ReFineVQA: Iterative Refinement of Video Description via Feedback Generation for Video Question Answering

Jeongwan Shin*,1,2,3 Chan Hur*,3 Seongmin Cho[4] Jaeho Choi†,1,2 Hyeyoung Park†,3
[1]DGIST [2]KAIST InnoCORE LLM [3]Kyungpook National University [4]4Genon
https://refinevqa.github.io/

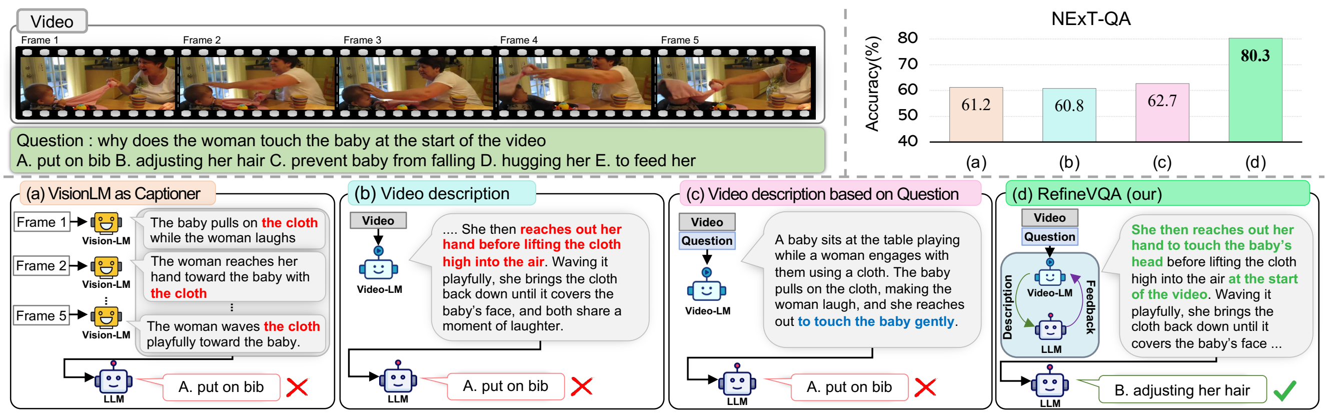


Figure 1. Qualitative and quantitative comparison of answer generation for VideoQA under different LLM inputs: (a) VisionLM-as-Captioner often produces redundant or question-irrelevant captions, (b) Video-level description captures the temporal continuity of actions but may misrepresent question-relevant visual evidence, (c) Question-based video description provides question-specific context yet remains insufficient for deriving the correct answer, and (d) RefineVQA (ours) iteratively refines video descriptions to better align with question-critical evidence, enabling accurate answer generation. The results on the upper right present VideoQA accuracy on the NExT-QA benchmark for methods (a)–(d).

Abstract

Video question answering is a non-trivial task that demands joint understanding of visual contents and linguistic questions as well as temporal reasoning across video frames. Recent agent-based approaches address this by conducting multi-step reasoning with large language models (LLMs) across frame-level captions generated by vision-language models, but encounter limited temporal coherence across frames. A possible direction based on video language models (VideoLMs) directly captures temporal dynamics via video-level descriptions, but often lacks fine-grained visual cues due to a restricted number of input frames and a large dependency on input prompts. To tackle these challenges, we propose RefineVQA, a training-free framework that can easily be plugged into existing VideoLMs with iterative, LLM-guided description refinements. Specifically, the VideoLM produces an initial description, followed by LLM feedback determining whether the description suffices for the question and guiding further visual extraction, which in turn enhances the description quality while preserving temporal context. Plugged into state-of-the-art VideoLMs, ReFineVQA yields consistent gains across diverse benchmarks–NExT-QA, EgoSchema, Video-MME, ActivityNet, and StreamingBench–even with a small external LLM of 3.8B parameters.

1. Introduction

Video Question Answering (VideoQA) [6, 20, 23, 34, 41] has recently achieved remarkable advancements, driven by vision-language models (VisionLM) [18, 46] trained on large-scale image-text pair datasets [19, 38]. Despite such progress, a core challenge remains: given the video and question pairs, VideoQA must jointly integrate visual-linguistic representations, and at the same time, perform temporal reasoning with limited computational resources.

*These authors contributed equally to this work.
†Corresponding authors, equal leading contribution

This, in turn, entails a need for approaches capable of preserving temporal context while retaining fine-grained, question-relevant cues from the video.

One line of approaches [5, 24] to address this employs large language model (LLM)-based multi-step reasoning over VisionLM outputs. As shown in Fig. 1(a), a VisionLM is used for producing frame-level captions from strategically selected frames in a video [21, 24, 30, 32]; these captions, together with a question, are then fed into an LLM [3, 25, 28] to generate answers. However, this frame-level captioning pipeline often yields redundant descriptions (e.g., repeatedly describing "the cloth"), where such repetition induces strong biases, thereby preventing the model from capturing coherent action sequences across frames. Moreover, it remains unclear whether the response generation derives from the visual grounding capability of the VideoLM or from the language priors of the LLM [35].

More recent works [19, 22, 43] directly adopt VideoLM to VideoQA, where they process multiple video frames simultaneously as input and generate answers. However, with a restricted frame budget, VideoLMs typically rely on uniform or heuristic sampling for their input, which risks omitting key information critical for answering the question. Moreover, the end-to-end answering pipeline remains a black box, making it unclear how the predicted answer is grounded in the underlying visual cues [35]. As exemplified in Fig. 1(b), VideoLMs often miss critical frames and consequently fail to capture motion-specific nuances related to the question (e.g., misdescribing the video as "She then reaches out her hand before lifting the cloth high into the air"). Since VideoLMs take multiple frames as input, it becomes crucial to determine which segments should be focused on to ensure that the most relevant visual information is reflected in the reasoning process. Misapplied focus leads the VideoLM to generate an incomplete description, upon which the LLM reasons insufficiently, returning an incorrect answer—this acts as one of the major reasons why video description-based approaches often perform worse even compared to the frame-level caption-based ones.

In response to this issue, we conducted an additional investigation by providing VideoLM with a video as well as *a question* to generate a question-specific description. With this simple approach, as shown in Fig. 1(c), VideoLM could produce more relevant descriptions such as "to touch the baby gently," which better reflects closer evidence toward answering the question, although it still did not lead to the correct answer in this example. Nevertheless, this approach achieved a 1.9% performance improvement over its conventional counterpart (i.e., VideoLM without the question), suggesting that further gains can be obtained by improving the description of VideoLM.

Motivated by this primary investigation, we propose ReFineVQA (shown in Fig. 1(d)), a framework that iteratively refines video descriptions using LLM feedback in the form of follow-up questions targeting missing information for the given query. VideoLM, inherently following auto-regressive text generation, generates video descriptions that differ according to the conditional probability of video frames and text inputs. Considering these characteristics, ReFineVQA uses an LLM to generate feedback to supplement insufficient information in the video description and adjust the frames and questions input to the VideoLM accordingly. Specifically, we first generate an initial video description and conduct a feedback process to determine whether it is sufficient to answer the question. Simultaneously, we generate a feedback question to extract insufficient information and select the most relevant frames accordingly. The input conditions of the VideoLM are modified using both the feedback question and the selected frames to extract additional details, which are then incorporated into the initial description to refine the description. This process is repeated iteratively until the description is sufficient to answer the question.

In our experiments, we validate the efficacy of ReFineVQA by plugging it into the state-of-the-art VideoLM [12, 24, 29, 30, 32, 45] across five representative benchmark datasets in VideoQA [6, 17, 23, 34, 40] : NExTQA, EgoSchema, ActivityNet, Video-MME and StreamingBench. Our results show that significant performance gains can be achieved for all benchmarks with even a single round of feedback-driven refinement. Furthermore, we found that iterating this refinement process can additionally improve the performance. We also design multiple feedback variants within ReFineVQA and provide an extensive ablation to quantify their impact.

In summary, our main contributions are as follows:

- We propose **ReFineVQA**, a training-free, plug-and-play framework that wraps existing VideoLMs with an LLM-guided feedback to produce question-specific, temporally grounded video descriptions, mitigating the intrinsic challenges of current VideoQA approaches.
- The video descriptions refined with our ReFineVQA are expected to be sufficiently informative for given questions, enabling robust and generalized performance even with small-sized LLMs.
- The extensive experiments on five popular benchmarks demonstrate the effectiveness of ReFineVQA, showing consistent performance improvements when applied to state-of-the-art VideoLM.

2. Related Work

2.1. Video Question Answering

The primary goal of the video question-answering task is to enable models to understand visual information in video data and generate appropriate responses [36, 37]. Recent studies have primarily focused on developing models to recognize actions and dynamics within videos [2, 11]. However, most of these efforts fall into the category of simple perceptual-level understanding, such as processing straightforward videos, and are based on an end-to-end approach [1, 13, 16, 47].

With the advent of Transformer-based language models [3, 28], several studies [19, 22, 43] have demonstrated strong performance across various vision–language tasks. In particular, employing a language model as a generative decoder in combination with a image/video encoder has recently emerged as a promising approach for video–language understanding models [12, 19, 22, 43]. For example, Flipped-VQA [9] enhanced the performance of VisionLM by integrating a visual encoder with LLaMA-Adapter [44] and training it to align visual embeddings with text embeddings, enabling the model to effectively understand and process both textual and visual inputs. In this line, VideoLM takes video inputs along with question inputs and performs end-to-end generation, achieving state-of-the-art performance across various VideoQA tasks. Despite these advancements, Xiao et al. [35] questioned whether answers from end-to-end generative approaches genuinely capture video understanding or simply mirror the language model's reasoning. In contrast, RefineVQA leverages a video-understanding-based description generation process, enabling more faithful alignment between video evidence and the final answer.

2.2. Video Understanding with LLM as Agents

A new paradigm in VideoQA has emerged with the use of LLMs as agents, in which a VisionLM generates frame captions to convey visual content to the LLM. Agents-based approach [30] explores the use of natural language text as an intermediate representation between large-scale multimodal models and language models, effectively combining the reasoning ability of LLMs with the image analysis capabilities of VisionLM. Min et al. [24] and Wang et al. [32] achieved significant performance gains by employing a method where a VisionLM first generates image captions, selects frames from the video that are directly relevant to the question, and then combines these captions with the reasoning capabilities [5, 33] of an LLM such as ChatGPT [25]. Recent studies such as Liao et al. [15], Yu et al. [39], and Zhang et al. [42] utilize additional frames to select and aggregate information when the initially provided input is insufficient. However, these methods still rely on frame captions to represent visual content, leaving the interpretation of temporal information largely to the LLM.

Building on this observation, prior work such as Chen et al. [4] attempted to incorporate temporal information by leveraging large models (ChatGPT and BLIP-2 [13]) and performing iterative question-answering across multiple frames. Although the method generates video descriptions through iterative question-answering over multiple frames, it treats each frame independently, thereby failing to capture temporal dependencies grounded in visual continuity. Furthermore, the approach demands considerable computational resources and still requires resource-intensive LLMs to interpret sequential visual information. Even though temporal information plays a critical role, prior studies have not leveraged multi-frame descriptions generated by VideoLMs, leaving a gap in directly modeling temporal dynamics grounded in visual continuity. To overcome this shortcoming, our method dynamically updates video descriptions based on feedback, making them directly relevant to the given question. This ensures that video descriptions are not only temporally coherent but also question-aligned—an essential step for accurate reasoning and efficient answer generation, even when using smaller LLMs.

3. Method

We propose a framework consisting of five interconnected processes as illustrated in Fig. 2. In the proposed RefineVQA, a VideoLM first generates a video description that covers the entire video. This description is then iteratively refined through LLM feedback and question-specific frame selection to supplement insufficient information.

3.1. Preliminaries: Video Question Answering

The objective in VideoQA is to evaluate video-based reasoning within multimodal models. Specifically, the problem involves an input video $V = \{v_1, \dots, v_l\}$ consisting of l frames and a related question Q expressed in natural language. The task is to determine the correct answer A from either the question alone or a set of candidate answers $A \in \mathbf{A}_{\text{cands}}$, where $\mathbf{A}_{\text{cands}}$ is present in closed-set VideoQA [23, 34] and absent in open-ended VideoQA [37]. Finally, our objective is to design a VideoQA model M, which can be formally defined as

$$A = M(V, Q, \mathbf{A}_{\text{cands}}). \tag{1}$$

3.2. Initial Description Generation

Unlike prior works [5, 24] that caption frames independently, our method leverages a VideoLM that accepts multiple frames as input simultaneously. This enables the model to capture dynamic actions and events, leading to video-level descriptions that explicitly incorporate such information. We uniformly sample a set of frames from the video V and feed them into a VideoLM together with a prompt. The model then generates an initial description that captures the overall flow of the video V, formulated as

$$D_{\text{init}} = \text{VideoLM}(V, x_{D_{\text{init}}}), \tag{2}$$

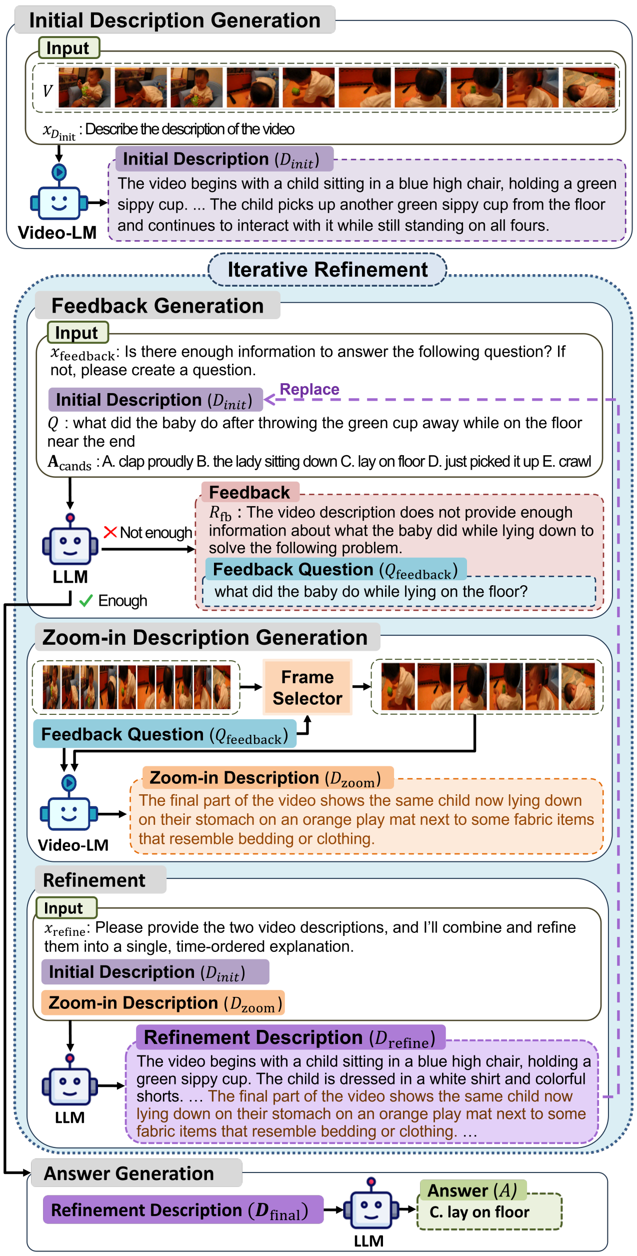


Figure 2. The overall process of the proposed RefineVQA, which iteratively refines video descriptions using LLM feedback for VideoQA. The framework generates the final answer through initial description, feedback generation, zoom-in description generation, iterative refinement, and final answer generation.

where $x_{D_{\text{init}}}$ is a textual prompt guiding the VideoLM, and D_{init} is the generated initial description. Note that D_{init} may not contain the information required to solve the VideoQA task, since the VideoLM has no access to the question at this stage. Therefore, the description must be checked for sufficiency and appropriately supplemented if lacking.

3.3. Iterative Refinement of Video Description

Since the initial description D_{init} may not contain sufficient information to answer the question, we employ an iterative refinement approach to progressively enrich the video description.

Feedback Generation. To accurately extract the required information, it is crucial to infer what is missing from D_{init} with respect to the question. We leverage an LLM that takes D_{init}, Q, and $\mathbf{A}_{\text{cands}}$ as input, together with a feedback prompt x_{feedback} and few-shot examples $\mathcal{E}_{\text{feedback}}$, to verify the available information and generate *Feedback Question* Q_{feedback} when it is insufficient. If the description is adequate, the process moves directly to the answer generation step. Formally, the decision of the LLM is defined as

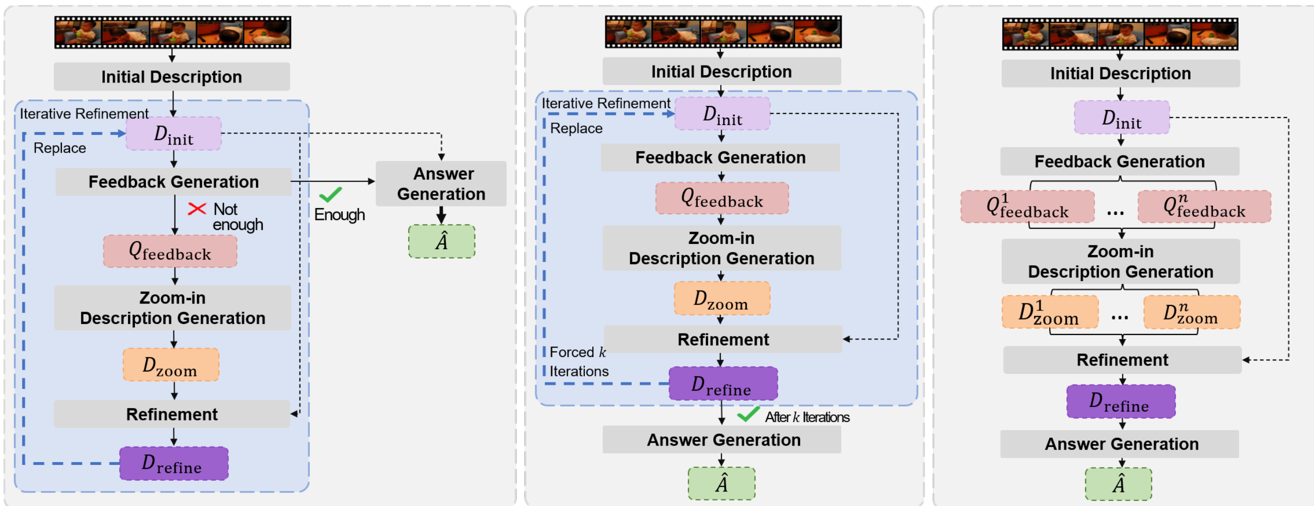
$$\begin{aligned}&\text{LLM}(x_{\text{feedback}},\ \mathcal{E}_{\text{feedback}},\ D_{\text{init}},\ Q,\ \mathbf{A}_{\text{cands}})\\&=\begin{cases}\text{Proceed to Answer Generation}, & \text{if Enough}\\ [R_{\text{feedback}}, Q_{\text{feedback}}]\,, & \text{otherwise}\end{cases}\end{aligned} \tag{3}$$

where the "Enough" case indicates that D_{init} contains sufficient information to generate the final answer A. The "Not Enough" case denotes that additional information is required: the LLM outputs a rationale R_{feedback} describing the missing information in D_{init} and generates a feedback question Q_{feedback}, which are then used to refine the video description for the next iteration. Building on prior works showing that rationales enhance the quality of feedback questions [8, 10, 14], we employ them to make feedback more precise and to steer the refinement of D_{init} toward the most relevant missing information.

Zoom-in Description Generation. The purpose of Q_{feedback} is to extract information that was not present in D_{init}. Therefore, instead of employing uniformly extracted frames at the Initial Description Generation stage, the Zoom-in Description Generation stage employs an adaptive frame selection strategy, which is more suitable for Q_{feedback} to find visual evidence for more detailed information. To achieve this, the frame selector FS is integrated into this stage. The FS calculates a matching score s between the feedback question Q_{feedback} and each frame in $v_t \in V$, expressed as $s_t = \phi(Q_{\text{feedback}}, v_t)$. The matching score s reflects the likelihood that each frame contains visual evidence relevant to Q_{feedback}. All frames are ranked based on their scores, and the top N frames are selected as samples to construct a Q_{feedback} based zoom-in video frames V'.

FS can be implemented in two ways: caption-based and image-based. In the caption-based setting, FS performs text-to-text matching between the feedback question Q_{feedback} and frame captions generated by a VisionLM[1], using a pretrained text encoder[2] inspired by [7]. In the image-

[1]https://huggingface.co/liuhaotian/llava-v1.5-7b
[2]https://huggingface.co/iarfmoose/bert-base-cased-qa-evaluator



(a) **AIoR**: Autonomous Iteration of Refinement (b) **GIoR**: Guided Iteration of Refinement (c) **BpoR**: Batch Processing of Refinement

Figure 3. Variants of iterative feedback refinement strategies: (a) AIoR: feedback generation judges whether the information is sufficient and decides whether to continue iteration. (b) GIoR: refinement is enforced over multiple iterations to improve the video description. (c) BPoR: multiple feedback questions are generated simultaneously and used to update the description in a single iteration.

based setting, FS instead performs image-to-text matching between frames and Q_{feedback} using a pretrained vision-language encoder such as CLIP [26]. Note that the number of comparison frames considered by FS is larger than the number of input frames N allowed in the VideoLM.

After the FS selects V', which contains visual evidence relevant to Q_{feedback}, the selected frames are combined with Q_{feedback} and fed into the VideoLM. This process generates a zoom-in description D_{zoom} that incorporates additional information beyond D_{init}:

$$D_{\text{zoom}} = \text{VideoLM}(V',\ Q_{\text{feedback}}). \tag{4}$$

Although D_{zoom} captures additional information relevant to Q_{feedback}, it lacks temporal grounding, i.e., it does not specify where in the video such information occurs. To address this limitation, we proceed to the next stage, Refinement, which incorporates temporal relations into the description.

Refinement. In this stage, the LLM refines D_{init} by incorporating additional information from D_{zoom}, thereby enhancing its relevance for answering the question. Specifically, the LLM integrates the details from D_{zoom} into the temporal flow of D_{init}, ensuring that the refined description preserves global context while adding localized evidence critical to the query. The refinement process is formally defined as follows:

$$D_{\text{refine}} = \text{LLM}(x_{\text{refine}},\ \mathcal{E}_{\text{refine}},\ D_{\text{init}},\ D_{\text{zoom}}), \tag{5}$$

where x_{refine} is the prompt that instructs refinement, and $\mathcal{E}_{\text{refine}}$ represents the few-shot examples. The refined description D_{refine} is fed back into the feedback generation process, replacing D_{init}, and this procedure is repeated until the information is deemed sufficient. We design several variants of the feedback refinement process, which are detailed in Sec. 3.5.

3.4. Answer Generation

Finally, the LLM performs reasoning to generate an answer based on the fully aggregated video description D_{final}. At this stage, both Q and D_{final} are provided as inputs to the LLM. Through this process, the model demonstrates how it understands the video and derives its response, thereby highlighting the advantage of explainability achieved through interpretable descriptions. The answer generation concludes with the LLM generating the final answer A, which can be expressed as:

$$A = \text{LLM}(x_{\text{answer}},\ \mathcal{E}_{\text{answer}},\ D_{\text{final}},\ Q,\ \mathbf{A}_{\text{cands}}), \tag{6}$$

where x_{answer} is the prompt that instructs question answering, $\mathcal{E}_{\text{answer}}$ provides examples demonstrating the answering process, and D_{final} represents the description that is determined to be sufficiently informative in the feedback process. We emphasize that through each step, the descriptions are refined to enhance the reasoning capabilities of the LLM.

3.5. Variants of Feedback Refinement

In this work, we employ three distinct variants of description refinement through feedback: autonomous iteration of refinement(AIoR), guided iteration of refinement(GIoR), and batch processing of refinement(BPoR). The overall process of each method is illustrated in Figure 3. Detailed ex-

Approach	Model	Video processing	LLM	NExT-QA	EgoSchema	ActivityQA	Video MME	Streaming Bench
(a) VisionLM-LLM	VideoAgent	Frame captions	GPT-4	71.3	54.1	-	-	-
	MoReVQA	Frame captions	GPT-4	69.2	51.7	45.3	-	-
	VideoTree	Frame captions	GPT-4	75.6	61.1	-	-	-
(b) VideoLM (E2E)	Tarsier-7B	Direct video input	✗	71.6	49.9	59.5	-	-
	Llava-OV-0.5B	Direct video input	✗	57.2	26.8	50.5	43.6	43.8
	Llava-OV-7B	Direct video input	✗	79.4	60.1	56.6	61.2	58.4
	Llava-V-7B	Direct video input	✗	$\underline{83.2}$	57.3	56.5	68.6	56.6
(c) VidDesc-LLM	Llava-OV-0.5B	Video Description	Phi-3-mini	58.9 (1.7$\uparrow$)	27.6 (0.8$\uparrow$)	48.5 (2.0$\downarrow$)	27.9 (15.7$\downarrow$)	32.4 (11.4$\downarrow$)
	Llava-OV-7B	Video Description	Phi-3-mini	60.8 (18.6$\downarrow$)	55.3 (4.8$\downarrow$)	54.1 (2.5$\downarrow$)	37.8 (23.4$\downarrow$)	43.6 (14.8$\downarrow$)
	Llava-V-7B	Video Description	Phi-3-mini	62.5 (20.7$\downarrow$)	54.8 (2.5$\downarrow$)	54.6 (1.9$\downarrow$)	38.1 (30.3$\downarrow$)	45.7 (10.9$\downarrow$)
(d) RefineVQA(Our)	Llava-OV-0.5B	Video Description	Phi-3-mini	61.2 (4.0$\uparrow$)	37.4 (10.6$\uparrow$)	55.2 (4.7$\uparrow$)	46.6 (3.0$\uparrow$)	45.2 (1.4$\uparrow$)
	Llava-OV-7B	Video Description	Phi-3-mini	80.3 (0.9$\uparrow$)	**62.4** (2.3$\uparrow$)	**59.4** (2.8$\uparrow$)	$\underline{63.1}$ (1.9$\uparrow$)	**60.2** (1.8$\uparrow$)
	Llava-V-7B	Video Description	Phi-3-mini	**83.7** (0.5$\uparrow$)	$\underline{61.2}$ (3.9$\uparrow$)	$\underline{59.2}$ (2.7$\uparrow$)	**70.7** (2.1$\uparrow$)	$\underline{59.7}$ (3.1$\uparrow$)

Table 1. Performance comparison of RefineVQA with different approaches:(a) VisionLM generating frame caption and LLM, (b) VideoLM end-to-end answer generation (E2E), (c) VideoLM generating description and LLM (VidDesc-LLM) and (d) our proposed RefineVQA (AIoR). Blue parentheses indicate performance gains, while red parentheses indicate drops on the same VideoLM models in (b). The last three columns represent accuracy (%). Bold and underline respectively indicate the best and second best.

amples of each feedback refinement variant are provided in the Supplementary Material.

Autonomous Iteration of Refinement (AIoR). In the AIoR, the feedback generation process determines whether the video description is sufficient to address the question. This decision is reflected in the Boolean output signal iteration stop. Termination following a positive signal mostly results in fewer iterations than the enforced maximum limit. While this accelerates the response by reducing exploration, it carries the risk of premature termination when faced with more complex queries.

Guided Iteration of Refinement (GIoR). GIoR adopts a strategy opposite to that of AIoR, performing refinement over a fixed number of iterations to prevent premature convergence. In GIoR, the process continues for a predefined number of steps k after which the LLM evaluates whether the video description is sufficient to derive the final answer. However, this approach involves additional computational overhead and may lead to redundant or repetitive iterations. This design improves robustness against premature convergence, but it also poses a limitation in that an appropriate value of k must be carefully chosen.

Batch Processing of Refinement (BPoR). BPoR generates multiple questions simultaneously during the feedback generation process, without iterative refinement. After multiple questions are generated, the frame selector extracts relevant frame sets for each question to provide responses. Independent zoom-in descriptions are then created for each question. The merging process for refinement is performed by the LLM within the allowable token limit, enabling the final description to be completed in a single iteration. This design offers advantages in terms of speed, but it does not support dynamic refinement and carries the risk of losing important information during the merging process.

4. Experiments

4.1. Performance on VideoQA

We demonstrate the effectiveness of RefineVQA by comparing it with existing approaches on the NExT-QA [34], EgoSchema [23], Activity-Net [40], Video-MME [6], and StreamingBench [17] benchmarks, as shown in Tab. 1. Notably, Video-MME and StreamingBench consist of long-form videos, which pose greater challenges for temporal reasoning. Further details on the datasets are provided in the Supplementary Material. We evaluate our method against three types of systems: (a) VisionLM generating frame captions, which are then processed by an LLM [24, 30, 32], (b) VideoLM end-to-end answer generation (E2E) [12, 29, 45] and (c) VideoLM generating description and LLM (VidDesc-LLM). Specifically, (a) utilizes an open-source VisionLM as a captioner, generating captions that are then fed into an LLM model (GPT-4) to generate an answer. In contrast, (b) employs a VideoLM to generate answers in an end-to-end manner, while (c) first generates video descriptions using a VideoLM and then feeds them into an LLM to produce the final answer. To ensure a fair evaluation, we employ the same small LLM, Phi-3-mini (3.8B), for both (c) and the proposed method (d).

RefineVQA, which employs a smaller LLM instead of a large-scale LLM like GPT-4, demonstrates significantly superior performance compared to the VisionLM captioner approach across the three benchmarks: NExT-QA, EgoSchema, and Activity-QA. Compared to the VideoTree model (the highest in (a)), RefineVQA outperformed it by 8.1% on NExT-QA when using the Llava-V 7B model and by 1.3% on EgoSchema when using the Llava-OV 7B model. However, since this performance improvement can be attributed to the enhanced video understanding ca-

VideoLM	AIoR	GIoR	BPoR
Llava-OV-0.5B	**61.2**	61.0	59.8
Llava-OV-7B	**80.3**	79.8	76.4
Llava-V-7B	**83.7**	83.4	79.3

Table 2. Comparative experiment on variants of iterative refinement in VideoQA on NExT-QA.

VideoLM	Caption-based	Image-based
Llava-OV-0.5B	**62.3**	61.2
Llava-OV-7B	**80.7**	80.3
Llava-V-7B	**83.9**	83.7

Table 3. Performance comparison of caption-based and image-based frame selection methods for VideoQA on NExT-QA.

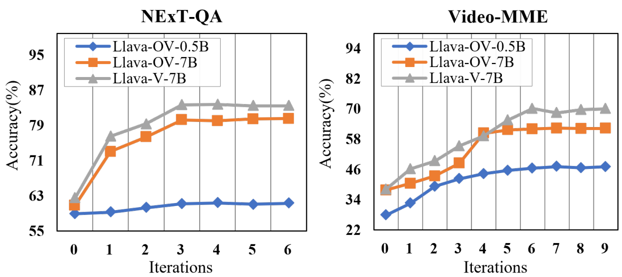


Figure 4. Performance changes depending on the iteration number of the refinement under AIoR setting.

pabilities of VideoLM, we compare RefineVQA with the end-to-end manner of VideoLM (b). In this comparison, the Llava-OV 0.5B model achieved improvements of 4%, 10.6%, 4.7%, 3.0%, and 1.4% across the five benchmarks when applying iterative refinement. Additionally, Llava-OV-7B achieves gains of 2.3% on EgoSchema and 2.8% on ActivityQA, while Llava-V-7B yields even larger improvements of 3.9% on EgoSchema and 3.1% on StreamingBench. These results demonstrate that iterative refinement remains effective, especially for egocentric and long-form video datasets, where limited perspectives or extended temporal contexts often lead to incomplete understanding in end-to-end models.

Note that the proposed framework emphasizes its ability to explain VideoLM's understanding of videos. Despite the impressive performance of VideoLM, the lack of explainability in the end-to-end approach limits the transparency of VideoLM's video understanding process. To verify this, experiment (c), where simple descriptions were generated and answers were inferred using an LLM, showed a significant performance drop. However, our approach compensates for this limitation and demonstrates superior performance compared to the end-to-end approach. The performance drop in experiment (c) was even more pronounced on long-form video datasets such as Video-MME and StreamingBench. In particular, Video-MME exhibited a degradation of over 30%. Nevertheless, our proposed method, which iteratively refines descriptions, outperforms the end-to-end approach, achieving performance gains of 3.0%, 0.9%, and 1.6% across different models. These results demonstrate that our method not only enhances explainability but also achieves superior performance.

4.2. Comparison of Feedback Refinement Variants

In this section, we compare and analyze three variations of feedback refinement on NExT-QA across different VideoLMs, as presented in Tab. 2. AIoR allows up to 3 iterations, averaging 2.1 in practice, whereas GIoR enforces exactly 3 iterations. BPoR generates 3 questions in a single step, corresponding to the three iterations, to ensure experimental consistency. Among the three variations, AIoR achieved the best performance, reflecting the intrinsic advantage of the proposed method in predicting subsequent information based on previously generated content. In addition, this demonstrates the necessity of re-selecting input frames to retrieve the required information during the feedback process. In the case of GIoR, although it does not show a significant difference compared to AIoR, its performance is affected by the generation of redundant information due to forced repetition and the degradation of initial information. BPoR shows faster progression without repetition, but does not achieve significant performance improvements compared to other variations. BPoR progresses quickly without repetition, but does not show significant performance improvements compared to other variations. This limitation can be attributed to its inability to harmoniously integrate information across frames through the LLM. In other words, since BPoR generates multiple questions in a single step, it fails to fully exploit inter-frame dependencies, resulting in less effective performance gains compared to iterative refinement methods such as AIoR.

4.3. Performance Improvements Across Refinement Iterations

Fig. 4 compares the performance of the proposed AIoR method across iterations and demonstrates the performance improvements at each refinement iteration for the NExT-QA and Video-MME datasets. Experimental results showed that utilizing the initial video description (without feedback) led to a significant performance drop compared to the end-to-end VideoLM approach. Both 7B models showed an approximate 20% performance drop compared to the end-to-end generation approach. This raises the possibility that the VideoLM model is optimized to generate answers without a deep perception of the video content. On NExT-QA, a single refinement iteration was sufficient to yield performance gains across all models, with the 7B model achieving

VideoLM	LLM	NExT QA	Ego Schema	Video MME
VAMBA	✗	78.1	–	57.8
	Phi-3(Our)	**80.4**	–	**63.2**
InternVideo	✗	–	63.9	65.1
	Phi-3(Our)	–	**65.3**	**68.3**

Table 4. Evaluation of end-to-end baselines (w/o LLM) and our refinement method using Phi-3-mini on alternative VideoLMs.

VideoLM	LLM	NExT QA	Ego Schema	Video MME
Llava-0.5B	Phi-3	61.2	37.4	55.2
	Qwen2.5	60.9	38.8	54.7
	Mistral	60.8	37.6	55.1
Llava-7B	Phi-3	83.7	61.2	59.2
	Qwen2.5	83.5	61.4	59.3
	Mistral	83.6	61.2	59.1

Table 5. Evaluation demonstrating the generalizability of our approach across various small-scale LLMs.

an improvement of about 12%. In contrast, on Video-MME, which contains longer video content, substantial gains began to appear from the third iteration, where improvements of more than 15% were observed. These results highlight the benefits of iterative refinement, as repeated iterations enable the extraction of necessary information from complex video content. Performance on NExT-QA converged after roughly three iterations, while on Video-MME, convergence was reached after around six iterations. This indicates that shorter videos allow faster convergence, whereas longer and more complex videos require additional iterations, leading to slower convergence in terms of refinement.

4.4. Comparison of Frame Selectors

Tab. 3 compares the performance of two frame selection approaches for identifying informative frames, caption-based and image-based, across different VideoLM models. The results demonstrate that the caption-based frame selector generally achieves consistently higher performance than the image-based approach for all models. Caption-based selection by leveraging textual representations enables better alignment with the question. However, it incurs additional computational costs, resulting in lower efficiency in terms of processing speed. These results suggest that caption-based frame selection offers consistent performance gains, but with a trade-off in computational complexity.

4.5. Generalizability of Proposed Method

To evaluate the robustness of our method across different VideoLMs, we conducted experiments using VAMBA [27] and InternVideo2.5 [31] instead of Llava. As shown in Tab. 4, we evaluate each model under two settings: the end-to-end baseline without an LLM and our refinement method that employs Phi-3-mini as the LLM. Across all configurations, our method consistently improves upon the end-to-end baselines. Specifically, VAMBA shows performance gains of 2.3% on NExT-QA and 5.4% on Video-MME, while InternVideo 2.5 demonstrates improvements of 1.4% on EgoSchema and 3.2% on Video-MME. These results indicate that the proposed framework is generalizable and effective across different underlying VideoLMs.

To further examine its robustness, we also evaluate the framework with different LLMs to assess its generalizability beyond VideoLMs. Tab. 5 demonstrates that providing descriptions refined via feedback yields consistent improvements over the base VideoLM, regardless of the choice of LLM. Notably, all three small-scale LLMs (Phi-3, Qwen2.5, and Mistral) deliver comparable gains, with Phi-3 showing slightly higher accuracy in most cases. These results indicate that the proposed framework is not tied to a specific LLM and can consistently enhance performance across different back-end language models. Furthermore, prior studies suggested that VQA performance is heavily dependent on the LLM's prior knowledge acquired during the LLM's training. [35]. In contrast, our results show that accurate and context-aware responses can be achieved by leveraging video-derived information, demonstrating strong performance even with smaller LLMs rather than relying solely on large knowledge-intensive models.

5. Conclusion

We propose **RefineVQA**, a training-free framework that augments the LLM agent paradigm with video descriptions, extending it beyond frame-level reasoning to exploit richer temporal context. By generating question-based zoom-in descriptions and performing iterative refinement, RefineVQA progressively enhances the alignment of video evidence, enabling more accurate reasoning and answer generation. ReFineVQA demonstrated strong performance across five benchmark datasets. A key strength of our approach lies in its ability to generate question-specific video descriptions that convey temporally grounded visual information to the LLM, enabling strong performance even with a lightweight language model. Additionally, our method is a robust, training-free approach that can be utilized regardless of pipeline components, specific domains, or data modalities. However, it acknowledges limitations in computational cost and complexity due to repeated refinement. Nevertheless, the core principle of iteratively refining video descriptions through feedback to better respond to questions provides the advantage of explainability in the process of understanding the video and generating answers.

Acknowledgments

This work was supported by the InnoCORE program of the Ministry of Science and ICT(No. N10250156), Institute of Information & Communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (No. RS-2025-02219277, AI Star Fellowship Support Project(DGIST)), and the National Research Foundation of Korea (NRF) Grant funded by the Korean Government (MSIT) under Grant NRF-2020R1A2C1010020. The authors would like to thank Seonghyeon Lee for his valuable advice, which greatly helped improve this work.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikoł aj Bińkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, pages 23716–23736. Curran Associates, Inc., 2022. 3

[2] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *Proceedings of the 38th International Conference on Machine Learning*, pages 813–824. PMLR, 2021. 3

[3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, and Dhariwal et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, pages 1877–1901. Curran Associates, Inc., 2020. 2, 3

[4] Jun Chen, Deyao Zhu, Kilichbek Haydarov, Xiang Li, and Mohamed Elhoseiny. Video chatcaptioner: Towards enriched spatiotemporal descriptions. *CoRR*, 2023. 3

[5] Hao Fei, Shengqiong Wu, Wei Ji, Hanwang Zhang, Meishan Zhang, Mong-Li Lee, and Wynne Hsu. Video-of-thought: Step-by-step video reasoning from perception to cognition. In *Forty-first International Conference on Machine Learning*, 2024. 2, 3

[6] Chaoyou Fu, Yuhan Dai, Yondong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*, 2024. 1, 2, 6

[7] Wei Han, Hui Chen, Min-Yen Kan, and Soujanya Poria. Self-adaptive sampling for accurate video question answering on image text models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2522–2534, Mexico City, Mexico, 2024. Association for Computational Linguistics. 4

[8] Zhongjian Hu, Peng Yang, Bing Li, and Fengyuan Liu. Prompting large language models with rationale heuristics for knowledge-based visual question answering. 2024. 4

[9] Dohwan Ko, Ji Lee, Woo-Young Kang, Byungseok Roh, and Hyunwoo Kim. 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, Singapore, 2023. Association for Computational Linguistics. 3

[10] Jaehyeok Lee, Keisuke Sakaguchi, and JinYeong Bak. Self-training meets consistency: Improving llms' reasoning with consistency-driven rationale evaluation. In *Proceedings of the 2025 NAACL Conference (Long Papers)*, pages 10519–10539. Association for Computational Linguistics, 2025. 4

[11] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara Berg. TVQA: Localized, compositional video question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1369–1379, Brussels, Belgium, 2018. Association for Computational Linguistics. 3

[12] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer, 2024. 2, 3, 6

[13] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, pages 19730–19742. PMLR, 2023. 3

[14] K. Li et al. Multimodal rationales for explainable visual question answering. 2024. 4

[15] Ruotong Liao, Max Erler, Huiyu Wang, Guangyao Zhai, Gengyuan Zhang, Yunpu Ma, and Volker Tresp. VideoINSTA: Zero-shot long video understanding via informative spatial-temporal reasoning with LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 6577–6602, Miami, Florida, USA, 2024. Association for Computational Linguistics. 3

[16] Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7082–7092, 2019. 3

[17] Junming Lin, Zheng Fang, Zihao Wan, Fuwen Luo, Chi Chen, Peng Li, Yang Liu, and Maosong Sun. Streamingbench: Assessing the gap for MLLMs to achieve streaming video understanding, 2025. 2, 6

[18] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, pages 34892–34916. Curran Associates, Inc., 2023. 1

[19] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26296–26306, 2024. 1, 2, 3

[20] Huabin Liu, Filip Ilievski, and Cees G. M. Snoek. Commonsense video question answering through video-grounded entailment tree reasoning. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pages 3262–3271, 2025. 1

[21] Ziyu Ma, Chenhui Gou, Hengcan Shi, Bin Sun, Shutao Li, Hamid Rezatofighi, and Jianfei Cai. Drvideo: Document retrieval based long video understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pages 18936–18946, 2025. 2

[22] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. Video-ChatGPT: Towards detailed video understanding via large vision and language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12585–12602, Bangkok, Thailand, 2024. Association for Computational Linguistics. 2, 3

[23] Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. 1, 2, 3, 6

[24] Juhong Min, Shyamal Buch, Arsha Nagrani, Minsu Cho, and Cordelia Schmid. Morevqa: Exploring modular reasoning models for video question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2, 3, 6

[25] OpenAI. Gpt-4 technical report, 2024. 2, 3

[26] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 5

[27] Weiming Ren, Wentao Ma, Huan Yang, Cong Wei, Ge Zhang, and Wenhu Chen. Vamba: Understanding hour-long videos with hybrid mamba-transformers, 2025. 8

[28] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017. 2, 3

[29] Jiawei Wang, Liping Yuan, Yuchen Zhang, and Haomiao Sun. Tarsier: Recipes for training and evaluating large video description models, 2024. arXiv:2407.00634 (v2, last revised 2024-09-24). 2, 6

[30] Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. Videoagent: Long-form video understanding withnbsp;large language model asnbsp;agent. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part LXXX*, page 58–76, Berlin, Heidelberg, 2024. Springer-Verlag. 2, 3, 6

[31] Yi Wang, Xinhao Li, Ziang Yan, Yinan He, Jiashuo Yu, Xiangyu Zeng, Chenting Wang, Changlian Ma, Haian Huang, Jianfei Gao, Min Dou, Kai Chen, Wenhai Wang, Yu Qiao, Yali Wang, and Limin Wang. Internvideo2.5: Empowering video mllms with long and rich context modeling, 2025. 8

[32] Ziyang Wang, Shoubin Yu, Elias Stengel-Eskin, Jaehong Yoon, Feng Cheng, Gedas Bertasius, and Mohit Bansal. Videotree: Adaptive tree-based video representation for LLM reasoning on long videos, 2024. 2, 3, 6

[33] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2024. Curran Associates Inc. 3

[34] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9777–9786, 2021. 1, 2, 3, 6

[35] Junbin Xiao, Angela Yao, Yicong Li, and Tat-Seng Chua. Can i trust your answer? visually grounded video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13204–13214, 2024. 2, 3, 8

[36] Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. Video question answering via gradually refined attention over appearance and motion. In *Proceedings of the 25th ACM International Conference on Multimedia*, page 1645–1653, New York, NY, USA, 2017. Association for Computing Machinery. 2

[37] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Just ask: Learning to answer questions from millions of narrated videos. In *ICCV*, 2021. 2, 3

[38] Qinghao Ye, Guohai Xu, Ming Yan, Haiyang Xu, Qi Qian, Ji Zhang, and Fei Huang. Hitea: Hierarchical temporal-aware video-language pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15405–15416, 2023. 1

[39] Shoubin Yu, Jaemin Cho, Prateek Yadav, and Mohit Bansal. Self-chained image-language model for video localization and question answering. *Advances in Neural Information Processing Systems*, 36:76749–76771, 2023. 3

[40] Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. Activitynet-qa: A dataset for understanding complex web videos via question answering. 33:9127–9134, 2019. 2, 6

[41] Chuanqi Zang, Hanqing Wang, Mingtao Pei, and Wei Liang. Discovering the real association: Multimodal causal reasoning in video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19027–19036, 2023. 1

[42] Ce Zhang, Taixi Lu, Md Mohaiminul Islam, Ziyang Wang, Shoubin Yu, Mohit Bansal, and Gedas Bertasius. A simple LLM framework for long-range video question-answering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21715–21737, Miami, Florida, USA, 2024. Association for Computational Linguistics. 3

[43] Hang Zhang, Xin Li, and Lidong Bing. Video-LLaMA: An instruction-tuned audio-visual language model for video understanding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 543–553, Singapore, 2023. Association for Computational Linguistics. 2, 3

[44] Renrui Zhang, Jiaming Han, Chris Liu, Aojun Zhou, Pan Lu, Yu Qiao, Hongsheng Li, and Peng Gao. LLaMA-adapter:

Efficient fine-tuning of large language models with zero-initialized attention. In *The Twelfth International Conference on Learning Representations*, 2024. 3

[45] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data, 2024. 2, 6

[46] Yutong Zhou and Nobutaka Shimada. Vision + language applications: A survey. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 826–842, 2023. 1

[47] Mohammadreza Zolfaghari, Kamaljeet Singh, and Thomas Brox. Eco: Efficient convolutional network for online video understanding. In *Computer Vision – ECCV 2018: 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part II*, page 713–730, Berlin, Heidelberg, 2018. Springer-Verlag. 3