

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

Supplementary Material

A. Datasets

NExT-QA [8] is a benchmark dataset designed to evaluate the ability of VideoQA systems to effectively answer questions across three types: causal (C), temporal (T), and descriptive (D), given both a video and a question. We follow the same multiple-choice (MC) setting reported in the prior single-stage modular reasoning work by Surís et al. [7]. In this setting, each video clip (with an average length of 43 seconds) contains one question and five candidate answers. For the experiment, we use 8,564 test video-question pairs.

EgoSchema [6] is a recently developed dataset designed to evaluate long video understanding, based on the Ego4D [2] benchmark with multiple-choice (MC) questions for long egocentric videos (180 seconds). The hidden test set consists of 5,000 videos and is evaluated through a server (Kaggle), with 500 videos publicly available for validation. We report accuracy results on the full 5k test set to compare with state-of-the-art VideoLM.

ActivityNet-QA [9] contains 5,800 videos, each accompanied by 10 annotated question-answer pairs designed to evaluate a model’s understanding of actions, objects, locations, and events. ActivityNet-QA features long videos (average length of 180 seconds, similar to EgoSchema) and uses an open-ended format. Following previous studies [3, 5, 10], we report test set results using GPT-based evaluation.

Video-MME [1] To evaluate the video understanding capabilities of Multi-modal Large Language Models (MLLMs), we utilize Video-MME, a recent benchmark spanning 900 videos (254 hours) across 6 visual domains and 30 subfields, with durations ranging from 11 seconds to 1 hour. It includes 2,700 expert-annotated QA pairs and supports multi-modal inputs such as frames, subtitles, and audio. Video-MME enables evaluation of both image- and video-based MLLMs, with prior studies showing that subtitles and audio improve performance, while longer videos remain challenging.

StreamingBench [4] To assess real-time video understanding in Multi-modal Large Language Models (MLLMs), we adopt StreamingBench, the first benchmark designed for streaming video comprehension. Unlike offline settings, it simulates real-time scenarios where models incrementally process video and audio inputs. StreamingBench evaluates three key capabilities: real-time visual recognition, synchronized multi-modal integration, and contextual reasoning over streaming content. It includes 900 videos and 4,500 human-annotated QA pairs, with five timestamped questions per video to assess understanding at different temporal stages.

B. Implementation Details

To evaluate the excellence of the proposed pipeline, various state-of-the-art VideoLM of different sizes and types are utilized. For VideoLM, we use the Llava-OneVision (0.5B, 7B) and Llava-Video (7B) models to generate both the Initial Description and Zoom-in Description. Due to resource limitations, we were unable to apply the proposed method to the 70B model. We use Phi-3-mini(3.8B)¹ to provide feedback, improvements, and generate answers. The frame selector model uses two variations, and both models uniformly select 200 frames from the entire video. The image-based model employs clip encoder² to measure the similarity between the image and the question. The caption-based model first generates a caption for each frame using a VisionLM. The similarity between the frame captions and the question is then measured using a text embedding model³.

¹<https://huggingface.co/microsoft/Phi-3-mini-128k-instruct>

²<https://huggingface.co/openai/clip-vit-large-patch14>

³https://huggingface.co/Marqo/dunzhang-stella_en_400M_v5

C. Inference Time Analysis

To complement the iteration statistics reported in Figure 4 of the main paper, we analyze the inference-time cost of our iterative refinement framework compared to direct answer generation in standard VideoLLMs. In our environment, direct prediction requires approximately 1 to 3 seconds depending on output length, whereas our method involves generating a long-form video description followed by iterative LLM-based refinement. Due to the autoregressive nature of text generation, each iteration takes on average around 70 seconds, as longer outputs naturally result in slower inference.

Despite the increased computational cost, the iterative procedure provides clear advantages for complex video reasoning tasks that require integrating information across multiple temporal segments. By progressively extracting and refining question-relevant visual evidence, our framework produces more accurate and better-grounded answers while offering improved interpretability. The explicit video descriptions generated during refinement reveal the model’s reasoning process, aligning with recent trends that prioritize transparency and reasoning quality over direct prediction.

D. Discussion on Sliding-Window Approaches

The key limitation we address is not merely the restricted number of frames processed by Video-LLMs, but the need to identify and incorporate *question-relevant key frames* into the video description. Our approach focuses on selecting visual evidence that is specifically relevant to the given question and integrating it into a refined textual representation, which is essential for accurate reasoning and answering. To examine the effectiveness of key-frame identification, we report results on NExT-QA using frames manually chosen by humans as an optimal reference. As shown in Table 1, our method achieves comparable or slightly better performance than optimal frame selection.

Model	Uniform	Frame Selector (CLIP)	Optimal
LLaVA-OV 0.5B	0.572	0.598	0.613
LLaVA-OV 7B	0.794	0.806	0.811
LLaVA-Video 7B	0.832	0.824	0.835

Table 1. Comparison of uniform sampling, CLIP-based frame selection, and human-selected optimal frames on NExT-QA.

Simply increasing the number of input frames through sliding-window sampling is limited by memory and context-length constraints, especially for long videos such as those in Video-MME (often around one hour). Models typically sample every 3–5 seconds, resulting in the loss of important temporal cues. Our approach provides a model-agnostic mechanism that enables effective reasoning across models of any scale while remaining robust under resource constraints.

E. Detailed Prompt Construction and System I/O

This section provides detailed prompts for each process of RefineVQA, along with the corresponding inputs, outputs, and a few examples. Additionally, it covers all variations of Feedback Refinement in detail, including Autonomous Iteration of Refinement(AIoR), Guided Iteration of Refinement(GIoR), and Batch Processing of Refinement(BPoR).

E.1. Autonomous Iteration of Refinement(AIoR)

The AIoR variation consists of five processes: 1) Initial Description, 2) Feedback Generation, 3) Zoom-in Description Generation, 4) Refinement, 5) Answer Generation. In process 2) Feedback Generation, the system checks whether the Initial Description contains sufficient information to answer the question. If it is insufficient, the process continues to process 4) Refinement, where the Initial Description is replaced with D_{Refine} . And then, Feedback Generation then re-evaluates whether the information is sufficient. If it remains insufficient, processes 2) to 4) are repeated iteratively to refine the description.

1) Initial Description Generation: Since Initial Description Generation is performed using the same prompt across all variants, this process remains identical in other variants as well.

Initial Description Generation

► **Input:**

<Input Video>  </Input Video>

<Task Instruction> Describe the overall content of the video </Task Instruction>

► **Output:**

<Initial Description Output> The video begins with a child sitting in a blue high chair, holding a green sippy cup. ... The child picks up another green sippy cup from the floor and continues to interact with it while still standing on all fours. </Initial Description Output>

2) Feedback Generation: In AIoR, two types of few-shot examples, Enough and Not Enough, are used for feedback generation

2-1) Few-shot Example: The Not Enough few-shot example is an example that generates a feedback question along with feedback indicating that the information in the video description is insufficient.

Few-shot example(Not Enough)

Descriptions: The video begins with a red car driving down a busy city street, stopping at a traffic light. ... The car then makes a sharp turn at an intersection and continues moving along the road.

Question: What did the car do after making the sharp turn near the end?

A. Stopped moving B. Parked nearby C. Drove straight D. Turned again E. Slowed down

Feedback: The video description does not provide enough information about what the car did after making the sharp turn to accurately answer the question.

Feedback Question: What did the car do after making the sharp turn?

The Enough few-shot example is an example where the video description contains sufficient information, outputting 'Enough' after the feedback question.

Few-shot example(Enough)

Descriptions: The video shows a golden retriever playing in a park. It chases a red ball, catches it in its mouth, and happily runs back to its owner. The owner then throws the ball again, and the dog sprints after it. Near the end, the dog catches the ball mid-air and wags its tail excitedly before returning to the owner.

Question: What did the dog do after catching the ball near the end?

A. Wagged tail B. Sat down C. Barked loudly D. Ran away E. Dropped ball

Feedback: The video description provides enough information to determine that the dog wagged its tail after catching the ball.

Feedback Question: Enough

2-2) Not Enough Information Feedback generation I/O:

Here, we examine the Not Enough system I/O example. The system receives both the Enough and Not Enough examples described above, along with instructions, the previously generated Initial Description, and the Question. The system then outputs the Feedback and Feedback Question.

Feedback generation

► Input:

Is there enough information to answer the following question? If not, please create a question.

<Few-shot example>

<Initial Description Output>

Descriptions: The video begins with a child sitting in a blue high chair, holding a green sippy cup. ... The child picks up another green sippy cup from the floor and continues to interact with it while still standing on all fours.

</Initial Description Output>

<Question>

what did the baby do after throwing the green cup away while on the floor near the end

A. clap proudly B. the lady sitting down C. lay on floor D. just picked it up E. crawl

</Question>

► Output:

Feed back: The video description does not provide enough information about what the baby did while lying down to solve the following problem.

Feed back question: what did the baby do while lying on the floor?

2-3) Enough Information Feedback generation I/O: Here, we demonstrate a case where the Video Description is refined to provide sufficient information. If the information is deemed sufficient, the system outputs 'Enough' after the feedback question.

Feedback generation

► Input:

<Task Instruction> Is there enough information to answer the following question? If not, please create a question.

</Task Instruction>

<Few-shot example>

<Refinement Description> The video begins with a child sitting in a blue high chair, holding a green sippy cup. The child is dressed in a white shirt and colorful shorts. ... The child picks up another green sippy cup from the floor and continues to interact with it while still standing on all fours. ... **The final part of the video shows the same child now lying down on their stomach on an orange play mat next to some fabric items that resemble bedding or clothing.**

</Refinement Description>

<Question> what did the baby do after throwing the green cup away while on the floor near the end

A. clap proudly B. the lady sitting down C. lay on floor D. just picked it up E. crawl </Question>

► Output:

Feed back: The video description does not provide enough information about what the baby did while lying down to solve the following problem.

Feedback Question: Enough

3) Zoom-in Description Generation: This process selects new frames through the Frame Selector based on the previously generated Feedback Question. The selected frames and the Feedback Question are then input into the LLM to generate the Zoom-in Description.

Zoom-in Description Generation

► **Input:**

<Input Query Based Frame Selected Video Frame>



</Input Query Based Frame Selected Video Frame>

<Feedback Question>

what did the baby do while lying on the floor?

</Feedback Question>

► **Output:**

<Zoom-In Description Output>

The final part of the video shows the same child now lying down on their stomach on an orange play mat next to some fabric items that resemble bedding or clothing.

</Zoom-In Description Output>

4) Refinement: Here, we present the few-shot examples used for the refinement process and the input-output examples of the LLM refining the Initial Description and Zoom-in Description.

4-1) Refinement Few-Shot Example: The Refinement Few-Shot Example demonstrates how the Initial Video Description and Zoom-in Description have been refined.

Refinement Few-Shot Example

Video Description:The video begins with a red car driving down a busy city street, stopping at a traffic light. ... The car then makes a sharp turn at an intersection and continues moving along the road.

Zoom-In Description:In the final part of the video, the red car is seen pulling into a parking lot near a convenience store. The car comes to a complete stop, and its headlights turn off as the driver exits the vehicle.

Refinement:The video begins with a red car driving down a busy city street, stopping at a traffic light. The car has a sleek design with tinted windows. ... The car then makes a sharp turn at an intersection and continues moving along the road. ... In the final part of the video, the red car is seen pulling into a parking lot near a convenience store. The car comes to a complete stop, and its headlights turn off as the driver exits the vehicle.

4-2) Refinement I/O Example: Refinement takes Task Instruction, Few-Shot Example, Initial Description Output, and Zoom-In Description Output as inputs to the LLM and generates the Refinement Description.

Refinement

► **Input:**

<Task Instruction>

Read the two video descriptions, then combine and refine them into a single, well-structured description in chronological order.

</Task Instruction>

<Few-Shot Example>

<Initial Description Output>

Video Description: The video begins with a child sitting in a blue high chair, holding a green sippy cup. ... The child picks up another green sippy cup from the floor and continues to interact with it while still standing on all fours.

</Initial Description Output>

<Zoom-In Description Output>

Zoom-In Description: [The final part of the video shows the same child now lying down on their stomach on an orange play mat next to some fabric items that resemble bedding or clothing.](#)

</Zoom-In Description Output>

► **Output:**

Refinement: The video begins with a child sitting in a blue high chair, holding a green sippy cup. The child is dressed in a white shirt and colorful shorts. ... The child picks up another green sippy cup from the floor and continues to interact with it while still standing on all fours. ... [The final part of the video shows the same child now lying down on their stomach on an orange play mat next to some fabric items that resemble bedding or clothing.](#)

5) Answer Generation: Here, if the Feedback Generation process determines that the information is Enough, the final step is to generate the Answer. This step is used consistently across other variants as well.

5-1) Answer Generation Few-Shot Example: The Few-Shot examples for Answer Generation include the Video Description, Question, and Answer.

Few-Shot Example

Video Description: The video showcases a lively scene at a playground on a sunny day. It begins with children playing on swings, one of which is attached to a central swing mechanism. The swings are blue and red, and the central mechanism has weights hanging from chains. The background features green grass, trees, and other playground equipment like umbrellas. ...

Question: how did the girl in blue moved the swing?

Options: A. running her feet on ground B. with its paws C. holds the baby close to her D. push him E. using headset

Answer: A. running her feet on ground

5-2) Answer Generation I/O: The LLM takes the Task Instruction, Few-Shot Example, Refinement Description, and Question as inputs to solve the VideoQA and generates the final Answer.

Answer Generation

► **Input:**

<Task Instruction>

You are a helpful assistant that answers questions about the description. You are given a Context and you choose one of five options to answer. You must select only one option as the answer and you should not provide any information that is not in the provided options.

</Task Instruction>

<Few-Shot Example>

<Refinement Description>

The video begins with a child sitting in a blue high chair, holding a green sippy cup. The child is dressed in a white shirt and colorful shorts. ... The child picks up another green sippy cup from the floor and continues to interact with it while still standing on all fours. ... **The final part of the video shows the same child now lying down on their stomach on an orange play mat next to some fabric items that resemble bedding or clothing.** </Refinement Description>

<Question>

what did the baby do after throwing the green cup away while on the floor near the end

A. clap proudly B. the lady sitting down C. lay on floor D. just picked it up E. crawl

</Question>

► **Output:**

C. lay on floor

E.2. Guided Iteration of Refinement(GIoR)

GIoR follows the same process as AIoR, with the only difference being that step 2) Feedback Generation and step 4) Refinement are repeated for a predefined number of iterations. Specifically, the Few-Shot Example that allows selecting Enough Information in step 2) Feedback Generation is excluded, ensuring that only the Feedback Question is generated.

E.3. Batch Processing of Refinement(BPoR)

BPoR follows the same process structure as AIoR. (1) Initial Description and (5) Answer Generation remain the same, while (2) Feedback Generation to (4) Refinement is performed only once and is not iterated.

1) Feedback Generation: In BPoR, multiple feedback questions are generated at once. Additionally, it does not determine whether it is "sufficient" or "insufficient," nor does it perform iteration.

1-1) Feedback Generation Few-Shot Example:

A few-shot example of Feedback Generation in BPoR includes Descriptions, a Question, Feedback, and multiple Feedback Questions.

Few-Shot Example

Descriptions: The video begins with a red car driving down a busy city street, stopping at a trafficlight. . . . The car then makes a sharp turn at an intersection and continues moving along the road.

Question: What did the car do after making the sharp turn near the end?

A. Stopped moving B. Parked nearby C. Drove straight D. Turned again E. Slowed down

Feedback: The video description does not provide enough information about what the car did after making the sharp turn to accurately answer the question.

Feedback Question: What did the car do after making the sharp turn?

Did the car park after making the sharp turn?

Did the car stop, park, keep going straight, turn again, or slow down?

1-2) Feedback Generation I/O: The LLM generates multiple Feedback Questions using Task Instruction, Few-shot Examples, Initial Description Output, and the given question as inputs.

Feedback Generation

► **Input:**

<Task Instruction> Is there enough information to answer the following question? If not, please create a question.

</Task Instruction>

<Few-shot example>

<Initial Description Output>

Descriptions: The video begins with a child sitting in a blue high chair, holding a green sippy cup. ... The child picks up another green sippy cup from the floor and continues to interact with it while still standing on all fours.

</Initial Description Output>

<Question>

what did the baby do after throwing the green cup away while on the floor near the end

A. clap proudly B. the lady sitting down C. lay on floor D. just picked it up E. crawl

</Question>

► **Output:**

Feed back: The video description does not provide enough information about what the baby did while lying down to solve the following problem.

Feed back question: what did the baby do while lying on the floor?

Did the baby clap, interact with someone, stay lying down, pick up the cup, or crawl?

What action did the baby take while lying on the floor?

2) Zoom-in Description Generation: The Zoom-in Description Generation process generates each description independently from multiple Feedback Questions. Since its I/O follows the same format as AIoR, it is omitted.

3) Refinement: Here, we refine the initial description and zoom-in descriptions in BPoR. We present a few-shot examples and input-output examples used in the refinement process.

3-1) Refinement Few-Shot Example: The Refinement Few-Shot Example shows how the Initial Video Description and multiple Zoom-in Descriptions have been refined.

Few-Shot Example

Video Description: The video begins with a red car driving down a busy city street, stopping at a traffic light. ... The car then makes a sharp turn at an intersection and continues moving along the road.

Zoom-In Description 1: In the final part of the video, the red car is seen pulling into a parking lot near a convenience store. The car comes to a complete stop, and its headlights turn off as the driver exits the vehicle.

Zoom-In Description 2: after making the sharp turn, the car continued for a short distance before stopping completely and parking in a designated parking lot.

Zoom-In Description 3: Following the sharp turn, the car gradually slowed down, came to a full stop, and parked in a nearby parking lot instead of continuing straight or making another turn.

Refinement: The video begins with a red car driving down a busy city street, stopping at a traffic light. After the light changes, the car moves forward and makes a sharp turn at an intersection. Instead of continuing straight or making another turn, the car gradually slows down and pulls into a nearby parking lot near a convenience store. It comes to a complete stop, and its headlights turn off as the driver exits the vehicle.

3-2) Refinement I/O: Refinement takes Task Instruction, Few-Shot Example, Initial Description Output, and multiple Zoom-In Description Outputs as inputs to the LLM and generates the Refinement Description.

Refinement

► **Input:**
 <Task Instruction> If you provide multiple video descriptions, I will combine and refine them into a single, coherent description in chronological order. </Task Instruction>
 <Few-Shot Example>
 <Initial Description Output>
 Video Description: The video begins with a child sitting in a blue high chair, holding a green sippy cup. . . . The child picks up another green sippy cup from the floor and continues to interact with it while still standing on all fours.
 </Initial Description Output>
 <Zoom-In Description Output>
 Zoom-In Description 1: The final part of the video shows the same child now lying down on their stomach on an orange play mat next to some fabric items that resemble bedding or clothing.
 Zoom-In Description 2: At the end, the baby lies down on the floor and remains still without making any further movements.
 Zoom-In Description 3: The baby appears to be lying down for a while without showing any noticeable actions or movements.
 </Zoom-In Description Output>
 ► **Output:**
 Refinement: The video begins with a child sitting in a blue high chair, holding a green sippy cup. After some interaction, the child picks up another green sippy cup from the floor and continues engaging with it while standing on all fours. Toward the end of the video, the child is seen lying on their stomach on an orange play mat next to some fabric items that resemble bedding or clothing. The baby remains still for a while, showing no noticeable actions or movements.

F. Qualitative analysis

This section presents additional qualitative analysis of the proposed method, RefineVQA, and compares it with VideoLM end-to-end generation, VidDesc-LLM and RefineVQA(AIoR)

F.1. End-to-End Generation with VideoLM

In the example of “How did the girl show excitement near the middle of the video?”, the VideoLM outputs A. pick up toy, while the correct answer is D. jumps. Visual inspection shows that the toy is salient across many frames, whereas the jump occurs briefly around the middle. The model appears to over-weight persistent object cues (the toy) and under-weight transient motion cues (a short hop), leading to an object-centric prediction rather than an action-centric one.

VideoLM End to End I/O

► **Input:**



how did the girl show excitement near the middle of the video
 A. pick up toy B. put finger in mouth C. standing D. jumps E. walking

► **Output:**
 A. pick up toy

F.2. Based on Video Description and LLM

The example illustrates two critical limitations of video-level descriptions. First, while they capture overall context and general activities (e.g., a child playing with toys in a home setting), they omit fine-grained actions that are central to the question, such as “jump.” Second, they lack temporal grounding, failing to specify what action occurs near the middle of the video, which is precisely what the question asks about. Consequently, the LLM answer generation is restricted to generic behaviors mentioned in the description, leading to the incorrect prediction “pick up toy” rather than the correct action. This demonstrates how background-focused narratives overlook question-relevant evidence and temporal cues, thereby reducing answer accuracy. Such shortcomings underscore the importance of RefineVQA, which employs question-guided zoom-in descriptions and iterative refinement to recover missing details, establish temporal grounding, and better align video evidence with the given question.

VideoLM Description Generation I/O

► **Input:**



Describe the overall description of the video.

► **Output<Video Description>:**

The video features a young child, dressed in a black and white sweater with a star pattern, interacting with various toys. The child is seen playing with a red and blue toy rocket ship on the table, which has orange wheels. The background includes a bookshelf filled with books and DVDs, creating an environment that appears to be both playful and cozy. As the video progresses, the child moves around the room while holding different toys such as another red and blue toy rocket ship in their hands. They also interact with other objects like teddy bears and stuffed animals placed around them. Towards the end of the video, one of these toys is transformed into an action figure by being turned into this character using its own hands. This transformation adds an element of imaginative play to their interaction. Throughout all frames in this sequence, there are no visible titles or subtitles within these images; they are purely focused on capturing moments of childhood playtime in a home setting.

LLM Answer Generation I/O

► **Input:**

<Video Description>

how did the girl show excitement near the middle of the video

A. pick up toy B. put finger in mouth C. standing D. jumps E. walking

► **Answer Output:**

A. pick up toy

F.3. Autonomous Iteration of Refinement(AIoR)

In this example we begin from feedback generation—not from a fresh video description—since the global description has already been shown in D.2. The question is: “How did the girl show excitement near the middle of the video?”

Stage 1. The LLM first points out that the available description lacks explicit action evidence and requests clarification about movements. The subsequent zoom-in yields motion-centric content, and the refined description is updated to: “At one point, when the child becomes excited, their body seems to shake, showing their excitement through movement.” This introduces temporal anchoring and foregrounds motion, but the action remains unnamed.

Stage-1: LLM Feedback Generation I/O

► **Input:**

Is there enough information to answer the following question? If not, please create a question. <Video Description> how did the girl show excitement near the middle of the video
A. pick up toy B. put finger in mouth C. standing D. jumps E. walking

► **Output:**

Feedback:The Video Description does not explicitly mention how the girl shows excitement, nor does it describe specific actions like jumping, putting a finger in the mouth, standing, or walking.
Feedback Question:Did the girl display any movements when she was excited?

Stage-1: VideoLM Zoom-in Description Generation I/O

► **Input:**



Did the girl display any movements when she was excited?

► **Output:**

Yes, when the girl was excited, her body seemed to shake, showing her excitement through movement.

Stage-1: LLM Refinement I/O

► **Input:**

Read the two video descriptions, then combine and refine them into a single, well-structured description in chronological order.

<Video Description>

Yes, when the girl was excited, her body seemed to shake, showing her excitement through movement.

► **Output<Refinement Description(Stage-1)>:**

The video features a young child, dressed in a black and white sweater with a star pattern, interacting with various toys. The child is seen playing with a red and blue toy rocket ship on the table, which has orange wheels. The background includes a bookshelf filled with books and DVDs, creating an environment that appears to be both playful and cozy. As the video progresses, the child moves around the room while holding different toys, such as another red and blue toy rocket ship in their hands. They also interact with other objects like teddy bears and stuffed animals placed around them. At one point, when the child becomes excited, their body seems to shake, showing their excitement through movement. Towards the end of the video, one of these toys is transformed into an action figure, as the child uses their own hands to turn it into this character. This transformation adds an element of imaginative play to their interaction. Throughout all frames in this sequence, there are no visible titles or subtitles within these images; they are purely focused on capturing moments of childhood playtime in a home setting.

Stage-2: LLM Feedback Generation I/O

► Input:

Is there enough information to answer the following question? If not, please create a question. <Refinement Description(Stage-1)>

how did the girl show excitement near the middle of the video

A. pick up toy B. put finger in mouth C. standing D. jumps E. walking

► Output:

Feedback: The video description mentions that the girl expressed excitement, but it only states that her body was shaking. It does not provide details about specific actions such as jumping or standing.

Feedback Question: Did the girl exhibit specific actions like jumping or standing while expressing excitement?

Stage 2. Prompted again, the LLM asks for specific actions (e.g., jumping or standing). The zoom-in then recovers the decisive cue and the refined description becomes: “At one point, when the child becomes excited, their body seems to shake, showing their excitement through movement. Additionally, while expressing excitement, the girl jumped while looking at the red and blue toys.” This edit resolves the ambiguity by explicitly naming jumped and retaining the relevant context, counteracting the earlier object-salience bias toward “pick up toy.”

Stage-2: VideoLM Zoom-in Description Generation I/O

► Input:



Did the girl exhibit specific actions like jumping or standing while expressing excitement?

► Output:

Yes, while expressing excitement, the girl jumped while looking at the red and blue toys.

Stage-2: LLM Refinement I/O

► Input:

Read the two video descriptions, then combine and refine them into a single, well-structured description in chronological order.

<Refinement Description(Stage-1)>

Yes, while expressing excitement, the girl jumped while looking at the red and blue toys.

► Output<Refinement Description(Stage-2)>:

The video features a young child, dressed in a black and white sweater with a star pattern, interacting with various toys. The child is seen playing with a red and blue toy rocket ship on the table, which has orange wheels. The background includes a bookshelf filled with books and DVDs, creating an environment that appears to be both playful and cozy. As the video progresses, the child moves around the room while holding different toys, such as another red and blue toy rocket ship in their hands. They also interact with other objects like teddy bears and stuffed animals placed around them. At one point, when the child becomes excited, their body seems to shake, showing their excitement through movement. Additionally, while expressing excitement, the girl jumped while looking at the red and blue toys. Towards the end of the video, one of these toys is transformed into an action figure, as the child uses their own hands to turn it into this character. This transformation adds an element of imaginative play to their interaction. Throughout all frames in this sequence, there are no visible titles or subtitles within these images; they are purely focused on capturing moments of childhood playtime in a home setting.

Stage-3: LLM Feedback Generation I/O

► **Input:**

Is there enough information to answer the following question? If not, please create a question. <Refinement Description(Stage-2)>

how did the girl show excitement near the middle of the video

A. pick up toy B. put finger in mouth C. standing D. jumps E. walking

► **Output:**

Feedback:The video description mentions that the girl expressed excitement and includes specific information about her jumping.

Feedback Question:Enough

Stage 3. With action and timing now explicitly grounded, the LLM marks the evidence as sufficient and produces the correct answer: D. jumps. Overall, this sequence illustrates how Autonomous Iteration of Refinement (AIoR) transforms a verbose, object-centric summary into action-anchored, temporally grounded evidence through targeted feedback questions and question-conditioned zoom-in descriptions—precisely the information needed to answer the query correctly.

Stage-3: LLM Answer Generation I/O

► **Input:**

<Refinement Description(Stage-2)>

how did the girl show excitement near the middle of the video

A. pick up toy B. put finger in mouth C. standing D. jumps E. walking

► **Answer Output:**

D. jumps

References

- [1] 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] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Zhongcong Xu, Chen Zhao, Siddhant Bansal, Dhruv Batra, Vincent Cartillier, Sean Crane, Tien Do, Morrie Doulaty, Akshay Erapalli, Christoph Feichtenhofer, Adriano Fragomeni, Qichen Fu, Abrham Gebrelesiasie, Cristina González, James Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie Khoo, Jáchym Kolář, Satwik Kottur, Anurag Kumar, Federico Landini, Chao Li, Yanghao Li, Zhenqiang Li, Karttikeya Mangalam, Raghava Modhugu, Jonathan Munro, Tullie Murrell, Takumi Nishiyasu, Will Price, Paola Ruiz, Mery Ramazanova, Leda Sari, Kiran Somasundaram, Audrey Southerland, Yusuke Sugano, Ruijie Tao, Minh Vo, Yuchen Wang, Xindi Wu, Takuma Yagi, Ziwei Zhao, Yunyi Zhu, Pablo Arbeláez, David Crandall, Dima Damen, Giovanni Maria Farinella, Christian Fuegen, Bernard Ghanem, Vamsi Krishna Ithapu, C. V. Jawahar, Hanbyul Joo, Kris Kitani, Haizhou Li, Richard Newcombe, Aude Oliva, Hyun Soo Park, James M. Rehg, Yoichi Sato, Jianbo Shi, Mike Zheng Shou, Antonio Torralba, Lorenzo Torresani, Mingfei Yan, and Jitendra Malik. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18995–19012, 2022. 1
- [3] KunChang Li, Yanan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding, 2024. 1
- [4] 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. 1
- [5] 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. 1
- [6] 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

- [7] Dídac Surís, Sachit Menon, and Carl Vondrick. Vipergpt: Visual inference via python execution for reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11888–11898, 2023. [1](#)
- [8] 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](#)
- [9] 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. [1](#)
- [10] 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. [1](#)