StreamEQA: Towards Streaming Video Understanding for Embodied Scenarios

Supplementary Material

In this supplementary material, we provide additional details and analyses to complement the main paper. Section 6 presents a comprehensive comparison of our proposed StreamEQA with existing video understanding benchmarks, along with extended data statistics. Section 7 describes our experiments, including specific experimental implementations and more quantitative results on the performance of the models. Section 8 dives deeper into the construction process of StreamEQA, specifically focusing on the information extraction from meta annotations. Finally, Section 9 shows more data examples of each sub-task.

6. Comparison with Existing Benchmarks

As shown in Table 4, StreamEQA sets itself apart from existing video understanding benchmarks by integrating both streaming and embodied task characteristics. Unlike StreamingBench [18], which only focuses on streaming tasks, or EgoThink [6], which focuses on embodied scenarios, StreamEQA uniquely combines both aspects, providing a more comprehensive challenge for video-LLMs. With an average video length of 67.5 seconds, 42 subtasks categories and 21K QA pairs, it surpasses other benchmarks in terms of task variety and dataset scale, offering a more rigorous and holistic evaluation of model generalization for real-world embodied agent application.

7. More Experiments Results

7.1. Evaluation Protocol

This section supplements the settings from Section 4.1 to ensure full reproducibility of results reported in Section 4.2 and a fair comparison across all models.

Our evaluation is based on a zero-shot, single-round inference paradigm, executed on NVIDIA A6000 GPUs. As previously mentioned, video frames are sampled in fixed 64 frames (GPT-5 and Gemini-2.5-Pro are evaluated at 1 fps). To guarantee deterministic and reproducible outputs, we adhere to the official inference strategies of these MLLMs.

Furthermore, for multiple-choice questions, the prompt illustrated in Figure 6 instructs the model to return its answer selected from A, B, C, or D.

7.2. Sub-tasks Evaluation

This section provides detailed results for all three embodied levels: Perception, Interaction, and Planning, broken down by task capabilities across different levels. The results are presented in Table 5, Table 6, and Table 7.

Tasks of Perception. Most offline video-MLLMs perform relatively better in tasks of perception. Gemini-2.5-Pro and GPT-5 achieve the highest average accuracy, followed by EgoGPT. We can observe that Gemini-2.5-Pro performs better in ARB, RRB, ARR, RRR, and SUR, which contribute the most to the overall accuracy, while GPT-5 has the best performance in ARF, RRF, and SUF, indicating that the proprietary model is good at predicting objects and actions that may appear in the future. Notably, in addition to EgoGPT, some other offline open-sourced MLLMs outperform in some specific sub-tasks, such as MiniCPM-V and Qwen3VL, and their average accuracy is similar to proprietary models. This may reveal that the challenges of such tasks are limited.

Tasks of Interaction. Compared to perception tasks, all MLLMs face significant challenges in interaction tasks, with average scores below 61%. In the settings, EgoGPT considerably outperforms open-source offline video-MLLMs. For instance, in the Episodic Memory task, EgoGPT surpasses Qwen3VL, which serves as the baseline for video understanding, by approximately 31% (78.3% vs. 47.6%). Furthermore, EgoGPT also far exceeds GPT-5 in certain realtime and forward tasks, such as OER (73.0% vs. 49.6%), ARC (84.3% vs. 61.2%), ERD (49.5% vs. 39.6%), and ARP (85.7% vs. 84.9%). This indicates that explicitly designed embodied video-MLLMs for embodied scenarios hold significant advantages in interaction tasks.

Tasks of Planning. All offline Video-MLLMs exhibit relatively low performance in tasks of planning, facing challenges in embodied reasoning. However, it is noteworthy that online video-MLLMs, such as Dispider and FlashVStream, have shown certain advantages compared with most offline Video-MLLMs, especially in forward tasks. For example, Dispider outperforms Qwen3VL by approximately 11% in Steps Merging and 16% in Moving-up Steps. Ultimately, this granular analysis confirms the primary bottleneck: the key challenge facing video-MLLM practical applications is embodied scenario analysis and temporal reasoning capabilities.

Dataset	Video Length	# Test	# Categories	Streaming	Embodied
Video-MME [11]	180s	2.7K	12	✗	✗
EgoThink [6]	-	750	12	✗	✓
EgoSchema [21]	3 min	500	-	✗	✓
StreamingBench [18]	(3s, 24 min)	4.5K	18	✓	✗
OVO-Bench [17]	7 min	2.8K	12	✓	✗
StreamEQA	67.5s	21K	42	✓	✓

Table 4. Overview of each dataset's characteristics, including average video length, number of test examples, number of categories, and question types (Streaming or Embodied).

Models	Frames	Backward				Realtime					Forward				Avg.
		OPB	ARB	RRB	SUB	OPR	ORQ	ARR	RRR	SUR	OPF	ARF	RRF	SUF	
Offline Video Understanding MLLMs															
Gemini	1fps	64.2	**76.3**	**82.8**	62.0	67.1	58.7	**76.1**	**82.4**	**56.4**	58.4	79.6	82.6	64.7	**70.1**
GPT5	1fps	60.4	75.8	83.2	53.6	61.4	58.6	73.4	83.5	52.2	60.3	**80.6**	**84.5**	**65.7**	69.3
InternVL3	64	60.8	71.8	70.2	51.0	63.8	58.8	71.2	70.7	45.8	**62.1**	73.8	73.3	51.6	63.5
LongVA	64	49.0	58.4	64.8	**62.4**	48.8	46.6	59.2	67.5	57.0	41.1	63.8	68.5	56.0	57.2
MiniCPM-V	64	63.0	68.6	70.6	56.8	62.8	**62.4**	67.0	71.9	50.8	61.7	62.8	77.9	54.6	63.9
Qwen3VL	64	56.6	69.2	78.8	55.6	53.2	51.2	62.4	75.7	52.0	47.9	72.0	79.7	64.5	63.0
VideoLLaMA3	64	58.4	65.8	63.8	48.8	58.2	53.8	66.8	63.5	47.6	53.3	56.2	65.7	53.2	58.1
EgoGPT	64	**69.0**	69.2	69.6	61.6	**67.2**	61.2	66.0	70.5	**56.4**	58.7	75.4	72.5	64.5	66.3
EgoVLPv2	64	22.4	22.4	27.2	22.4	25.2	26.6	25.8	25.5	22.4	23.9	21.8	23.5	26.1	24.2
Online Video Understanding MLLMs															
Dispider	64	56.4	66.6	59.6	47.2	59.2	55.6	64.6	61.3	40.2	56.9	58.0	59.8	40.6	55.9
FlashVStream	64	55.8	65.2	69.2	57.8	51.2	50.0	62.6	68.9	53.6	49.7	62.8	71.08	56.6	59.6
TimeChatOnline	64	38.8	46.4	56.8	44.4	40.2	45.4	45.8	63.1	48.8	38.9	53.0	62.9	38.8	47.9
VideollmOnline	64	28.2	43.4	32.6	40.4	25.8	28.4	40.8	39.5	57.4	28.7	20.6	39.0	44.8	36.1

Table 5. Evaluation on Sub-tasks about Perception.

StreamEQA Inference Prompt

You are an advanced video question-answering AI assistant. You have been provided with some frames from the video and a multiple-choice question related to the video. Your task is to carefully analyze the video and provide the best answer to question, choosing from the four options provided. Do not Respond any other words except for the letter (A, B, C, or D).

Question: {}

Options:
{}
{}
{}
{}

Figure 6. Prompt of StreamEQA Inference

8. Data Construction Details

This section provides a more detailed description of the data construction process to ensure reproducibility, and offer deeper insight into how StreamEQA is systematically synthesized.

8.1. Extraction of Meta Information

Accurate meta informations is essential for QA construction. We extract such information from the meta annotations sourced from the HD-EPIC dataset [23], which provide dense narrations with precise timestamps, event-level time ranges, eye-gaze priming, and spatial trajectories of objects. Specifically, we determine the temporal boundaries of each event (comprising multiple actions) within the video based on event time ranges and eye-gaze priming. The spatial relationships of objects are derived from the objects tracks, while meta informations of interaction and planning are inferred from the dense narrations. We will detail each key step of this extraction process in the following.

Object Spatial Relationship. The spatial relationship between two objects is computed directly from their bounding boxes, obtained as top-left and bottom-right coordinates from the objects' spatial tracks. For each video frame, all co-occurring object pairs are evaluated. If two bounding boxes are identical or one is fully enclosed by the other, the relationship is assigned as "same place". Otherwise, we calculate the center point of each box (via the average of its diagonal coordinates) and compare their horizontal and vertical offsets to derive directional relations such as "above",

Models	Frames	Backward					Realtime					Forward						Avg.
		ASR	EMM	ART	AIR	BRC	AR	OER	ARC	AIU	OMP	NAP	EPD	ERD	AIP	ARP	FRC	
Offline Video Understanding MLLMs																		
Gemini	1fps	55.2	73.7	51.9	64.6	44.8	63.1	49.7	60.7	61.6	**62.2**	37.2	60.2	37.7	57.0	73.0	45.5	56.4
GPT5	1fps	**62.1**	72.4	50.0	71.8	47.8	**71.3**	49.6	61.2	**76.6**	52.2	38.0	**66.7**	39.6	**67.9**	84.9	44.4	59.8
InternVL3	64	45.5	44.9	39.8	48.0	32.3	60.5	36.6	36.9	44.2	32.9	37.4	36.5	16.8	41.1	65.8	26.2	40.3
LongVA	64	43.1	14.9	23.0	26.1	35.4	58.5	8.6	10.0	27.4	32.9	37.6	15.7	12.3	22.0	41.3	15.2	26.5
MiniCPM-V	64	43.1	41.5	43.8	40.2	55.8	58.6	33.6	30.7	39.4	52.0	34.4	33.3	9.1	31.9	62.6	41.6	39.9
Qwen3VL	64	50.3	47.3	47.6	50.2	**72.2**	62.7	45.0	50.9	53.8	33.5	**44.0**	35.3	23.6	44.6	69.2	**67.0**	49.9
VideoLLaMA3	64	36.9	29.2	30.0	32.2	53.2	60.9	19.6	27.9	35.2	30.0	33.0	17.1	11.9	26.4	42.05	45.8	33.2
EgoGPT	64	43.1	**78.3**	**59.0**	**72.4**	60.8	60.1	**73.0**	**84.3**	68.8	37.8	40.6	59.0	**49.5**	61.7	**85.7**	34.2	**60.5**
EgoVLPv2	64	10.7	26.5	26.9	26.6	24.1	28.3	26.2	21.5	26.8	25.1	11.4	25.3	2.4	25.8	26.8	24.2	22.4
Online Video Understanding MLLMs																		
Dispider	64	29.4	26.1	38.8	36.5	15.7	63.4	24.5	27.6	34.8	26.4	28.0	20.2	6.1	30.0	57.4	10.7	29.5
FlashVStream	64	49.1	36.5	33.5	44.2	65.2	64.6	27.4	39.0	44.6	27.3	36.5	19.6	4.2	35.5	63.1	56.4	40.7
TimeChatOnline	64	48.1	18.2	23.6	23.0	56.8	62.5	15.0	17.9	20.4	24.0	40.4	13.1	15.0	17.9	39.7	28.4	29.5
VideollmOnline	64	26.7	33.0	17.8	38.2	42.2	57.3	30.2	24.5	37.8	6.9	34.6	22.5	0.2	23.6	50.5	36.6	30.2

Table 6. Evaluation on Sub-tasks about Interaction.

Models	Frames	Backward			Realtime		Forward								Avg.
		PRF	BPO	BPR	RPO	RPR	FPS	FAC	FPA	RPD	SMG	FPR	MUS	FPO	
Offline Video Understanding MLLMs															
Gemini	1fps	65.8	53.6	51.2	54.5	54.3	**46.7**	47.9	55.3	**34.4**	53.3	52.7	49.2	60.7	52.5
GPT5	1fps	**81.5**	64.6	64.6	**61.0**	53.8	45.4	**48.9**	**62.1**	30.7	59.1	46.8	50.0	**67.7**	55.3
InternVL3	64	73.0	41.6	51.8	47.2	48.2	40.0	43.3	59.6	28.9	54.7	52.4	38.8	43.4	46.5
LongVA	64	53.0	34.01	33.9	42.2	36.5	22.6	27.3	50.1	20.0	29.7	33.5	37.9	33.1	33.8
MiniCPM-V	64	63.5	42.0	47.5	39.8	39.7	33.3	33.8	48.7	26.3	34.9	41.6	35.6	37.6	39.2
Qwen3VL	64	68.7	46.6	46.8	43.4	43.6	34.8	37.1	53.9	26.3	42.7	48.8	32.0	40.2	42.0
VideoLLaMA3	64	55.1	28.8	41.6	34.6	40.6	33.4	35.9	44.3	27.3	39.1	41.4	32.2	35.9	36.7
EgoGPT	64	62.7	**67.8**	**68.4**	53.8	**62.8**	42.0	40.7	59.8	31.1	**62.5**	**74.1**	**50.4**	53.4	**55.7**
EgoVLPv2	64	27.7	22.4	23.8	23.0	27.1	26.4	20.0	25.5	23.0	28.3	24.5	25.4	27.5	24.8
Online Video Understanding MLLMs															
Dispider	64	59.7	44.1	60.5	37.9	54.5	31.2	34.4	46.5	30.6	53.9	65.1	48.2	35.7	45.9
FlashVStream	64	69.7	39.3	51.8	46.3	48.2	37.3	40.8	59.8	22.1	47.5	52.7	43.5	45.8	45.2
TimeChatOnline	64	55.2	38.4	40.9	44.6	36.8	19.9	23.3	53.7	21.5	32.5	37.3	32.9	41.1	35.6
VideollmOnline	64	53.0	20.8	45.0	22.6	48.0	27.0	28.1	50.7	22.5	38.3	71.5	47.4	20.7	37.2

Table 7. Evaluation on Sub-tasks about Planning.

"below", "left", or "right". When both axes indicate displacement, composite relations (e.g., "below-left") are resolved using a predefined mapping to the closest semantically meaningful descriptor. Finally, unreasonable or visually implausible relations are filtered out using cues from narrative annotations, ensuring that the resulting spatial labels remain consistent with the actual scene dynamics.

Object Semantic Relationship. We leverage GPT-5 to extract semantic relationships between objects from the meta annotations, with the prompt shown in Figure 7. Specifically, for each event, we first identify the five most frequent nouns from its dense narrations. We then analyze all pairwise combinations and consider two categories of relationships: functional relationships (e.g., "contains") and interaction relationships (e.g., "assists"). For each object pair, GPT-5 is instructed to output a hexatuple consisting of the two object names, the actions occurring between them, and the corresponding relationship types.

Actions Motivation. Figure 8 illustrates the prompt for actions motivation extraction. The prompt is provided with the top 5 objects with the highest occurrence frequency within the event, and requires the model to extract the actions related to these objects, as well as analyze the motivations and outcomes of such actions.

Procedure Optimization. The Procedure Optimization Prompt comprises two components. As illustrated in Figure 9, the swap/merge steps analysis prompt instructs GPT-5 to analyze whether two operations described in the context

Objects Relationships Extraction Prompt

Analyze all pairwise relationships between the top 5 most frequent nouns in the following video description.

Top 5 nouns: {(top5_noun_names)}

All noun pairs to analyze (each pair must be evaluated): {(a) & (b)}

Focus on these relationship categories:

1. Functional relationships: {(RELATIONSHIP_TYPES['functional'])}
2. Interaction relationships: {(RELATIONSHIP_TYPES['interaction'])}

For each valid relationship between every pair, provide:
- noun_pair: list of two related nouns (from the pairs listed above)
- relationship_type: the specific relationship from the above categories
- original_verb: the original verb from the narration describing their interaction

Return results in English according to the following NEW format (6 elements, separated by semicolons):
{{Original name of noun A; Original name of noun B; main action related to noun A; main action related to noun B; relationship_type; Original description when the verb happens}}

Notes:
- 'relationship_type' must be one of the specified relationship categories
- 'main action related to noun A/B' refers to the main behavior or action that the object participates in.
- 'Original name of noun A'+'main action related to noun A'+'Original name of noun B' must be sentence
- 'Original name of noun B'+'main action related to noun B'+'Original name of noun A' must be sentence
- The generated six elements must be enclosed in curly braces '{{...; ...; ...; ...; ...; ...}}'.
- If no valid relationship exists between a pair, you can omit it, but you have to at least analyse 5 pairs.

Video description: {context}

Figure 7. Prompt of Object Semantic Relationship Extraction

Actions Motivation Extraction Prompt

Extract Five-tuple information for each high-frequency noun from the following video description.

List of high-frequency nouns (must all be analyzed):
{(top5_noun_names)}

Five-tuple format (separated by semicolons, enclosed in curly braces):
{{Original name of noun A; main action related to noun A; motivation of the action related to noun A; result of the action related to noun A; Original description when the verb happens}}

Element explanations:
- 'Original name of noun A': Must be the original noun from the list
- 'main action related to noun A': Core action involving the noun (from the narration text)
- 'motivation of the action related to noun A': Motivation for the action, must be a complete sentence, must NOT use pronouns like 'it', and be concisely expressed
- 'result of the action related to noun A': the outcome or result of the action related to noun A
- 'Original description when the verb happens': Original text fragment from the narration that describes the action

Requirements:
1. Each high-frequency noun must correspond to one Five-tuple; generate at least 5 Five-tuple
2. The motivation part must not omit the subject (use the noun itself or clear reference)
3. All elements must come from the video description; do not fabricate information

Video description: {context}

Figure 8. Prompt of Actions Motivation Extraction

can be merged or swapped to enhance efficiency. Additionally, we have designed a second prompt for single-operation optimization (Figure 10). For a given event, the prompt is designed to identify inefficient operations within the event and propose an optimized procedure for achieving the target.

Swap/Merge Steps Analysis Prompt

Extract five-tuple information for operation pairs from the following video narration.

Five-tuple format (separated by |, enclosed in curly braces):
{{Operation A | Operation B | Is it possible to swap order/merge (swap, merge, No) | how to conduct | impacts of swap or merging}}

Element explanations:
- 'Operation A' and 'Operation B': Must be original sentences from the narration, and Operation A appears before Operation B in the narration. mergeable operations must not be adjacent but swappable operations can be adjacent.
- 'Is it possible...': Determine if the two operations can be swapped (fill 'swap'), merged (fill 'merge'), or not (fill 'No') without affecting the task goal. 'merging' refers to the fact that two actions can be expressed as a single action, rather than the simple concatenation of two ADJACENT narrations. Examples of merging include: 'combine cutting onions and cutting green peppers' or 'take out all the eggs needed for the task from the refrigerator at once instead of multiple times'. CRITICAL: swapping or merging must increase efficiency. If swapping or merging would REDUCE efficiency, you must output 'No' for this field.
- 'how to conduct': If swap/merge is possible, create a COMPLETE SENTENCE that modifies Operation A to incorporate the swap or merge. This sentence should clearly show how to perform the swap or merge based on Operation A. If not possible, fill 'no need to conduct'.
- 'impacts of swap or merging': If swap/merge is possible, explain the impacts (e.g., efficiency, complexity); if not, explain the reason why they can't be swapped/merged

Requirements:
1. Extract at least 5 operation pairs
2. At least 2 of them must be swappable ('swap') and at least 2 of them must be mergeable ('merge')
3. All elements must be derived from the narration; do not fabricate information
4. Operation pairs must be relevant (adjacent or related steps in the same task flow)

Video narration: {context}

Figure 9. Prompt of Swap/Merge Steps Analysis

Operation Optimization Analysis Prompt

Extract six-tuple information related to efficiency improvement from the following video narration.

FOCUS EXCLUSIVELY ON THE MOST CRITICAL OPERATIONS THAT REQUIRE OPTIMIZATION. Ignore minor improvements
that only slightly reduce steps or time. Prioritize steps where optimization would:
- Result in SIGNIFICANT overall efficiency gains (not marginal improvements)
- Effectively avoid risks or mitigate major drawbacks
- Have a substantial impact on the entire task flow

Six-tuple format (separated by |, enclosed in curly braces):
{{Operation A | In which aspect of Operation A can efficiency be improved | How to improve efficiency of Operation A | impacts of improving efficiency | Practices that neither improve nor reduce efficiency | Inefficient practices}}

Element explanations:
- 'Operation A': Original sentence from the narration describing a specific operation (focus on critical steps only)
- 'In which aspect of Operation A can efficiency be improved': Specific aspects where optimization would bring substantial gains (not marginal ones)
- 'How to improve efficiency of Operation A': Concrete methods that would result in significant efficiency improvements
- 'impacts of improving efficiency': Significant effects on overall task efficiency, risk reduction, or workflow improvement
- 'Practices that neither improve nor reduce efficiency': Practices that do not increase or decrease efficiency
- 'Inefficient practices': Practices that would reduce efficiency

Requirements:
1. Extract 3-5 most critical six-tuples (quality over quantity)
2. All elements must be derived from the narration; no fabrication
3. Only include operations where optimization would have a meaningful impact on the entire process
4. Exclude trivial improvements (e.g., minor hand movement adjustments, negligible time savings)

Video narration: {context}

Figure 10. Prompt of Operation Optimization Analysis

Plan Reflection and Adjustment. To generate QA pairs for evaluating models' capabilities to reflect on past procedures and adjust future plans accordingly, we have designed two prompts. As illustrated in Figure 11, the Plan Reflection

Plan Reflection Prompt

Please analyze the following action sequence and identify the MOST CRITICAL missing steps and their impacts.

The input contains numbered actions with timestamps in the format: "1.Action description.(timestamp);2.Action description.(timestamp);..." Please keep the numbering and timestamps in your analysis.

IMPORTANT: Focus only on the most meaningful and critical dependencies. Do not analyze every single action - select only the most important ones that would cause significant problems if missing.

For each selected critical action A, please analyze:
1. The next action B that is directly and critically dependent on action A
2. What significant impact would occur if action A is missing (focus on major problems, not minor inconveniences)
3. Whether action B can still be performed normally when action A is missing
4. What practical remedial methods are available if action B cannot be performed

Selection criteria for meaningful actions:
- Actions that are prerequisites for safety (e.g., turning off gas, checking equipment)
- Actions that are essential for the main goal (e.g., key cooking steps, critical preparation)
- Actions that would cause significant delays or failures if missing
- Actions involving dangerous or expensive materials/equipment

Please output the analysis results in the following format:

```
{
  "Action A (with number and timestamp)": {
    "Related Next Action B": "Specific description with number and timestamp",
    "Impact of Missing A": "Detailed impact description focusing on significant problems",
    "Is B Feasible Without A": "Yes/No",
    "Remedial Method": "Specific practical remedial measures"
  }
}
```

Focus on realistic, practical problems that could occur in real-world scenarios. Prioritize quality over quantity - analyze fewer actions but with deeper, more meaningful insights.

Action sequence: {action_sequence}

Figure 11. Prompt of Plan Reflection

Forward Plan Analysis Prompt

You are given a cooking activity with continuous, numbered narrations including timestamps, and a brief recipe hint listing possible ingredients and key steps.

Goal: Identify ingredients that are added or used in the activity. For EACH ingredient, you MUST provide at least one 'add' entry. 'use' entries are optional.
For each entry, provide: narration used (keep original numbering and timestamp), risk (short), prevention (short).

Requirements:
- Output compact JSON only, no extra text.
- Use the exact numbering and timestamps from the input for narration used.
- Prefer 3-6 ingredients when possible.
- Structure by category: for each ingredient, split entries into 'add' and 'use' arrays.

Activity label: {activity_label}

Recipe hint:
{recipe_hint}

Continuous narrations:
{continuous_narration}

Output format example (structure only):

```
{
  "items": [
    {
      "ingredient": "milk",
      "add": [ { "narration_used": "4.Open the milk bottle...(25.50)", "risk": "Spilling", "prevention": "Hold bottle steadily" } ],
      "use": [ { "narration_used": "7.Pour milk into frother...(32.10)", "risk": "Overfilling", "prevention": "Check max line" } ]
    }
  ]
}
```

Figure 12. Prompt of Forward Plan Analysis

Prompt analyzes an action sequence to identify the most critical actions, assessing their significant impacts if missing, the feasibility of subsequent actions, and practical remedial measures. In addition, we have designed the Forward Plan Analysis Prompt (Figure 12), which is tailored to cooking activities to identify ingredients incorporated in the cooking process. For each ingredient, the prompt requires GPT-5 to analyze three key aspects in a structured JSON format: when and how to adjust the ingredient usage, the risks associated with relevant operations, and corresponding preventive measures.

Refinement Prompt

Please help modify the incorrect options of the following QA to make them more relevant to the context, while keeping the correct option unchanged. The modified options should be confusing but still related to the context.
For increasing the difficulty of distinguishing from the correct answer, Ensure modified options are plausible but must not have the same meaning as the correct option.

Requirements:
1. Only modify incorrect options (options that are not the correct answer).
2. The correct option is at index {gt_idx} (0-based), which must remain unchanged.
3. All options must be related to the question and context.
4. Ensure modify incorrect options are not completely illogical, but will ultimately be excluded due to contradiction with video behaviors.
5. Each modify incorrect options must contain a seemingly reasonable optimization action or reason, but actually has hidden flaws or conflicts with video details.
6. Keep the original format: options should be complete sentences.
7. The output must be a JSON object with the same structure as the input QA.

Context (related activities): {context}
QA to modify: {qa_item}

Figure 13. Prompt of QA Refinement

8.2. QA Refinement

To ensure StreamEQA can effectively evaluate capabilities about video understanding and reasoning, QA Refinement is introduced to address the limitations of originally generated distractors (incorrect options), such as contextual irrelevance or weak, misleading nature. When conducting QA refinement, several key aspects need to be highlighted. Firstly, only distractors should be modified, while the correct option (gt_idx) remains unchanged. Furthermore, modified incorrect options must be contextually related and plausible but ultimately excluded due to contradictions with video context, and distractors should not share the same meaning as the correct option. In addition to QA pairs (qa_item) that need to be refined, we provide context to supplement the details and relationships of relevant activities, as shown in Figure 13. Besides, Figure 14 summarizes statistics of objects and action categories appearing in the benchmark.

8.3. Quality Control

StreamEQA performs multi-round quality control (QC). In each quality control round, two inspectors each sample 10 QA from each of the 42 topics. Any QA is refined or discarded if it is illogical, inconsistent with the video content, or violates common human intuition. After each round, additional QA are generated to fill the vacancy. The QC process stops when the inspectors complete three consecutive rounds without discarding any QA. In total, approximately 6K QA are discarded over about 40 QC rounds.

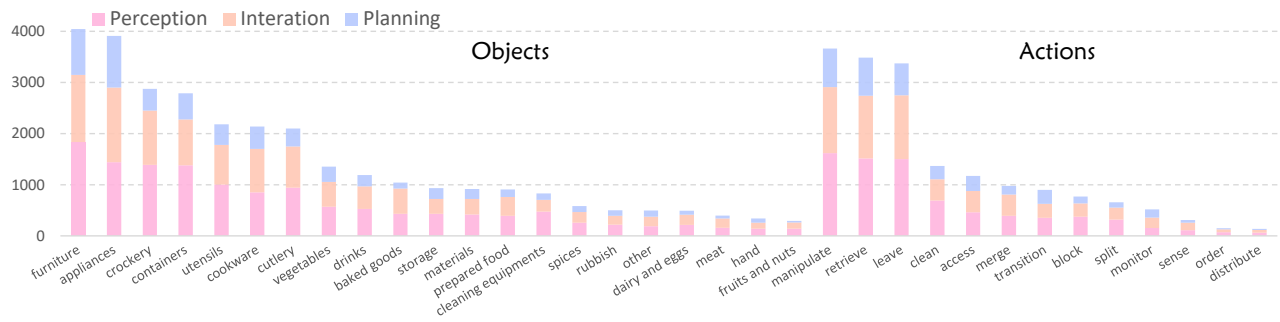


Figure 14. Statistics of objects and action categories appearing in the benchmark

9. More Data Examples

As shown in Figure 15 to Figure 19, we provide more examples extracted from our benchmark. We try to cover different tasks to offer a holistic overview of StreamEQA. Furthermore, we offer four hard samples to show the difficulties of StreamEQA in Figure 20.

Figure 15. Data examples for Attribute Recognition Backward, Attribute Recognition Forward, Attribute Recognition Realtime, Object Perception Backward, Object Perception Forward, Object Perception Realtime, Object Requirement, Relationship Recognition Backward, Relationship Recognition Forward, SpaTial Understanding Forward tasks.

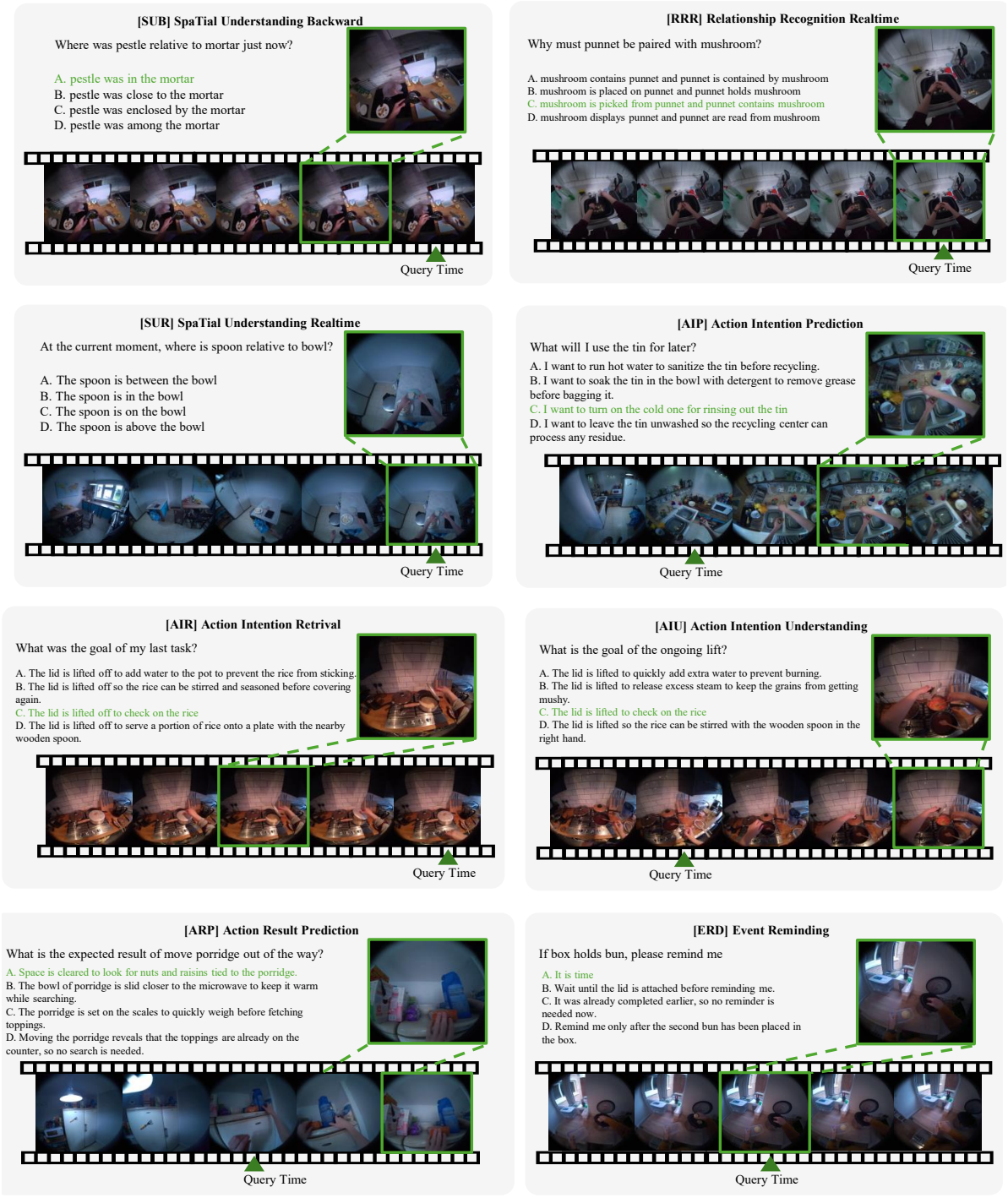


Figure 16. Data examples for SpaTial Understanding Backward, Relationship Recognition Realtime, SpaTial Understanding Realtime, Action Intention Prediction, Action Intention Retrival, Action Intention Understanding, Action Result Prediction, Event Reminding tasks.

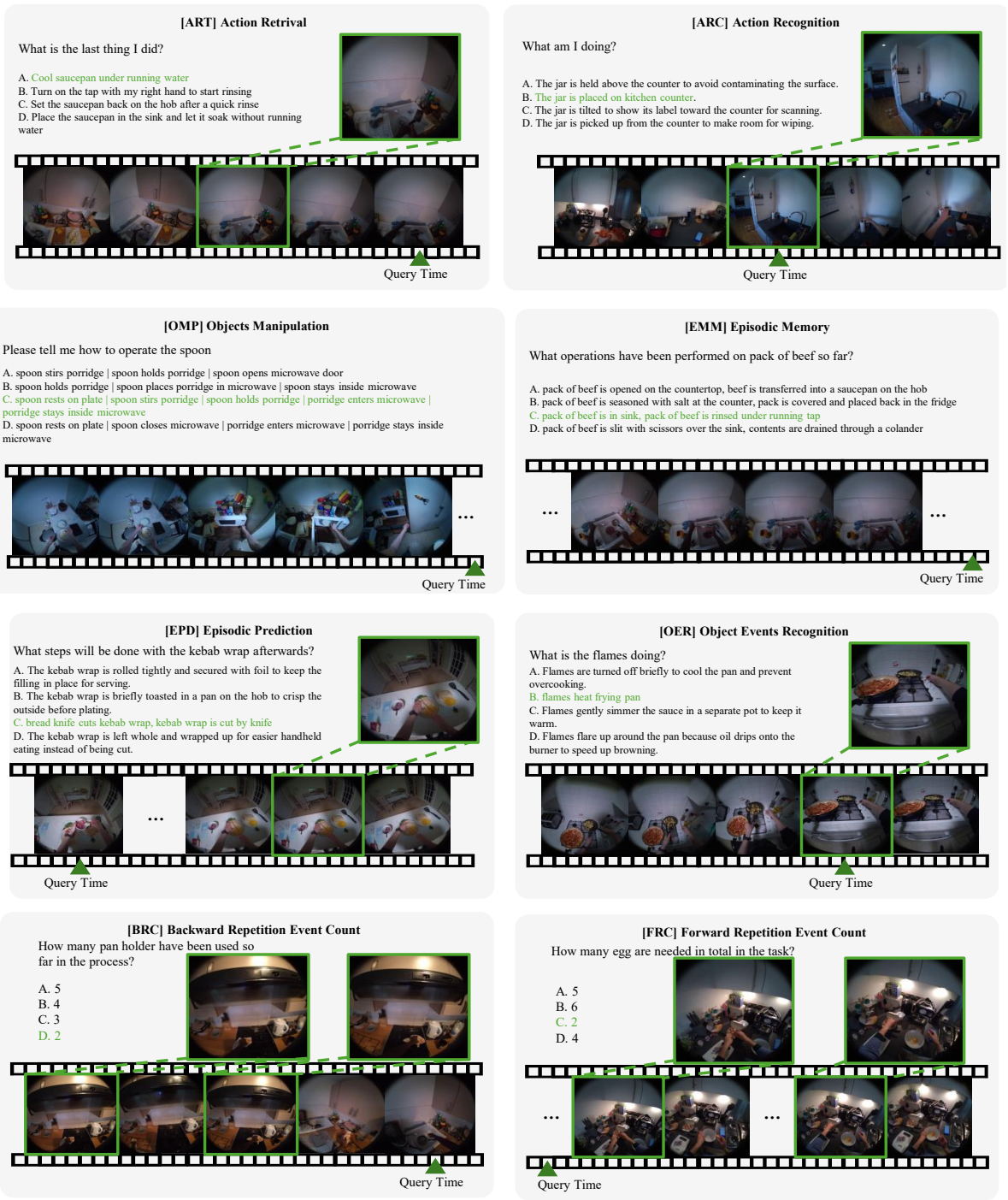


Figure 17. Data examples for Action Retrieval, Action Recognition, Objects Manipulation, Episodic Memory, Episodic Prediction, Object Events Recognition, Backward Repetition Event Count, Forward Repetition Event Count tasks.

Figure 18. Data examples for Actions Relationship, Action Sequence Retrieval, Next Action Prediction, Backward Procedure Optimazation, Backward Procedure Replacement, Forward Procedure Optimazation, Forward Procedure Replacement, Move up Steps tasks.

[RPO] Real-time Procedure Optimazation

Can you help me improve the efficiency of the operation right now?

A. Open the cutlery drawer using the left hand (where the knife is actually retrieved) instead of checking the drying rack
B. Transfer the serrated knife from the left hand to the right hand
C. After rinsing each container, place it inverted to drain over the sink in one step, then dry hands once with the cloth instead of shaking hands multiple times.
D. Looking for the serrated knife in the utensils drying rack when it is not there

Query Time

[RPR] Real-time Procedure Replacement

I want to Stir onion in frying pan with spatula distribute oil split onion chunks with spatula top first instead of left hand flick switch extractor hood turn on extractor fan. Can you tell me how to do it?

A. No, before you stir the onion in the frying pan, keep the extractor fan off to reduce noise and improve focus, this can avoid distraction while operating the spatula.
B. Yes, Before you stir the onion in the frying pan by moving them with the spatula so that the oil is well distributed among the onion, with my left hand, turn on the extractor fan by flicking the switch on the extractor hood.
C. Yes, Before you stir the onion in the frying pan, with both hands, gather onion scraps from the worktop and drop them into the bin in one trip to save steps and avoid contamination, this can keep the area tidy and facilitate continuous cooking.
D. Yes, Before you stir the onion in the frying pan, with both hands, scoop the onion to the kitchen scale after tapping its screen to turn it on, this can improve portion accuracy and facilitate even heating.

Query Time

[SMG] Steps Merging

Can we merge 'place plate next to kitchen sink on countertop with left hand' and 'pick up large plate with left hand from countertop' together to save time?

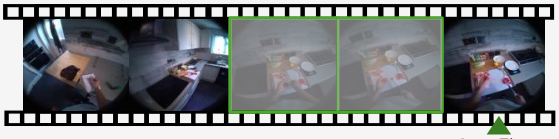
A. This can reduce repositioning and keep one continuous left-hand motion, slightly improving efficiency while the fork stays in place.
B. This can cut an extra trip between stations and increase efficiency, requiring only minor load management during the carry.
C. Avoid merging to keep the area dry and improve safety, since placing then regripping can prevent slips.
D. Combine placing the plate next to the sink and picking it up into one motion by keeping hold of it briefly.

Query Time

[FPA] Forward Plan Adjustment

If I forgot to do Make a thin slice at the top of the tomato to remove the stem area., how should I remedy it?

A. Trim the stem end or create a small flat base before proceeding; alternatively, use a tomato corer to remove the core. Use a serrated knife for better grip and a damp towel under the board to enhance stability.
B. Reposition the tomato with a stable flat side down; if none exists, trim a small flat off the curved side to create a stable base. Use a proper claw grip and ensure the cutting board is stable (damp towel or non-slip mat under the board).
C. Place a damp towel or non-slip mat under the board; create a flat face on the onion (trim end, then place cut-face down) and ensure the onion is stable before cutting. Use a claw grip to secure the onion.
D. Before proceeding, create a flat surface: cut the carrot in half lengthwise or trim an end now to stabilize it. Use a claw grip and place a damp towel or non-slip mat under the board to improve stability. Consider a cut-resistant glove if available.



Query Time

[FAC] Future Actions

After completing "opening the paper sugar bag", what steps still need to be done?

A. Turn off the cold water tap and place the saucepan on the hob while picking up the lid
B. Turn on the scale, unscrew the sugar jar, and place the lid on the scale while recalibrating
C. Retrieve the paper bag of sugar, open it, turn on the scale, and unscrew the sugar jar
D. Retrieve the sugar jar, place it on the counter, open the bag, unscrew the jar lid, and set jar and lid down

Query Time

[PRF] Plan Reflaction

What impact would occur if Missing: Proper washing of citrus juicer and guard with soap/water is not performed?

A. The bin cannot be rinsed, leaving organic residues and bacteria inside. This can cause strong odors, attract pests, and pose hygiene risks when the bin is reused. It also makes later drying ineffective because contaminants remain.
B. A used lemon half left in hand or on the board can drip juice onto surfaces, knives, and ingredients (e.g., garlic on the board), causing sticky contamination, attracting pests, and transferring microbes to high-touch points like the fridge handle. This degrades hygiene and can lead to cross-contamination throughout the kitchen.
C. Drying without washing merely spreads residual juice pulp, sugars, and microbial contamination onto the towel and the tool. Stored equipment will be sticky, attract pests, develop odors, and pose hygiene risks when next used.
D. Without detergent, grease and biofilm will not be effectively broken down, leading to visibly and microbiologically unclean surfaces. This risks cross-contamination when the strainer is used next and may require a full re-wash, wasting time and water.

Query Time

[FPS] Future Plan Summarization

Please summarize the main steps that need to be done next.

A. Open the drawer fully, move items like the colander to access contents, then select the small plastic measuring cup.
B. Move to the kitchen area, access the drawer, unscrew the blender, retrieve the measuring cup from inside, and begin examining it.
C. After closing the drawer, bend down to open the cabinet door and retrieve the measuring cup from the cupboard.
D. Close the top cupboard, retrieve a wooden spoon from the middle drawer, then close the drawer and pick up a cup of water from the kitchen top.

Query Time

[RPD] Risk Prediction

What are the possible risks of the current operation? How should I respond to these risks?

A. Risks: Metal shavings or contamination. Response: Inspect grater; clean before use
B. Risks: Bacterial transfer. Response: Use clean towel.
C. Risks: Dust transfer. Response: Check padding is clean before use.
D. Risks: Lint transfer or non-microwave-safe material. Response: Use clean, microwave-safe paper towel.

Query Time

Figure 19. Data examples for Real-time Procedure Optimazation, Real-time Procedure Replacement, Steps Merging, Forward Plan Adjustment, Future Actions, Plan Reflection, Future Plan Summarization, Risk Prediction tasks.

Realtime Procedure Optimazation (RPO)→Realtime Planning
Q: Can you help me improve the efficiency of the operation right now?
GT: Open the cutlery drawer using the left hand (where the knife is actually retrieved) instead of checking the drying rack
A: After rinsing each container, place it inverted to drain over the sink in one step, then dry hands once with the cloth instead of shaking hands multiple times. (Qwen3VL)

Objects Manipulation (OMP) →Realtime Interaction
Q: How should I handle the spatula
GT: spatula breaks mixture of dates | spatula breaks biscuits | spatula stirs content of bowl | bowl contains mixture of dates | bowl contains content of bowl | mixture of dates mixes with biscuits
A: spatula stirs content of bowl | spatula avoids breaking pieces | bowl contains mixture of dates and nuts | bowl contains content of bowl (Qwen3VL)

Action Sequence Retrival (ASR) → Backward Interaction
Q: After which step is "hold pot handle and stir“ performed?
GT: place empty bowl on counter
A: stir coconut milk (Qwen3VL)

Future Actions (FAC) →Forward Planning
Q: Please tell me what steps need to be done in the short term next.
GT: Close the drawer, pick up the tin opener, and transfer it to the right hand
A: Pick up the white plastic serving spoon, close the drawer, and transfer the spoon to the right hand (Qwen3VL)

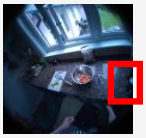


Figure 20. Representative hard examples from our benchmark.