# Conan: Progressive Learning to Reason Like a Detective over Multi-Scale Visual Evidence

Supplementary Material

## A. Related Work

### A.1. Video Reasoning Tasks

Recent advances in multimodal large language models (MLLMs) such as Qwen2.5-VL [2], Kimi-VL [25], MiMo-VL [29], and GPT-4o [13], have substantially improved video understanding, including captioning [26], question answering [1], and temporal grounding [9]. However, these capabilities mainly reflect perceptual understanding [20], whereas video reasoning [17], which demands multi-hop deduction and causal inference across frames, remains insufficiently explored and evaluated. To address this gap, several benchmarks have been introduced to assess the reasoning capabilities of MLLMs, such as Video-Holmes [6], VideoReasonBench [21], VideoEspresso [11], MMR-V [37], VRBench [33], and VCRBench [23]. Unlike conventional video understanding tasks focused on recognizing visual content, these benchmarks require models to actively locate, connect, and reason over multiple relevant clues, thereby demanding a deeper comprehension of temporal dependencies and causal structures in dynamic visual narratives.

### A.2. Video Reasoning Models

Inspired by the reasoning advancements of the DeepSeek-R1 [10], several studies [7, 16] adopt reinforcement learning with verifiable rewards [10] (RLVR) to promote video reasoning in MLLMs. While these approaches [7, 16, 22] encourage step-by-step reasoning, most are limited to text-only chain of thought, lacking explicit grounding in visual evidence, which often leads to unverified or hallucinated reasoning. To bridge this gap, concurrent works like Video-MTR [30] and Frame-Thinker [12] incorporate frame retrieval actions into the reasoning process, enabling dynamic evidence gathering and long-form understanding. Despite great improvements in long video understanding [27, 28, 36], these methods partially depend on benchmark-specific training sets and still lack a reliable evidence identification mechanism, rendering the retrieval actions less reliable. Motivated by this, we aim to develop a framework named Conan that incentivizes deductive-like reasoning abilities in MLLMs, combining precise evidence identification, logical multi-step reasoning, and confident action decision towards robust video reasoning.

## B. Dataset Construction

**Multi-scale Frame Categorization**. Frames are categorized into three types (*i.e*. *evidence*, *context*, and *distractor*), based on their frame-level scores in GenS-Video-150K [32]. Specifically, frames with scores below 3 are labeled as *distractor*, those equal to 3 as *context*, and those above 3 as *evidence*.

**Reasoning Trace Construction**. To establish the dynamic answering threshold, we adopt a double-verification mechanism. In particular, the accumulated frame descriptions and QA pair are first provided to K2 to assess whether the current evidence suffices to derive the correct answer. If so, K2 is prompted again without the answer to verify whether it can independently arrive at the correct option using the accumulated evidence. In addition, the prompt templates used to guide Kimi K2 in generating reasoning traces are illustrated in Figure 5.

**Dataset Statistics**. The Conan-91k dataset comprises $91,350$ samples, including $60,000$ for Conan-CoT-60k and $31,350$ for Conan-RLVR-31k. The duration distribution of each subset is shown in Figure 6, and the distribution of QA types is presented in Figure 7, respectively.

You are an assistant tasked with generating a step-by-step, concise reasoning process. You should act as a multimodal large language model to think based on the provided input and make the final action.

Input:
- Question: {question}
- Frame(s): {caption}
- Frame type(s): {type}

Guidelines for reasoning:
1. Begin by analyzing the question, clarifying what kind of evidence is required.
2. Analyze the relevant frames (evidence/contextual frames) with high scores that help answer the question, referencing the corresponding score if necessary.
3. Compare the available evidence across frames, giving a summary.
4. Explain why to retrieve the additional clip(s) or frame(s).
5. Conclude with a structured justification that motivates the action: {action}. The action should be rephrased concisely and fluently while maintaining all timestamps of retrived clip(s) at the end of the reasoning process.
The reasoning process is:

(a) Prompt template for free-form question.

You are an assistant tasked with generating a step-by-step, concise reasoning process. You should act as a multimodal large language model to think based on the provided input and make the final action.

Input:
- Question: {question}
- Option: {option}
- Frame(s): {caption}
- Frame type(s): {type}
Guidelines for reasoning:
1. Begin by analyzing the question and option, clarifying what kind of evidence is required.
2. Analyze the relevant frames (evidence/contextual frames) with high scores that help answer the question, referencing the corresponding score if necessary.
3. Compare the available evidence across frames, giving a summary.
4. Explain why to retrieve the additional clip(s) or frame(s).
5. Conclude with a structured justification that motivates the action: {action}. The action should be rephrased concisely and fluently while maintaining all timestamps of retrived clip(s) at the end of the reasoning process.
The reasoning process is:

(b) Prompt template for multi-choice question.

Figure 5. Prompt templates for reasoning trace construction.

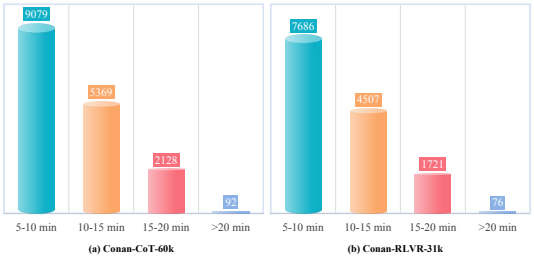



Figure 6. Duration distribution of (a) Conan-CoT-60k and (b) Conan-RLVR-31k.

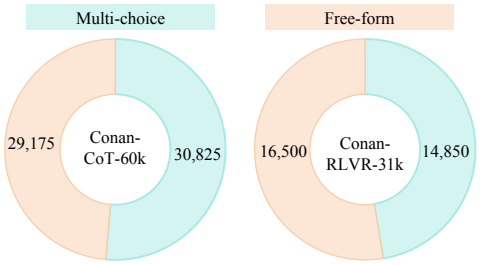



Figure 7. QA types distribution of Conan-CoT-60k and Conan-RLVR-31k.

# C. Experiment

## C.1. Benchmarks & Baselines

**Benchmark Descriptions.** 1) The six challenging video multi-step reasoning benchmarks.

- **MMR-V** [37] is a benchmark for multimodal deep reasoning in videos, containing 317 videos and 1,257 QA tasks designed to assess complex visual reasoning.
- **Video-Holmes** [6] evaluates high-level video reasoning across 1,837 questions derived from 270 manually annotated suspense short films, covering seven reasoning task types.
- **VRBench** [33] is a long-form video reasoning benchmark with 960 curated narrative videos spanning eight languages and seven categories, featuring 8,243 human-labeled multi-step QA pairs targeting temporal and causal reasoning.
- **VCRBench** [23] assesses video chain-of-thought reasoning over 859 videos with 1,034 high-quality QA pairs. We adopt its multiple-choice subset for evaluation.
- **Human-P&C** [15] includes over 6,000 human-verified multiple-choice questions evaluating nine reasoning dimensions, covering both perceptual recognition and higher-level visual comprehension integrating commonsense or domain knowledge.
- **LongVideoReason** [3] consists of 1,000 long videos designed to comprehensively evaluate reasoning along four dimensions: temporal, goal-oriented, spatial, and narrative understanding.

The four long-video understanding benchmarks.

- **LongVideoBench** [28] evaluates long-context multimodal video understanding with 6,678 multiple-choice questions derived from videos up to one hour long, covering diverse real-world scenarios. We use its validation set and remove video subtitles for fair evaluation.
- **MLVU** [36] is a multi-task benchmark containing 3,102 questions across nine categories, specifically designed for long-video comprehension. We evaluate the multiple-choice subset from the dev split (2,593 samples).
- **LVBench** [27] assesses models' ability to understand and extract information from long videos of up to two hours, comprising 1,549 QA pairs in total.
- **Video-MME** [8] is a comprehensive benchmark for general video understanding, including 900 videos and 2,700 high-quality multiple-choice questions across diverse scenarios. Subtitles are excluded during evaluation for consistency.

**Baseline Descriptions.**

- **GPT-4o** [13] is a state-of-the-art MLLM developed by OpenAI, demonstrating strong performance across diverse vision-language tasks.
- **Gemini 1.5 Pro, Gemini 2.0 Flash** are advanced models from Google's Gemini family, achieving leading performance on video understanding. Particularly, Gemini 2.0 Flash exhibits improved complex reasoning capability.
- **LLaVA-OneVision-7B** [14] integrates the Qwen2 [31] language backbone with the SigLIP [34] vision encoder, achieving strong open-source performance in fine-grained visual understanding.
- **Kimi-VL-A3B-Instruct** [25] is an efficient Mixture-of-Experts (MoE) MLLM built upon the Moonlight [19] LLM and the high-resolution MoonViT [25] vision encoder.
- **InternVL3-8B** [4] employs the InternViT-300M-448px-V2_5 [5] vision encoder and Qwen2.5-7B [31] backbone, delivering competitive open-source performance.
- **Video-R1** [7] enhances Qwen2.5-VL-7B-Instruct with CoT SFT and T-GRPO-based RLVR to improve video reasoning.
- **VideoChat-R1** [16] applies RLVR to Qwen2.5-VL-7B-Instruct, achieving strong spatiotemporal reasoning performance.
- **Video-MTR** [30] introduces multi-turn reasoning and a gated bi-level reward mechanism to enhance long-video understanding of Qwen2.5-VL-7B-Instruct.
- **Rewatch-R1** [35] augments RLVR with an Observation & Reasoning reward to improve multi-step visual reasoning based on Qwen2.5-VL-7B-Instruct.
- **FrameThinker** [12] interleaves textual reasoning with visual frames, enabling multimodal

chain-of-thought reasoning for long video understanding built upon Qwen2.5-VL-7B-Instruct. We do not include this model in the main results table, as it is not open-sourced prior to our submission.

- **Qwen2.5-VL-3B-Instruct, Qwen2.5-VL-7B-Instruct, Qwen2.5-VL-72B-Instruct** [2] are part of the Qwen2.5-VL series, which combine the Qwen2.5 [31] language model with a redesigned Vision Transformer (ViT) architecture for enhanced visual grounding and understanding.

## C.2. More comparison results & analyses

| | LongVideoBench | MLVU | LVBench | Video-MME |
|---|---|---|---|---|
| Qwen2.5-VL-7B-Instruct [2] | 48.9 | 52.8 | 34.4 | 55.8 |
| Video-R1 [7] | 55.6 | 62.5 | 38.3 | 58.6 |
| VideoChat-R1 [16] | 54.3 | 60.5 | 38.0 | 56.9 |
| Rewatch-R1 [35] | 50.5 | 55.2 | 37.2 | 58.9 |
| Video-MTR [30] | 56.4 | 59.7 | 38.6 | 59.3 |
| FrameThinker [12] | 52.9 | 59.1 | 36.6 | - |
| Conan | **56.6** (↑ 7.7 ) | **63.4** (↑ 10.6 ) | **39.2** (↑ 4.8 ) | **60.5** (↑ 4.7 ) |

Table 3. Evaluation results on long video understanding. The results of FrameThinker are taken from its original paper, which does not evaluate the entire Video-MME benchmark.

**Long Video Understanding.** Beyond multi-step reasoning, Conan exhibits strong generalization to long video understanding tasks. As presented in Table 3, Conan consistently outperforms Qwen2.5-VL-7B-Instruct across LongVideoBench [28], MLVU [36], LVBench [27], and Video-MME [8], achieving state-of-the-art performance compared with both text-CoT and video-CoT models. These results indicate that the high-quality, multi-scale evidence reasoning data and progressive training curriculum not only enhance multi-step reasoning but also effectively boost long video understanding.

**Frame Efficiency Analysis.** We compare the efficiency of Conan with Qwen2.5-VL-7B-Instruct under varying numbers of initial input frames. As shown in Figure 8, Conan with only 8 initial input frames outperforms the base model that uses 64 frames, demonstrating its superior efficiency and accuracy in visual reasoning. However, we observe a slight performance decline on the Human-P&C benchmark as the number of input frames increases. This is reasonable, as introducing more initial frames also introduces additional distractors that are irrelevant to answering the question, thereby increasing the difficulty of frame identification.

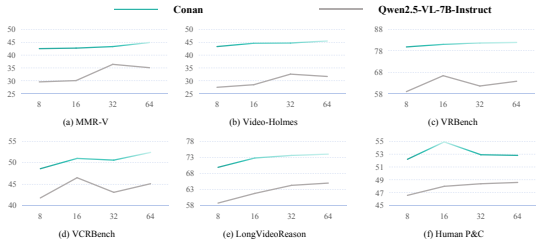



Figure 8. Performance variations of Conan and Qwen2.5-VL-7B-Instruct across different input frame numbers.

**Generalizability across model sizes and architectures.** To evaluate the generalizability of our dataset and training framework, we apply them to different base models: Qwen2.5-VL-3B-Instruct and InternVL3-8B. As shown in Table 4, both Conan-3B and Conan-Intern-8B achieve consistent performance improvements over their respective baselines, demonstrating the strong extensibility and scalability of the proposed Conan-91k dataset and two-phase training framework across varying model sizes and architectures. Notably, although InternVL3-8B demonstrates stronger baseline performance than Qwen2.5-VL-7B-Instruct, Conan-Intern-8B underperforms Conan-7B in multi-step reasoning. This discrepancy likely arises because InternVL3-8B employs a perception-oriented visual encoder and a relatively shallow cross-modal fusion module, which constrain its ability to integrate multi-step temporal cues and causal dependencies during reasoning [4]. In contrast, Qwen2.5-VL-7B-Instruct inherits reasoning priors and instruction-following capabilities from its text-based foundation model, thereby facilitating more effective compositional and multi-hop reasoning [2].

| | #Params | MMR-V | Video-Holmes | VRBench | VCRBench∗ | Human-P&C | LVR | Overall |
|---|---|---|---|---|---|---|---|---|
| Qwen2.5-VL-3B-Instruct [2] | 3B | 29.0 | 28.8 | 59.7 | 38.0 | 41.7 | 57.6 | 42.5 |
| Conan-3B | 3B | 33.1 (↑ 4.1 ) | 31.9 (↑ 3.1 ) | 68.0 (↑ 8.3 ) | 42.7 (↑ 4.7 ) | 44.3 (↑ 2.6 ) | 61.3 (↑ 3.7 ) | 46.9 (↑ 4.4 ) |
| InternVL3-8B [4] | 8B | 35.5 | 32.8 | 75.8 | 45.7 | 51.0 | 68.0 | 51.5 |
| Conan-Intern-8B | 8B | 43.3 (↑ 7.8 ) | 41.8 (↑ 9.0 ) | 81.7 (↑ 5.9 ) | 48.4 (↑ 2.7 ) | 50.2 (↓ 0.8 ) | 70.8 (↑ 2.8 ) | 56.0 (↑ 4.5 ) |

Table 4. Performance comparison on multi-step reasoning of Conan variants and their corresponding base models. **Overall** represents the average results across six multi-step reasoning benchmarks. ∗ indicates the multiple-choice subset.

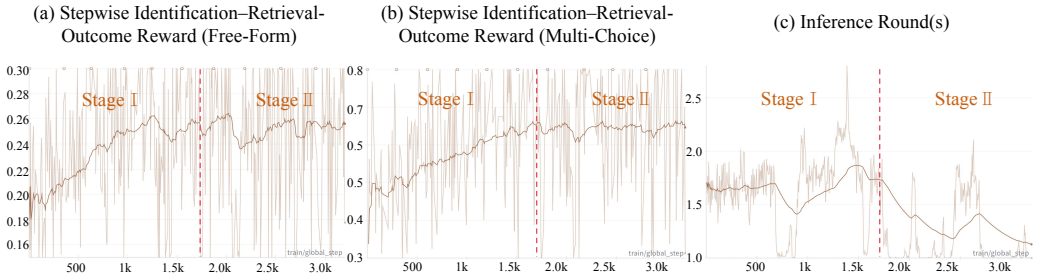



Figure 9. Training dynamics in the AIR RLVR process of Conan.

## C.3. Training Dynamics of AIR RLVR

To gain a deeper understanding of how the model's behavior evolves during end-to-end reinforcement learning, we perform a fine-grained analysis of its training dynamics. As shown in Figure 9, the training process can be divided into two stages:

**Stage I: Accuracy-Oriented Evidence Exploration.** Building upon the solid foundation established through multi-stage progressive cold-start, the model initially enters a phase characterized by frequent yet progressively more accurate frame retrieval. During this stage, it actively queries additional clips to maximize the joint identification–retrieval–outcome reward across tasks. The steadily rising reward curve indicates that the model adopts an exploratory strategy: Frequently engaging with the visual context to accumulate sufficient evidence for accurate reasoning. This stage marks a transitional period in which the model learns the value of comprehensive evidence collection before optimizing retrieval efficiency.

**Stage II: Efficient Evidence Retrieval.** As training progresses, the model transitions to a more refined and selective retrieval policy. It significantly reduces retrieval frequency while maintaining high reward accuracy, indicating that Conan has internalized a compact and efficient multi-step reasoning strategy, retrieving evidence only when necessary, much like a detective who strategically gathers key clues rather than exhaustively examining all information.

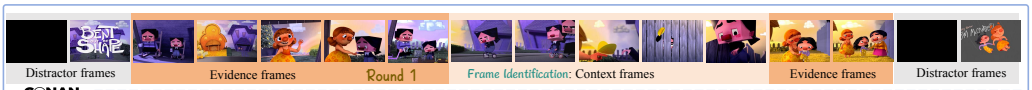
## C.4. More qualitative analyses

As shown in Figure 10, in case (a), while Qwen2.5-VL-7B-Instruct incorrectly selects option "E" by relying on superficial textual associations, Conan correctly identifies option "D" through structured, evidence-grounded reasoning. During the *Frame Identification* stage, Conan identifies context and evidence frames that emphasize the contrast in lighting and tone between the man and the girl. It then performs explicit evidence reasoning, linking the recurring visual pattern, bright, warm lighting accompanying the girl's appearance, to the symbolic themes of joy and hope she brings to the man's life. Finally, Conan synthesizes this reasoning into a confident and accurate conclusion. In case (b), despite Rewatch-R1 producing a plausible reasoning process, it fails to anchor its analysis in explicit visual evidence and ultimately predicts the

Question: The film uses completely different lighting and tones when depicting the man and the girl together. What does this symbolize?

(A) It symbolizes the inherent conflict between light and darkness. (B) It illustrates the environmental impact on human emotions. (C) It shows the contrast between urban and rural lifestyles. (D) The daughter brings joy and hope into the man's life. (E) It symbolizes the contrasting emotional perspectives between the man and the girl. (F) It depicts the balance between dreams and reality. (G) It represents the historical differences within the community. (H) It highlights societal roles during different eras. (I) It represents the diversity within the neighborhood. (J) It is symbolic of generational wisdom versus youthful ignorance.

Qwen2.5-VL-7B-Insturct: E



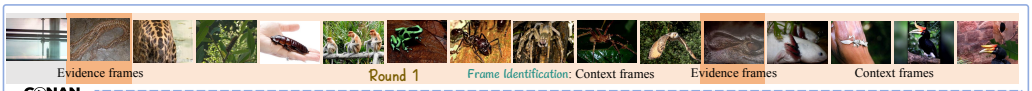

CONAN

Evidence Reasoning:1. Question focus We need to identify what the stark difference in lighting and tone between the man and the girl's scenes symbolizes...The man is outside with the girl, both happy; the lighting is bright and cheerful.3. Evidence summary The lighting and tone shift whenever the man and the girl are together. When they are apart, the man is often shown in a dim, somber setting (chair, wheelchair, sad expressions). When they are together, the scenes are consistently bright, warm, and hopeful. This pattern is repeated multiple times, emphasizing the positive impact the girl has on the man's mood.4. Option evaluation A – No direct conflict between light and darkness is highlighted. B – No environmental impact is discussed. C – No urban vs. rural contrast is mentioned. D – The girl's presence consistently brings joy and hope, aligning perfectly with the observed lighting changes. E – No contrasting emotional perspectives are explicitly stated. F – No balance between dreams and reality is emphasized. G – No historical or generational differences are highlighted. H – No societal roles during different eras are addressed. I – No diversity within the neighborhood is discussed. J – No generational wisdom vs. youthful ignorance is mentioned.5. Conclusion The consistent visual cue of brighter, warmer lighting whenever the man and the girl are together symbolizes the girl bringing joy and hope into the man's life. Therefore, the correct answer is D.
Action Decision: Confident Question Answering D

(a) A qualitative example from MMR-V.

Question: The video's creators aimed to connect one of the depicted animals to a prehistoric creature through contextual clues and visual storytelling. Which animal\'s presentation most strongly implies this intentional link to an extinct species?
A. Jaguar resting in tropical grass (0:00:00-0:00:10) B. Crocodile labeled "TITANOBOLA" (0:00:10-0:00:20)
C. Large coiled snake fossil in a museum (0:00:40-0:00:50) D. Proboscis monkey in Bornean rainforest (0:04:00-0:04:10)

Rewatch-R1: Let me consider each option carefully. The Jaguar, (A), shows an animal still alive today. The Proboscis monkey (D) is also a living primate. For an explicit connection to a prehistoric creature, neither of these fits the bill perfectly. Now, let's look at the crocodile. It's explicitly labeled \"TITANOBOLA\", which immediately brings to mind the prehistoric reptile Titanoboa, one of the largest. However, a label alone doesn't directly relate to its current physical state or morphology. Finally, we have the large, coiled snake fossil (C). While the jaguar and monkeys are present, I'm drawn to this fossil because it's a direct prehistoric representation. Unlike the modern animals, it's an inanimate object, clearly showing characteristics of an ancient species that would be classified with dinosaurs during the Mesozoic era. Its presence strongly implies an evolutionary lineage connecting primitive terrestrial reptiles or early tetrapods to modern snakes, even if dinosaurs weren't depicted in a specific way. Therefore, between the jaguar, monkey, crocodile, and the fossilised snake, the fossil clearly exemplifies the creators' intent to show an explicit connection to a prehistoric creature through its very nature as a 'fossil' and a display meant to represent extinction.
Answer: B




CONAN

Evidence Reasoning:1. Clarify the question: We need the animal whose presentation explicitly links it to a prehistoric creature, using contextual clues and visual storytelling. 2. Examine high-scoring frames that contain animals - 00:00:48– "large coiled snake fossil … in a museum setting." 00:02:25 – "a close-up of a caterpillar on a leaf." - 00:03:13– "a large cockroach on a hand." - 00:04:02 – "proboscis monkeys in a forest." - 00:04:50– "a green and black poison dart frog." - 00:05:39– "ants on a tree trunk." - 00:06:27– "a spider on a leaf." - 00:08:04– "an owl in flight." - 00:08:52– "a crocodile with the text 'TITANOBOLA'." - 00:09:41 – "a white axolotl."3. Compare the evidence - 00:00:48 shows snake fossils in museums, explicitly labeled as "fossil" and "prehistoric." - 00:08:52 shows a crocodile labeled "TITANOBOLA," but the label is modern and the context is a zoo, not a museum. - Other frames either depict living animals or do not emphasize a prehistoric link.4. Reasoning The snake fossils are the only animals whose presentation directly connects them to prehistoric creatures via museum displays and explicit labeling. The crocodile's label is modern and the setting is a zoo, so it does not fulfill the "prehistoric link" requirement as strongly.5. Conclusion The animal whose presentation most strongly implies a connection to an extinct species is the large coiled snake fossil. Answer: C
Action Decision: Confident Question Answering C

(b) A qualitative example from LongVideoReason.

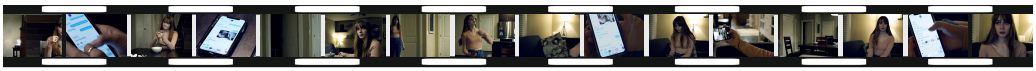
Figure 10. Two qualitative examples from MMR-V and LongVideoReason are presented, illustrating the reasoning traces and final answers of Rewatch-R1 and Conan for comparison.

wrong answer due to misalignment between reasoning (answer C) and action (answer B). In contrast, Conan accurately identifies the correct option through structured reasoning grounded in evidence frames. In the *Frame Identification* stage, Conan localizes key frames featuring the fossil exhibit, then performs evidence reasoning to associate the fossil's visual features and its museum context, explicitly symbolizing extinction, with the concept of a prehistoric creature. Conan then integrates this reasoning into a coherent and confident response, showcasing robust visual grounding and strong reasoning–action alignment. Both of the two cases highlight Conan's ability to perform compositional reasoning over visual cues, aligning symbolic interpretation with concrete visual evidence rather than

relying on spurious correlations.

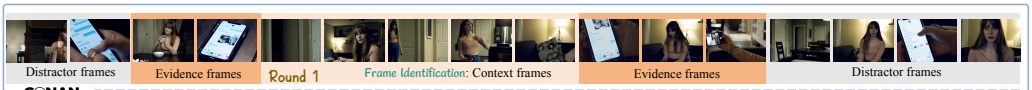
**Error Analyses.** Figure 11 presents two representative error cases from Video-Holmes and VCR-Bench. In error case (a), Conan accurately identifies and aligns the key evidence frames showing the masked man's repeated appearance on the phone screen. However, it incorrectly infers that the masked man is the boyfriend in disguise (*Option B*) rather than recognizing the correct explanation of supernatural infiltration (*Option E*). This misjudgment arises from a hallucinated causal link, interpreting visual recurrence as intentional disguise, while neglecting contextual cues, such as the absence of any indication of the boyfriend's presence. The case reveals Conan's limitation in distinguishing plausible narrative causality from repetitive visual patterns. In error case (b), Conan interprets the monkey's seemingly gentle behavior toward the pigeon as an intention to befriend it (*Option B*). While the model accurately focuses on consistent proximity and interaction frames, it overlooks subtle affective cues, such as the monkey's frustrated expressions and its act of using a scarecrow to drive the pigeon away, which point to the correct answer (*Option C*). This error underscores Conan's limitation in capturing fine-grained emotional and social dynamics, leading to an over-reliance on surface-level behavioral patterns.

Collectively, these cases reveal Conan's ongoing challenges in integrating implicit visual evidence and affective reasoning beyond superficial visual and temporal correlations.



Question: What is the root cause of the phone displaying a masked man when the girl took a photo for the second time?

A. Mobile phone camera failure.
B. Boyfriend disguised as a masked man.
C. Photo post-processing.
D. Light refraction illusion.
E. The masked man has infiltrated the room.
F. The girl had hallucinations.

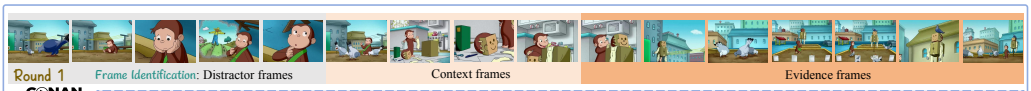



Evidence Reasoning: Step-by-step reasoning 1. Question focus We need the *root cause* of the masked man appearing on the phone screen the second time the girl takes a photo.2. Key evidence from high-score frames - 00:00:09 – The first masked-man image appears on the phone. - 00:00:18 – The girl is now holding the phone and the same masked-man image is still visible. - 00:01:05 – The girl is again holding the phone; the masked man is still on the screen. - 00:01:24 – The girl is now taking a second photo, and the masked man is still visible on the phone. - 00:01:33 – The girl is still holding the phone; the masked man is still on the screen. - 00:02:20 – The girl is looking at the phone; the masked man is still visible.3. Comparison & summary Across every high-score frame, the masked man's image is consistently shown on the phone screen while the girl is actively taking photos. No other explanations (camera failure, post-processing, light refraction, hallucinations, or infiltration) are supported by the visual evidence. 4. Option evaluation A: Camera failure – no technical issues are mentioned. B: Boyfriend disguised as a masked man – the repeated appearance of the same masked figure suggests a deliberate act rather than a one-time prank. C: Post-processing – no editing tools or filters are visible. D: Light refraction – no such effect is observed. E: Infiltration – no intruder is seen. F: Hallucinations – the masked man is repeatedly visible in the same image, not just once.5. Conclusion The only consistent and supported explanation is that the boyfriend is disguised as the masked man and is deliberately appearing in the photos.Answer: B
Action Decision: Confident Question Answering B

(a) Error case from Video-Holmes.

Question: What is the monkey's attitude towards the pigeon?

A. Catching pigeons
B. Make friends with pigeons
C. Want to drive away pigeons
D. Hope to take care of the injured pigeon.




Evidence Reasoning: Step-by-step reasoning 1. Question focus We need to know the monkey's attitude toward the pigeon(s). The key is to identify any overt actions or expressions that show the monkey's feelings.2. High-score frames that contain both the monkey and the pigeon - 00:00:00 – 00:00:02: The monkey is simply sitting on the ground; no interaction yet. - 00:00:05 – 00:00:06: The monkey is now actively talking to the pigeon while it is on the ground. - 00:00:08 – 00:00:10: The monkey is still talking to the pigeon while it is on the ground. - 00:00:12 – 00:00:15: The monkey is now standing next to the pigeon, which is perched on a box, and continues to talk to it.3. Evidence summary Across these frames, the monkey's consistent behavior is to approach the pigeon, sit or stand beside it, and speak to it. There is no attempt to catch, drive away, or care for the injured pigeon; the monkey's sole focus is communication.4. Option comparison A. Catching pigeons – no attempt to grab or harm the bird. B. Make friends with pigeons – the monkey's repeated interaction and conversation suggest friendship. C. Want to drive away pigeons – no aggressive gestures or movements. D. Hope to take care of the injured pigeon – no physical care or concern shown.5. Conclusion The monkey's attitude is best described as wanting to make friends with the pigeon. Answer: B
Action Decision: Confident Question Answering B

(b) Error case from VCRBench.

Figure 11. Two error cases of Conan from Video-Holmes and VCRBench.