the model. The model is trained to generate both the reasoning and the correct answer sequentially. The objective function is:

$$\mathcal{L}_{st} = \mathcal{C}(f(v, q), \hat{r}_a). \quad (1)$$

where $\mathcal{C}(\cdot)$ denotes the cross-entropy loss. However, this method proves challenging because, even after refinement, there may still be inconsistencies within $\hat{r}_a$, making it difficult for the model to learn as well as reconcile discrepancies.

An alternative, more effective method is the MTL approach. In this way, the VideoQA model $f(\cdot)$ learns to simultaneously perform the VideoQA task and reconstruct the reasoning process $\hat{r}$. To enhance flexibility in model training, we use a weighted sum of two loss functions:

$$\mathcal{L}_{mt} = \alpha * \mathcal{C}_{qa}(f(v,q),a) + \beta * \mathcal{C}_{rea}(f(v,q),\hat{r}). \tag{2}$$

where α and β are weights such that $0 < \alpha, \beta < 1$ and $\alpha + \beta = 1$. $\mathcal{C}_{qa}(\cdot)$ and $\mathcal{C}_{rea}(\cdot)$ represent the cross-entropy losses for QA and reasoning generation, respectively. By integrating reasoning generation as an auxiliary task, MTL allows the model to benefit not only from the direct supervision of the true answers but also from learning the logical reasoning processes, thus enhancing its understanding and reasoning capabilities in complex video scenarios. Moreover, adjusting the balance between $\mathcal{C}_{qa}(\cdot)$ and $\mathcal{C}_{rea}(\cdot)$ with different weights helps mitigate the propagation of any residual error in the reasoning, ensuring that the model can still learn effectively even with imperfect reasoning.

4 Experiments

In this section, we present our experiments on various VideoQA tasks. First, we describe the datasets we used and the implementation details. Then, we evaluate our method, compare ReasVQA with other SOTAs, and provide a comprehensive analysis.

4.1 Datasets and Baselines

Datasets and Evaluation Metrics We conduct experiments on three popular VideoQA datasets: NExT-QA, STAR, and IntentQA (Xiao et al., 2021; Wu et al., 2021; Li et al., 2023a), which demand both causal and temporal reasoning abilities. For all these tasks, we employ the most used answer accuracy as the evaluation metric. A higher accuracy score indicates better model performance.

Implemententation For the **RG** phrase, we use InternVL(v1.5) (Chen et al., 2024a,b) as the reasoning generator, since it is currently the most powerful open-source MLLM. It has strong reasoning capabilities and matches the performance of commercial closed-source models such as GPT-4V, GPT-4O, and Gemini Pro (Team et al., 2023; OpenAI, 2024a,b) across various benchmarks. We uniformly sample N (N = 4 for faster generation here) frames from the video, then feed these frames to InternVL(26B) to generate complete reasoning processes for each dataset's training set via prompts.

For the **LR** phrase, we use BLIP-FlanT5 (Li et al., 2023b) as our model. Specifically, we employ ViT-G (Fang et al., 2023) as the visual encoder and initialize FlanT5 (Chung et al., 2022) (3B parameters) as the LLM. We only finetune the modality projection layers and LoRA weights of LLM (Hu et al., 2021) during training. We use this setting by default for experiments unless otherwise specified. See Appendix A.1 and A.2 for more details.

4.2 Model Performance Evaluation

Overall Reasults Tables 1 and 2 provide a comprehensive comparison of our method (ReasVQA, 3B) with existing methods across multiple benchmarks. In Table 1, which presents performance on the NExT-QA and STAR, our model consistently surpasses other methods. Our model achieves a total accuracy of 77.0% on NExT-QA and 74.5% on STAR, significantly outperforming other models. Notably, compared to methods that also utilize the 3B Flan-T5, our approach demonstrates a clear performance advantage, exceeding them by +2.9 and +7.3, respectively. ReasVQA even surpasses models such as LLaMA-VQA, MotionEpic, and Vamos, which rely on larger LLMs as answer generators. This highlights the robust capability of ReasVQA in tackling complex VideoQA tasks.

Table 2 further illustrates results on the IntentQA. Similar to the previous results in Table 1, our model achieves an impressive total accuracy of 77.0%, surpassing SOTA models by a significant margin, outperforming them by +5.9 (77.0% vs. 71.1%), even when they utilize larger LLMs. ReasVQA excels across all question types, with particularly strong performances in the 'How' (93.3%) and 'Tem.' (69.1%) categories. This demonstrates ReasVQA's effectiveness in handling a diverse range of question types and its superior capability in understanding and generating accurate answers.

The results above demonstrate the effectiveness of our approach, particularly in leveraging reasoning processes to enhance video understanding. Our model consistently achieves significantly better performance across various benchmarks, underscoring its robustness and versatility in VideoQA tasks. The detailed improvements in specific categories further validate the strengths of our methodology.

Adapting ReasVQA to Different Model Architectures In addition to integrating ReasVQA with encoder-decoder architecture LLMs like FlanT5, we further adapt ReasVQA to decoder-

Model	LLM Arch.	NExT-QA				STAR				
		Tem.	Cau.	Des.	Tot. ↑	Int.	Seq.	Pre.	Fea.	Tot. ↑
Non-LLM Models										
Just Ask (Yang et al., 2021)	-	51.4	49.6	63.1	52.3	-	-	-	-	-
All-in-One (Wang et al., 2023a)	-	48.6	48.0	63.2	50.6	47.5	50.8	47.7	44.0	48.9
MIST (Gao et al., 2023)	-	56.6	54.6	66.9	57.1	55.5	54.2	54.2	44.4	54.0
HiTeA (Ye et al., 2023)	-	58.3	62.4	75.6	63.1	-	-	-	-	-
InternVideo (Wang et al., 2022b)	-	58.5	62.5	75.8	63.2	62.7	65.6	54.9	51.9	58.7
LLM-based Models										
LLaMA-VQA (Ko et al., 2023)	LLaMA 7B	69.2	72.7	75.8	72.0	66.2	67.9	57.2	52.7	65.4
MotionEpic (Fei et al., 2024)	Vicuna 7B	74.6	75.8	83.3	76.0	71.5	72.6	66.6	62.7	71.0
Vamos (Wang et al., 2023b)	LLaMA2 7B	72.3	74.8	81.6	75.0	-	-	-	-	-
LSTP (Wang et al., 2024)	FlanT5 3B	66.5	72.8	81.2	72.1	-	-	-	-	-
SeViLA (Yu et al., 2023)	FlanT5 3B	67.0	73.8	81.8	73.8	66.4	70.3	61.2	55.7	67.2
VidF4 (Liang et al., 2024)	FlanT5 3B	69.6	74.2	83.3	74.1	68.4	70.4	60.9	59.4	68.1
ViLA (Wang et al., 2023c)	FlanT5 3B	71.4	73.6	81.4	74.1	70.0	70.4	61.2	55.7	67.2
ReasVQA (ours)	FlanT5 3B	**73.0**	**77.7**	**82.8**	**77.0**	**75.9**	**76.6**	**67.3**	**62.0**	**74.5**

Table 1: **Model comparison on NExT-QA and STAR**. Specifically, Tem., Cau., Des., and Tot. denote Temporal, Causal, Description, and Total accuracy, respectively. Int., Seq., Pre., and Fea. denote Interaction, Sequence, Prediction, and Feasibility, respectively.

Model	LLM	**IntentQA**			
		Why	How	Tem.	Tot. ↑
HQGA (Xiao et al., 2022a)	-	48.2	54.3	41.7	47.7
VGT (Xiao et al., 2022b)	-	51.4	56.0	47.6	51.3
BlindGPT (Ouyang et al., 2022)	GPT3	52.2	61.3	43.4	51.6
CaVIR (Li et al., 2023a)	-	58.4	65.5	50.5	57.6
Vamos (Wang et al., 2023b)	LLaMA3 8B	69.5	70.2	65.0	68.5
MotionEpic (Fei et al., 2024)	Vicuna 7B	-	-	-	70.8
LVNet (Park et al., 2024)	GPT-4o	75.2	71.6	60.8	71.1
ReasVQA (ours)	FlanT5 3B	**75.6**	**93.3**	**69.1**	**77.0**

Table 2: Model comparison on IntentQA.

Setting	**NExT-QA**	**STAR**	**IntentQA**	**Avg.** ↑
BLIP-FlanT5(3B)	74.5	71.0	73.0	72.8
w. ReasVQA	77.0	74.5	77.0	76.2(+3.4)
LLaVA-OV(0.5B)	64.2	60.0	58.9	61.0
w. ReasVQA	65.8	61.5	63.2	63.5(+2.5)
LLaVA-OV(7B)	77.5	66.2	74.1	72.6
w. ReasVQA	78.9	68.2	77.6	74.9(+2.3)

Table 3: ReasVQA improves the performance of different models.

Model	NExT-QA	STAR	IntentQA	Avg. ↑
InternVL(4B)	59.8	60.3	61.9	60.7
InternVL(26B)	67.0	69.8	69.0	68.6

Table 4: InternVL's accuracies on three VideoQA tasks.

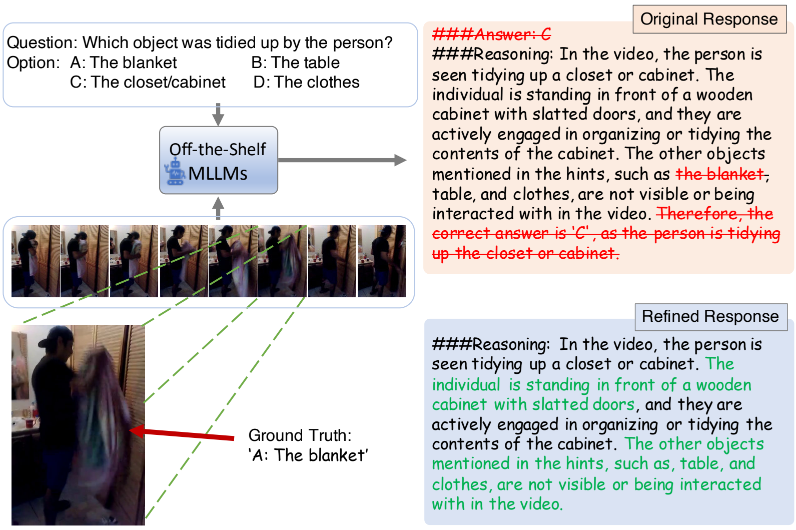
only architecture LLMs to demonstrate its generalizability and robustness. Specifically, we apply ReasVQA to LLaVA-OV (Li et al., 2024), one of the SOTA models in MLLMs that owns strong performance across various tasks. We conduct experiments using two different model scales of LLaVA-OV, namely LLaVA-OV (0.5B) and LLaVA-OV (7B), to validate our method.

The results in Table 3 clearly demonstrate the effectiveness of integrating our method across different model architectures and scales. ReasVQA consistently achieves better performance across models of varying sizes, with an average improvement of +2.5 and +2.3 for LLaVA-OV (0.5B) and LLaVA-OV (7B), respectively, across three tasks. Results indicate that our method is widely and easily adapted to different model architectures, and it scales well with larger models, making it suitable for both smaller and more complex architectures.

4.3 Data Impact Analysis

Effectiveness of Reasoning Refinement Even though existing MLLMs are strong, when used as zero-shot reasoning generators, they inevitably introduce errors in the reasoning steps due to inherent limitations. To mitigate the potential impact of such errors in the reasoning, we propose to process and refine these reasoning steps. To further explore it, we conduct experiments validating the effect of data refinement on model performance.

Setup We categorize the generated reasoning into Correct Reasoning (**CR**) and Incorrect Reasoning (**IR**) as described in Section 3.1. The proportion of CR serves as an indicator of reasoning accu-


Question: Which object was tidied up by the person?
Option: A: The blanket B: The table C: The closet/cabinet D: The clothes

~~###Answer: C~~
###Reasoning: In the video, the person is seen tidying up a closet or cabinet. The individual is standing in front of a wooden cabinet with slatted doors, and they are actively engaged in organizing or tidying the contents of the cabinet. The other objects mentioned in the hints, such as ~~the blanket,~~ table, and clothes, are not visible or being interacted with in the video. ~~Therefore, the correct answer is 'C', as the person is tidying up the closet or cabinet.~~

###Reasoning: In the video, the person is seen tidying up a closet or cabinet. The individual is standing in front of a wooden cabinet with slatted doors, and they are actively engaged in organizing or tidying the contents of the cabinet. The other objects mentioned in the hints, such as, table, and clothes, are not visible or being interacted with in the video.

Figure 2: An example of a response generated by an MLLM: the predicted final answer is “C”, while the true answer is “A: The blanket”. Although the final answer is incorrect, some of the reasoning steps still offer valuable learning elements for the model. For instance, sentences highlighted in green provide a partial description of the video and eliminate the possibility of two other options, providing meaningful insights.

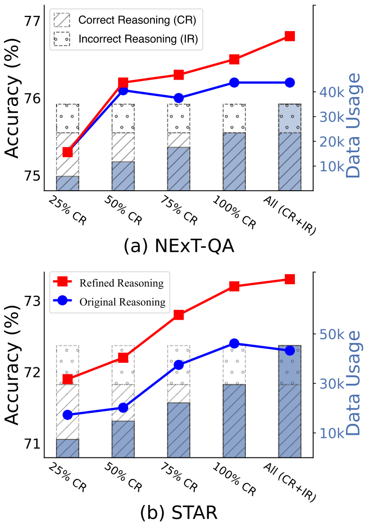


Figure 3: The impact of data quantity on performance and the comparison of Original and Refined Reasoning. Experiments are conducted using the multi-task learning approach, with $\alpha = \beta = 0.5$.

racy, and the results are shown in Table 4. For example, on NExT-QA, InternVL(26B) achieves an answer accuracy of 67%, indicating that 33% of the generated reasoning steps are flawed. In our experimental setup, we test ReasVQA trained using both CR and **All**, where **All** refers to using all data without distinguishing between CR and IR. For CR, we further experiment with using subsets of 25%, 50%, and 75% of the CR data.

The results shown in Figure 3 illustrate the impact of using refined reasoning versus original reasoning on the overall performance across a varying amount of reasoning. For NExT-QA, the performance using original reasoning shows minor fluctuations, with accuracy starting at 75.3% and slightly increasing to 76.2% as more correct reasoning data is used. In contrast, the refined reasoning demonstrates a steady upward trend, culminating in an accuracy of 76.8%, highlighting the effectiveness of reasoning refinement in enhancing performance consistently across different data usage levels. Similarly, the results on STAR exhibit the same trends.

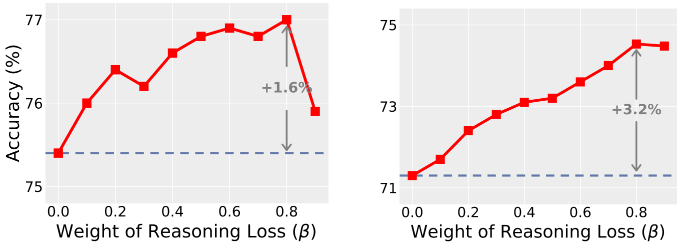
Based on the results, we can derive the following conclusions: (1) **refined reasoning consistently outperforms original reasoning** : In both datasets, the refined reasoning line shows a clear upward trajectory, indicating that refining reasoning substantially improves the model's performance. This effect is more pronounced as the amount of reasoning data increases. (2) **greater benefits at higher data usage levels**: While improvements are observed at lower levels (e.g., 25% and 50%), the differences become more substantial at 75% and 100%. This suggests that refinement is especially beneficial when a larger amount of reasoning steps is used, effectively leveraging the available correct data to enhance overall accuracy. (3) **plateau effect in original reasoning**: For both datasets, using the original reasoning without refinement results in performance that plateaus or slightly fluctuates at higher levels, implying that the presence of incorrect reasoning in the data might counteract the potential benefits of increased data volume.

We can draw consistent conclusions when using InternVL(4B) as the reasoning generator. For more experiments, see Appendix B.3.

4.4 Methodological Insights

Hyper-parameter Tuning To increase flexibility in model training, we introduce weighting parameters α and β into our framework, as shown in Equation 3.3, to control the influence of reasoning processes in the model's training objective. We perform a hyperparameter search to evaluate the model's performance as the weight β for the reasoning generation cross-entropy loss is varied.

Figure 4 presents the results on NExT-QA and STAR. When $\beta = 0$, only the QA loss contributes



(a) NExT-QA

(b) STAR

Figure 4: Hyper-parameter tuning for the weight β of the Reasoning Generation Loss $\mathcal{C}_{rea}(\cdot)$, with the corresponding $\alpha = 1 - \beta$.

Setting	**NExT-QA**	**STAR**	**IntentQA**	**Avg.** ↑
STL$_{QA}$	75.4	71.0	73.3	73.2
STL$_{CR}$	74.2	71.9	74.4	73.5
STL$_{All}$	73.3	70.6	70.8	71.6
MTL$_{CR}$	76.2	72.9	76.1	75.1
MTL$_{All}$	76.8	73.2	77.0	75.7

Table 5: The impact of different training approaches on the model performance. All experiments are conducted with $\alpha = \beta = 0.5$ for MTL setups.

to training, serving as a baseline for comparison. The results clearly show that integrating reasoning generation loss improves model performance. Specifically, on NExT-QA, accuracy steadily increases as β rises from 0 to 0.8, with accuracy peaking at 77.0% compared to the baseline of 75.4%, which suggests that incorporating the reasoning generation loss helps the model better understand video-based questions by enhancing contextual comprehension and reasoning capabilities. Performance stabilizes between $\beta = 0.5$ and $\beta = 0.8$, indicating an optimal balance between reasoning generation and QA objectives. However, when β exceeds 0.8, performance slightly declines to 75.9% at $\beta = 0.9$, indicating that overemphasizing reasoning generation may negatively impact VideoQA performance. Similarly, results on STAR exhibit the same trends and conclusions.

Overall, these results demonstrate the importance of balancing reasoning generation and QA training objectives. Notably, this flexibility offers substantial benefits when carefully managed, allowing the model to better leverage the strengths of both reasoning and question answering. Additionally, it enables customization based on different datasets or task complexities, further improving the model's adaptability and overall performance.

Single Task or Multi-Task? To further understand the impact of different training approaches, as mentioned in Section 3.3, we compare single-task learning (STL) and multi-task learning (MTL). In the STL setting, we evaluate three types of supervision signals: (1) using only the original true answer, denoted as $_{QA}$, (2) concatenating the true answer with the Correct Reasoning, meaning only samples with Correct Reasoning are used, denoted as $_{CR}$, and (3) concatenating all samples with their corresponding reasoning, denoted as $_{All}$.

The results are shown in Table 5. In the STL setting, we observe a minor performance improvement of +0.3 (73.2% vs. 73.5%) when correct reasoning is concatenated with the true answer. However, performance drops significantly by +1.6 (73.2% vs. 71.6%) when all reasoning samples, correct or not, are used. Conversely, in the MTL setting, using correct reasoning achieves a higher performance of 75.1%. After refining the reasoning, incorporating all samples leads to further improvements, reaching a performance of 75.7%.

These results highlight several key insights regarding the effectiveness of incorporating reasoning data: (1) Incorrect reasoning, when used in single-task training, can significantly degrade performance, indicating the importance of filtering incorrect data. (2) Multi-task learning consistently outperforms single-task setups, as it enables the model to better leverage reasoning processes, even when both correct and incorrect reasoning are present. (3) Joint learning of QA and reasoning generation provides substantial gains, highlighting the importance of multi-task learning for VideoQA.

See Appendix B for the abltion of each component and the impact of various linguistic materials.

4.5 Complex Reasoning Tasks

To further evaluate the complex reasoning ability of our method, we conduct experiments using the ATP-hard subset (Buch et al., 2022) of the NExT-QA validation set, which is for more challenging causal and temporal questions. This subset filters out those 'easy' questions that can be answered with a single frame. The results are in Table 6.

These results show that our method improves over VideoAgent (Wang et al., 2025) by 9.3% on this challenging subset, demonstrating its ability to enhance complex reasoning tasks and mitigate potential issues related to single-frame biases.

Model	Cau.	Tem.	Tot. ↑
Temporal[ATP] (Buch et al., 2022)	38.4	36.5	38.8
GF (Bai et al., 2024)	48.7	50.3	49.3
VideoAgent (Wang et al., 2025)	57.8	58.8	58.4
BLIP-FlanT5 (w. ReasVQA)	**69.2**	**65.7**	**67.7**(+9.3)

Table 6: Model comparison on ATP-hard subset.

5 Conclusion

In this paper, we introduce **ReasVQA** (Reasoning-enhanced Video Question Answering), a novel approach that utilizes reasoning processes generated by MLLMs to improve the performance of smaller VideoQA models. Through extensive experiments on three benchmarks, ReasVQA achieves new SOTA results, demonstrating the effectiveness of leveraging reasoning processes in enhancing video understanding. Our in-depth analysis validates each step of ReasVQA, highlighting the value of incorporating refined reasoning data and multi-task learning to enhance the model's capabilities. Results indicate that by guiding models with intermediate reasoning steps, we can significantly boost performance, particularly on complex reasoning tasks. Future work will explore further refinement strategies and integration processes, aiming to extend the approach to other multimodal tasks.

Limitations

Our method is straightforward but relies on the quality of reasoning generated by MLLMs. Although the performance of the VideoQA model improves significantly after refinement, the incorrect or biased reasoning produced by these MLLMs can still negatively affect the overall performance. In future work, we will continue to investigate how to generate higher-quality reasoning or develop better refinement strategies to enable the model to extract more consistent and meaningful information from the reasoning data.

Acknowledgements

We would like to express our sincere gratitude to the anonymous reviewers for their thorough review, insightful comments, and constructive suggestions, which have significantly improved the quality of this manuscript. We also appreciate the valuable guidance and support provided by the AI Data Technology Laboratory of Huawei Noah's Ark Lab. This work paper is supported by the State Key Laboratory of General Artificial Intelligence.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Ziyi Bai, Ruiping Wang, and Xilin Chen. 2024. Glance and focus: Memory prompting for multi-event video question answering. *Advances in Neural Information Processing Systems*, 36.

Tom B Brown. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Shyamal Buch, Cristobal Eyzaguirre, Adrien Gaidon, Jiajun Wu, Li Fei-Fei, and Juan Carlos Niebles. 2022. Revisiting the "Video" in Video-Language Understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. 2024a. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. 2024b. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. 2023. Eva: Exploring the limits of masked visual representation learning at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19358–19369.

Hao Fei, Shengqiong Wu, Wei Ji, Hanwang Zhang, Meishan Zhang, Mong-Li Lee, and Wynne Hsu. 2024. Video-of-thought: Step-by-step video reasoning from perception to cognition. In *Forty-first International Conference on Machine Learning*.

Difei Gao, Luowei Zhou, Lei Ji, Linchao Zhu, Yi Yang, and Mike Zheng Shou. 2023. Mist: Multi-modal iterative spatial-temporal transformer for long-form video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14773–14783.

Jiaxian Guo, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Boyang Li, Dacheng Tao, and Steven Hoi. 2023. From images to textual prompts: Zero-shot visual question answering with frozen large language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10867–10877.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

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.

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

Aisha Urooj Khan, Amir Mazaheri, Niels Da Vitoria Lobo, and Mubarak Shah. 2020. Mmft-bert: Multimodal fusion transformer with bert encodings for visual question answering. *arXiv preprint arXiv:2010.14095*.

Dohwan Ko, Ji Lee, Woo-Young Kang, Byungseok Roh, and Hyunwoo Kim. 2023. Large language models are temporal and causal reasoners for video question answering. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4300–4316.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L Berg, Mohit Bansal, and Jingjing Liu. 2021. Less is more: Clipbert for video-and-language learning via sparse sampling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7331–7341.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2024. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

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.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023b. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. 2023c. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*.

Jianxin Liang, Xiaojun Meng, Yueqian Wang, Chang Liu, Qun Liu, and Dongyan Zhao. 2024. End-to-end video question answering with frame scoring mechanisms and adaptive sampling. *arXiv preprint arXiv:2407.15047*.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let's verify step by step. *arXiv preprint arXiv:2305.20050*.

Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. 2023. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. 2023. Egoschema: A diagnostic benchmark for very long-form video language understanding. *Advances in Neural Information Processing Systems*, 36:46212–46244.

OpenAI. 2024a. Gpt-4o system card. `https://cdn.openai.com/gpt-4o-system-card.pdf`.

OpenAI. 2024b. Gpt-4v(ision) system card. `https://cdn.openai.com/papers/GPTV_System_Card.pdf`.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Junting Pan, Ziyi Lin, Yuying Ge, Xiatian Zhu, Renrui Zhang, Yi Wang, Yu Qiao, and Hongsheng Li. 2023. Retrieving-to-answer: Zero-shot video question answering with frozen large language models. *arXiv preprint arXiv:2306.11732*.

Jongwoo Park, Kanchana Ranasinghe, Kumara Kahatapitiya, Wonjeong Ryoo, Donghyun Kim, and Michael S Ryoo. 2024. Too many frames, not all useful: Efficient strategies for long-form video qa. *arXiv preprint arXiv:2406.09396*.

Devshree Patel, Ratnam Parikh, and Yesha Shastri. 2021. Recent advances in video question answering: A review of datasets and methods. In *Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10–15, 2021, Proceedings, Part II*, pages 339–356. Springer.

David Romero and Thamar Solorio. 2024. Question-instructed visual descriptions for zero-shot video question answering. *arXiv preprint arXiv:2402.10698*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Jonathan Uesato, Nate Kushman, Ramana Kumar, Francis Song, Noah Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. 2022. Solving math word problems with process-and outcome-based feedback. *arXiv preprint arXiv:2211.14275*.

Jinpeng Wang, Yixiao Ge, Rui Yan, Yuying Ge, Kevin Qinghong Lin, Satoshi Tsutsui, Xudong Lin, Guanyu Cai, Jianping Wu, Ying Shan, et al. 2023a. All in one: Exploring unified video-language pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6598–6608.

Shijie Wang, Qi Zhao, Minh Quan Do, Nakul Agarwal, Kwonjoon Lee, and Chen Sun. 2023b. Vamos: Versatile action models for video understanding. *arXiv preprint arXiv:2311.13627*.

Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. 2025. Videoagent: Long-form video understanding with large language model as agent. In *European Conference on Computer Vision*, pages 58–76. Springer.

Xijun Wang, Junbang Liang, Chun-Kai Wang, Kenan Deng, Yu Lou, Ming Lin, and Shan Yang. 2023c. Vlap: Efficient video-language alignment via frame prompting and distilling for video question answering. *arXiv preprint arXiv:2312.08367*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022a. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, et al. 2022b. Internvideo: General video foundation models via generative and discriminative learning. *arXiv preprint arXiv:2212.03191*.

Yuxuan Wang, Yueqian Wang, Pengfei Wu, Jianxin Liang, Dongyan Zhao, and Zilong Zheng. 2024. Lstp: Language-guided spatial-temporal prompt learning for long-form video-text understanding. *arXiv preprint arXiv:2402.16050*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Bo Wu, Shoubin Yu, Zhenfang Chen, Joshua B Tenenbaum, and Chuang Gan. 2021. Star: A benchmark for situated reasoning in real-world videos. In *Thirty-fifth conference on neural information processing systems datasets and benchmarks track (Round 2)*.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. 2021. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9777–9786.

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

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

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 *Proceedings of the 25th ACM international conference on Multimedia*, pages 1645–1653.

Zenan Xu, Xiaojun Meng, Yasheng Wang, Qinliang Su, Zexuan Qiu, Xin Jiang, and Qun Liu. 2023. Learning summary-worthy visual representation for abstractive summarization in video. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, pages 5242–5250. International Joint Conferences on Artificial Intelligence Organization. Main Track.

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.

Zekun Yang, Noa Garcia, Chenhui Chu, Mayu Otani, Yuta Nakashima, and Haruo Takemura. 2020. Bert representations for video question answering. In

Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pages 1556–1565.

Qinghao Ye, Guohai Xu, Ming Yan, Haiyang Xu, Qi Qian, Ji Zhang, and Fei Huang. 2023. Hitea: Hierarchical temporal-aware video-language pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15405–15416.

Shoubin Yu, Jaemin Cho, Prateek Yadav, and Mohit Bansal. 2023. Self-chained image-language model for video localization and question answering. *arXiv preprint arXiv:2305.06988*.

Andy Zeng, Maria Attarian, Brian Ichter, Krzysztof Choromanski, Adrian Wong, Stefan Welker, Federico Tombari, Aveek Purohit, Michael Ryoo, Vikas Sindhwani, et al. 2022. Socratic models: Composing zero-shot multimodal reasoning with language. *arXiv preprint arXiv:2204.00598*.

Ce Zhang, Taixi Lu, Md Mohaiminul Islam, Ziyang Wang, Shoubin Yu, Mohit Bansal, and Gedas Bertasius. 2023a. A simple llm framework for long-range video question-answering. *arXiv preprint arXiv:2312.17235*.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2023b. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*.

Yaoyao Zhong, Junbin Xiao, Wei Ji, Yicong Li, Weihong Deng, and Tat-Seng Chua. 2022. Video question answering: Datasets, algorithms and challenges. In *The 2022 Conference on Empirical Methods in Natural Language Processing*.

A Experiments Setup

A.1 Dataset Details

NExT-QA (Xiao et al., 2021) is a VideoQA benchmark targeting the explanation of video content. The video in NExT-QA primarily encompasses aspects of daily life, social interactions, and outdoor activities, featuring three types of questions: *Temporal* (Tem), *Causal* (Cau), and *Descriptive* (Des). It contains 5.4k videos and about 52K manually annotated question-answer pairs, each QA pair comprises one question and five candidate answers.

STAR (Wu et al., 2021) is oriented towards real-world reasoning scenarios, encompassing four question types, namely *Interaction* (Int), *Sequence* (Seq), *Prediction* (Pre), and *Feasibility* (Fea). STAR contains 22K Situation Video Clips and 60K Situated Questions.

IntentQA (Li et al., 2023a) is a special kind of inference VideoQA dataset that focuses on intent reasoning which studies inference VideoQA beyond factoid VideoQA.

	NExT-QA	STAR	IntentQA
#videos	5.4k	22k	4.3k
#questions	52k	60k	16k

Table 7: Statistics of the datasets we used.

A.2 Experimental Details

Training details We finetune the modality projection layers and LLM(LoRA) (Hu et al., 2021) during training in NVIDIA H800(80GB) GPU × 1. We use AdamW with a cosine learning rate scheduler, whose max learning rate is 3e-5, and a batch size of 8, We train our model within 10 epochs. Our training code is implemented based on LAVIS [1] and transformers [2] libraries, and will be later open sourced.

Baselines We compare our method with two types of baselines: non-LLM and LLM-based models. For non-LLM methods, we use recent SOTA models, including Just Ask (Yang et al., 2021), All-in-One (Wang et al., 2023a) and MIST (Gao et al., 2023), HiTeA (Ye et al., 2023) and Intern-Video (Wang et al., 2022b). For LLM-based models, we use SOTA models such as BLIP-2 (Li et al., 2023b), LLaMA-VQA (Ko et al., 2023), LSTP (Wang et al., 2024), SeVILA (Yu et al., 2023), VidF4 (Liang et al., 2024), ViLA (Wang et al., 2023c), Vamos (Wang et al., 2023b) and MotionEpic (Fei et al., 2024). Among these models, LLaMA-VQA, Vamos, and MotionEpic use 7B-parameter LLM as part of the model.

LLaMA-VQA is built based on LLaMA-7B (Touvron et al., 2023), enabling the model to understand the complex relationships between videos, questions, and answers by constructing multiple auxiliary tasks.

LSTP adopts the BLIP-2 architecture and uses optical flow for frame selection, followed by using LLM to generate answers.

SeVILA relies on a multi-stage training process and is trained on an additional dataset with temporal localization supervision. During the inference, it first utilizes BLIP-2 and LLMs for frame selection and then uses BLIP-2 and LLM again for answer generation.

MotionEpic breaks down the raw intricate video reasoning problem into a chain of simpler subproblems and solves them one by one sequentially.

[1] `https://github.com/salesforce/LAVIS`
[2] `https://github.com/huggingface/transformers`

Vamos (Wang et al., 2023b) generalizes the concept bottleneck model to work with tokens and nonlinear models, which uses hard attention to select a small subset of tokens from the free-form text as inputs to the LLM reasoner.

A.3 Prompts used by MLLMs.

We use prompt engineering to employ MLLMs to generate reasoning processes. We present the prompt template as follows:

```
"""
These frames are uniformly sampled from a video.  Given a question about the video, you should choose the correct answer option from a list of possible answers based on the video content and respond with the option in the format '###Answer: A'. You should also provide a detailed reasoning process explaining why the chosen answer is correct.  Cite specific details from the video frames to support your answer. Explain each step of the reasoning to ensure that the answer is logical and reliable.  ###Question: question, ###Hints: options."""
```

A.4 Reasoning Refinement Details

To address the conflict between process supervision and final result supervision during training, specifically targeting inconsistencies between generated reasoning and true answers, we apply reasoning refinement to the original reasoning generated by MLLMs. Our reasoning refinement phase removes any sentences in the OR that contain the conclusion or predicted answer. Specifically, we identify and remove the sentences with the following fixed patterns:

```
"""
'###Answer: ... '
'**Answer**: ... '
'###Conclusion: ... '
'**Conclusion**: ... '
'###Detailed Explanation ... '
'The correct answer ... '
'Thus, the correct answer is ... '
'Therefore, the correct answer is ... '
'Based on these observations ... '
'Given these observations and the context
... '"""
```

This ensures the reasoning focuses on the process rather than the final conclusion, reducing potential bias.

Setting	STAR				
	Int.	**Seq.**	**Pre.**	**Fea.**	**Tot.**
ReasVQA (3B)	75.9	76.6	67.3	62.0	74.5
w/o. Reasoning	71.4	73.5	63.0	62.7	71.0
w/o. LoRA	68.5	71.1	59.3	58.0	68.3
w/o. both	65.1	69.3	58.7	58.0	66.0

Table 8: Ablation study for each component of ReasVQA.

B More Anylysis

B.1 Ablations

Ablation Study for Each Component of ReasVQA To investigate the effect of each component in our framework, we conduct an extensive ablation study. Table 8 presents an ablation study evaluating the impact of various components on model performance, specifically focusing on the reasoning process (Rea) and Low-Rank Adaptation (LoRA). Our model, ReasVQA (3B), which integrates both reasoning processes and LoRA, achieves a Total accuracy of 74.5. This demonstrates the effectiveness of incorporating these elements, significantly improving performance over the reasoning processes alone.

When the reasoning process is omitted from ReasVQA, the Total accuracy decreases to 71.0. This drop highlights the significant contribution of the reasoning process to the model's overall performance. In the absence of LoRA, the model's Total accuracy is 68.3, showing that while LoRA improves performance, its effect is less pronounced compared to the reasoning process. The removal of both the reasoning process and LoRA results in the lowest Total accuracy of 66.0. This underscores the combined importance of both components, as their exclusion notably impairs the model's effectiveness.

Overall, the ablation study indicates that both the reasoning process and LoRA are crucial for optimal performance. The reasoning process has a more substantial impact, while LoRA also contributes positively but to a lesser degree. The highest performance is achieved when both components are utilized, demonstrating their complementary roles in enhancing the model's capabilities.

Setting	STAR				
	Int.	**Seq.**	**Pre.**	**Fea.**	**Tot.**
InternVL (26B)	66.0	68.9	63.8	62.2	67.0
QA only	71.4	73.5	63.0	62.7	71.0
w. Cap_{brief}	70.7	73.1	64.4	61.4	70.7
w. $\text{Cap}_{detailed}$	72.1	73.7	65.2	63.1	71.7
w. Reasoning	74.1	76.0	62.3	61.8	73.2

Table 9: Influence of different linguistic materials on ReasVQA. All experiments are conducted using a multi-task learning approach, with both α and β set to 0.5.

B.2 Effects of Different Linguistic Materials on ReasVQA

We also evaluate the model's performance when using different types of linguistic information as training data. Table 9 illustrates the results of different types of additional information on the multi-task learning of the model, including settings with QA only, brief captions, detailed captions, and reasoning processes.

The performance of InternVL (26B) is reported with a Total accuracy of 67.0, reflecting the quality of the reasoning processes it produces. InternVL is included not as a baseline for our model, but to provide a reference for the accuracy of the reasoning processes it generates.

The baseline model using only the QA loss function achieves a Total accuracy of 71.0. When brief captions are added, the accuracy across various categories decreases, resulting in a reduced Total accuracy of 70.7. This suggests that brief captions may provide insufficient contextual information, failing to significantly enhance model performance. In contrast, when detailed captions are used, accuracy improves across all categories, raising the Total accuracy to 71.7. This indicates that detailed captions offer richer contextual information, which contributes to better overall model performance.

When our approach is applied, utilizing reasoning processes generated by a multi-modal model to supervise the small model's reasoning generation, significant improvements are observed in Interaction (74.1) and Sequence (76.0) metrics. The Total accuracy further increases to 73.2, which is a 3.2% improvement compared to the QA-only setting (73.2 vs. 71.0). These results highlight that reasoning processes significantly enhance the model's understanding and answering capabilities, providing the most substantial performance gains.

Setting	NExT-QA↑			IntentQA↑		
	75% CR	**CR**	**All**	**75% CR**	**CR**	**All**
Original	75.0	75.9	75.5	75.0	75.7	75.7
Refined	75.9	76.2	76.5	76.3	76.3	76.4

Table 10: The impact of Original Reasoning and Refined Reasoning generated by InternVL(4B) on ReasVQA.

B.3 Impact of Reasoning Generated by InternVL(4B)

To further evaluate the impact of the reasoning generated by the MLLMs (reasoning generator) on our model, we employ InternVL-4B to generate reasoning processes. Following the methods outlined in Sections 3.1 and 4.3, we categorize the generated reasoning into "Correct Reasoning" and "Incorrect Reasoning." It is evident that when using all reasoning data, performance does not surpass that of using only correct reasoning; in fact, it is lower, which highlights the negative impact of incorrect reasoning on model training. After applying refinement, performance improves across all categories, especially with 75% correct reasoning, where the model shows a 1.3% improvement compared to the original reasoning. This indicates that refinement is beneficial, even for correct reasoning.

Importantly, after refinement, using all data (both correct and incorrect) achieves higher performance than using only correct reasoning. This suggests that, while incorrect reasoning may contain erroneous conclusions, it still offers some meaningful information.

Additionally, when combining the results from Figure 3, Table 4, and Table 10, we find that the stronger the MLLM, the higher the accuracy of the reasoning it generates. Moreover, incorporating more correct reasoning consistently enhances model performance.