This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

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

Kun Ouyang[1], Yuanxin Liu[1], Linli Yao[1], Yishuo Cai[1],
Hao Zhou[2], Fandong Meng[2], Jie Zhou[2], Xu Sun[1]
[1]State Key Laboratory for Multimedia Information Processing,
School of Computer Science, Peking University
[2]WeChat AI, Tencent Inc., China
kunouyang10@gmail.com, xusun@pku.edu.cn



Figure 1. **Top**: The reasoning process of our Conan on an example question. **Bottom**: Performance comparison across six multi-step reasoning benchmarks.

Abstract

Video reasoning, which requires multi-step deduction across frames, remains a major challenge for multimodal large language models (MLLMs). While reinforcement learning (RL)-based methods enhance reasoning capabilities, they often rely on text-only chains that yield ungrounded or hallucinated conclusions. Conversely, frame-retrieval approaches introduce visual grounding, yet still struggle with inaccurate evidence localization. To address these limitations, we present Conan, a framework for evidence-grounded multi-step video reasoning. Conan identifies context and evidence

*frames, reasons over cross-frame clues, and adaptively decides when to conclude or explore further. To achieve this, we **1)** construct Conan-91K, a large-scale dataset of automatically generated reasoning traces that include frame identification, evidence reasoning, and action decision, and **2)** design a multi-stage progressive cold-start strategy combined with an Identification-Reasoning-Action (AIR) RLVR training framework to progressively incentivize multi-step visual reasoning. Extensive experiments on six multi-step reasoning benchmarks demonstrate that Conan surpasses the baseline Qwen2.5-VL-7B-Instruct by an average of over 10% in accuracy, achieving state-of-the-art performance. Furthermore, Conan generalizes effectively to long video understanding tasks, validating its strong scalability and robustness*[1].

1. Introduction

Frontier multimodal large language models (MLLMs) [2, 13, 25, 29] have demonstrated remarkable progress on standard video understanding tasks such as question answering [1], temporal grounding [9], and captioning [26]. However, video reasoning [6, 21] remains a substantial challenge. Unlike conventional tasks, video reasoning [3, 6, 23, 33, 37] demands active visual information accumulation across temporal spans and multi-step logical inference to reach well-grounded conclusions.

Inspired by the success of reinforcement learning with verifiable rewards [10] (RLVR) in incentivizing reasoning ability of LLMs, recent works [7, 16, 22] have begun extending this paradigm to video reasoning, achieving promising gains. Nevertheless, these approaches primarily rely on pure-text reasoning without explicit grounding in visual evidence, often leading to superficial or hallucinated reasoning chains that fail to reflect the actual video content. To integrate visual evidence into the reasoning process, concurrent works [12, 30, 35] introduce frame retrieval mechanism to enable video chain-of-thought (Video-CoT) reasoning, boosting performance of long video understanding [28, 36]. However, these approaches usually suffer from inaccurate or implicit evidence localization, yielding unreliable reasoning paths. Additionally, some methods [12, 35] partially rely on benchmark-specific training data (*e.g.* Video-Holmes [6] and LongVideoReason [3]), making it difficult to disentangle solid reasoning improvements from in-domain overfitting.

Motivated by these limitations, we aim to equip MLLMs with multi-step, evidence-grounded video reasoning skills, analogous to how Conan acts as a detective (Figure 1): Specifically, our framework identifies relevant frames at multiple scales (context and evidence frames), reasons over cross-frame clues to form coherent chains of deduction, and decides whether to draw the final conclusion or continue exploring the video. Achieving this goal raises two core challenges: 1) **How to automatically construct a high-quality evidence-based reasoning dataset** that explicitly captures evidence localization, multi-step deductive reasoning, and confident action decision. 2) **How to design training curriculum for effective acquisition of visual reasoning across multi-scale evidence**.

To tackle the **first** challenge, we introduce Conan-91k, a large-scale dataset for Conan-style evidence reasoning. Built upon the key-frame identification dataset GenS-Video-150K [32], we develop an automated pipeline to generate interleaved video-text reasoning traces using the advanced LLM Kimi K2 [24], as illustrated in Figure 2. Each reasoning trace contains three key components: a) **Frame Identification** distinguishes among evidence, context, and distractor frames. b) **Evidence Reasoning** conducts textual reasoning over the question and accumulated visual clues. c) **Action Decision** decides whether to retrieve additional frames or reach the final conclusion. Furthermore, we propose an evidence difficulty-aware sampling strategy to facilitate progressive training from simple to complex reasoning, producing 60k samples (Conan-CoT-60k) for SFT and 31k samples (Conan-RLVR-31k) for RL.

To address the **second** challenge, we intro-

[1]Code, model and dataset are available at https://github.com/OuyangKun10/Conan.

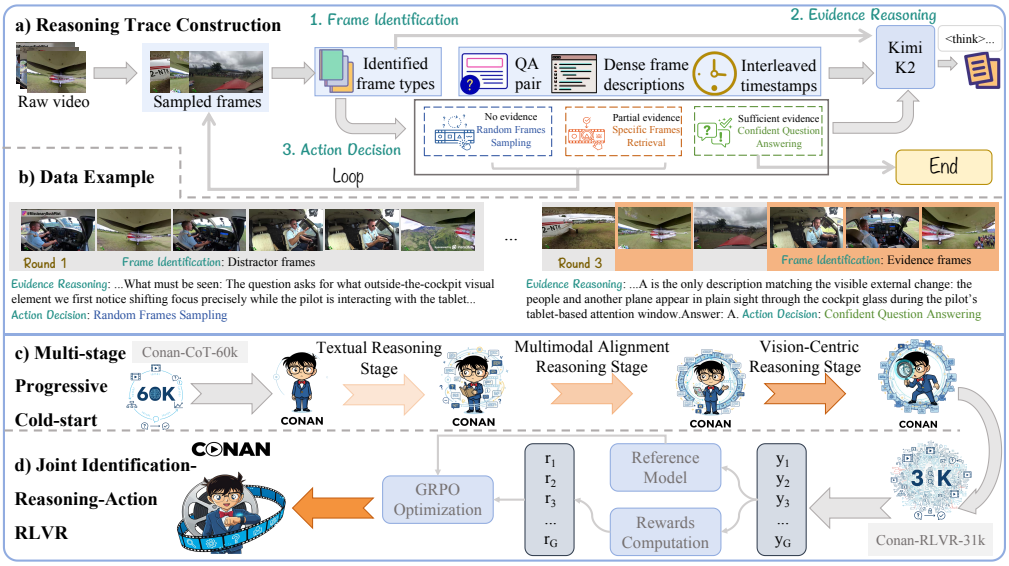


Figure 2. **a)** Reasoning Trace Construction. **b)** Data Example. **c)** Multi-stage Progressive Cold-start, including textual, multimodal alignment, and vision-centric reasoning stages. **d)** The Joint Identification-Reasoning-Action RLVR.

duce a two-phase training curriculum: a) A multi-stage progressive cold-start strategy incrementally activates the model's multi-step reasoning using Conan-CoT-60k, starting from textual reasoning, advancing to multimodal alignment, and culminating in vision-centric deduction. b) On this foundation, the joint Identification-Reasoning-Action (AIR) RLVR framework further refines the model's capacity to perform Conan-style reasoning on Conan-RLVR-31k by jointly optimizing frame identification, evidence reasoning, and action decision. Together, these components empower Conan to "seek, deduce, and act" across visual clues, achieving reliable multi-step reasoning.

Extensive experiments on six challenging multi-step reasoning benchmarks (*e.g.* MMR-V [37], Video-Holmes [6], VRBench [33], VCR-Bench [23], Human-P&C [15], and LongVideoReason [3]) demonstrate that Conan consistently surpasses the base model Qwen2.5-VL-7B-Instruct [2], with an average accuracy improvement exceeding 10%. Moreover, Conan achieves promising enhancement on long video understanding tasks (*e.g.* LongVideoBench [28], MLVU [36], LVBench [27], and Video-MME [8]), validating strong scalability and robustness.

In summary, our contributions are threefold: ♠ We introduce Conan-91k, the first large-scale dataset for multi-scale evidence reasoning, including Conan-CoT-60k and Conan-RLVR-31k. ♥ We propose a multi-stage progressive cold-start strategy and a joint AIR RLVR framework to foster gradual acquisition of evidence-based multi-step reasoning skills. ♣ We conduct extensive experiments, demonstrating that Conan achieves substantial improvements in both multi-step reasoning and long video understanding.

2. Dataset Construction

2.1. Data Collection & Processing

We collect source data from the GenS-Video-150K [32] dataset, which provides dense frame descriptions, multi-choice and free-form QA pairs, as well as frame-level relevance scores. Leveraging these relevance scores, we categorize video frames into three types: 1) Evidence frames, which are directly relevant to answering the question; 2) Con-

text frames, which offer auxiliary hints that may support the reasoning process; and 3) Distractor frames, which bear no relation to the question. This multi-scale frame categorization establishes the foundation for subsequent stepwise reasoning trace construction.

2.2. Reasoning Trace Construction

Starting from the processed data, we apply an automatic pipeline to construct Conan-style video-text interleaved reasoning traces with the assistance of a strong LLM, *Kimi K2* [24], as illustrated in Figure 2. The core loop proceeds as follows:

- We first sample 16 frames uniformly from the raw video and retain each frame's type label (evidence, context, or distractor).
- If all frames are distractors, we select the action *Random Frames Sampling*, which randomly samples 8 new frames and continue the loop.
- If some sampled frames are evidence or context but the evidence proportion does not exceed a dynamic answering threshold at which K2 can reach the answer based on the acquired clues, we select the action *Specific Frames Retrieval*, which uniformly retrieves 8 frames within a single clip or multiple clips that contain evidence/context, and continue the loop.
- If the proportion of evidence frames exceeds the threshold, we terminate the loop by selecting *Confident Question Answering* as the final action.

For every loop iteration, we prompt K2 with the chosen action, the frame types, the QA pair, dense frame descriptions, and timestamps to generate a coherent textual reasoning trace that a) analyzes the QA and sampled frames, and b) justifies the chosen action. More details are in Appendix B.

2.3. Evidence Difficulty-Aware Sampling

To facilitate a progressive training curriculum from simple to complex reasoning cases, we propose an evidence difficulty-aware sampling strategy. In particular, we define an Evidence Difficulty Index (EDI) to quantify the reasoning complexity based on the proportion and temporal dispersion of evidence frames. Let the evidence ratio be $P = m/N$, where m and N denote the numbers of evidence and total frames, respectively. The temporal variance of evidence $\mathrm{Var} = \frac{1}{m}\sum_{i=1}^{m}(x_i - \bar{x})^2$, where x_i represents the temporal position of i-th evidence frame, and $\bar{x}$ is the mean position of all evidence frames. The overall difficulty is defined as $\mathrm{EDI} = (1 - P)\,\mathrm{Var}$, where higher EDI value indicates sparser and more temporally dispersed evidence, reflecting greater difficulty.

Based on the EDI distribution and reasoning round, samples are stratified and adaptively allocated across training phases. During the **SFT** stage, we focus on lower-EDI samples to emphasize foundational reasoning skills, selecting 60k instances with up to three reasoning rounds: 25k one-round, 25k two-round, and 10k three-round samples that align with the overall round distribution. In contrast, the **RLVR** stage employs higher-EDI samples without constraining reasoning rounds, selecting 31k instances spanning diverse reasoning rounds. This difficulty-aware sampling scheme establishes a principled curriculum that transitions smoothly from low-difficulty grounding in SFT to high-difficulty multi-hop reasoning in RLVR, facilitating the gradual and robust acquisition of evidence-based reasoning capabilities. The resulting datasets, **Conan-CoT-60k** and **Conan-RLVR-31k**, together form the **Conan-91k** dataset. Dataset statistics and additional construction details are provided in Appendix B.

3. Two-Phase Training Curriculum

3.1. Multi-Stage Progressive Cold-Start

To progressively activate the multi-step reasoning abilities, we conduct a multi-stage progressive cold-start on Conan-CoT-60k, guided by a stage-wise incremental sampling that gradually expands data diversity and reasoning difficulty: **1) Textual Reasoning Stage.** In the initial stage, the model is trained on 10k relatively low-EDI samples from one-round subset, where frames are represented by dense textual descriptions and timestamps. This stage focuses on temporal and causal reasoning across ordered frame descriptions, establishing a structured reasoning foundation for subsequent multimodal learning. **2) Multimodal**

Alignment Reasoning Stage. Compared with the first stage, we utilize 25k one-round samples (including 15k new ones) and a new set of 10k relatively low-EDI samples from two-round subset, inserting timestamps and textual descriptions before visual frames. This incremental expansion ensures that the model continues learning on partially new data, preventing overfitting to previously seen questions while promoting stable adaptation from textual to multimodal reasoning. The two-round samples allow the model to collect more evidence for reasoning. **3) Vision-Centric Reasoning Stage.** In the final stage, the model is trained on the complete Conan-CoT-60k, including additional 15k two-round and 10k three-round samples. This stage compels the model to execute deep multi-step reasoning directly over visual frames with interleaved timestamps, thereby enhancing perceptual grounding and fostering vision-centric reasoning.

3.2. AIR RLVR

Building upon the model Conan-SFT obtained from the previous cold-start process, we further refine its multi-step reasoning capabilities via the Identification-Reasoning-Action (AIR) RLVR framework. Given that the model has already learned to produce reasoning traces consisting of: 1) frame identification, 2) evidence reasoning, and 3) action decision, AIR RLVR aims to optimize the exploration of effective reasoning trajectories through a set of carefully designed reward functions. We first introduce one format reward and two outcome rewards to ensure both structural consistency and answer accuracy.

Format Reward. To enforce structural consistency in model outputs y, we define a format reward R_{fmt} that verifies whether specific tags are correctly applied. And the model is restricted to performing only one action (*e.g.* random frames sampling, specific frames retrieval, confident question answering) at a time. The format reward R_{fmt} is defined as:

$$R_{fmt}(y) = \begin{cases} 0.5, & \text{if } y \text{ matches format,} \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

Multi-choice Reward. For multi-choice QA, the outcome reward R_{mc} is determined by exact match between the predicted answer y and ground truth $\hat{y}$:

$$R_{mc}(y, \hat{y}) = \begin{cases} 1, & \text{if } y = \hat{y}, \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

Free-form Reward. For free-form QA, the outcome reward R_{free} is computed as the average of ROUGE-1, ROUGE-2, and ROUGE-L scores [18] between the predicted answer y and the ground truth $\hat{y}$:

$$R_{free}(y, \hat{y}) = \frac{1}{3}\big(\mathrm{R}_1(y, \hat{y}) + \mathrm{R}_2(y, \hat{y}) + \mathrm{R}_\mathrm{L}(y, \hat{y})\big) \tag{3}$$

To evaluate the quality of the multi-scale frames identification, and frames retrieval actions, we design an identification reward R_{ide} and a retrieval reward R_{ret}, respectively: 1) The **identification reward** R_{ide} measures the average accuracy of identified evidence/context frames across reasoning rounds. 2) The **retrieval reward** R_{ret} evaluates the quality of retrieved frames by computing the average ratio of evidence/context frames among all retrieved frames. And the final joint identification-retrieval-outcome reward R_J is formulated as:

$$R_J = \begin{cases} R_{fmt} + R_o + R_{ide} + R_{ret}, & \text{if } R_o > 0, \\ R_{fmt} + R_o, & \text{otherwise,} \end{cases} \tag{4}$$

where R_o is the outcome reward and $o \in \{mc, free\}$. This reward shaping encourages the model to generate structurally valid, evidence-grounded, and accurate reasoning traces while improving retrieval efficiency. Finally, we prompt the model generate a group of responses $\{y_1, y_2, \cdots, y_G\}$, where G is the number of generated responses, and we adopt the GRPO [10] algorithm for reinforcement optimization to stabilize training and refine the reasoning policy.

4. Experiment

4.1. Evaluation Setups

Implementation Details 1) Training. We adopt Qwen-2.5-VL-7B-Instruct [2] as the base model. During the cold start, the model is trained for up

	#Params	MMR-V	Video-Holmes	VRBench	VCRBench*	Human-P&C	LVR	Avg. VR	Avg. LU
Closed-source Models									
GPT-4o [13]	-	44.0	42.0	76.7	54.0	48.4	63.1	54.7	-
Gemini 1.5 Pro	-	-	41.2	-	-	52.6	69.3	-	-
Gemini 2.0 Flash	-	42.6	30.6	-	-	56.1	-	-	-
Open-source Models									
LLaVA-OneVision-7B [14]	7B	6.5	-	-	30.7	48.5	-	-	-
InternVL3-8B [4]	8B	35.5	32.8	75.8	45.7	51.0	68.0	51.5	52.5
Kimi-VL-A3B-Instruct [25]	3B/16B	32.4	32.4	60.5	34.3	42.0	64.6	44.4	-
Qwen2.5-VL-72B-Instruct [2]	72B	39.1	40.2	72.7	50.8	55.7	72.3	55.1	53.4
Qwen2.5-VL-7B-Instruct [2]	7B	30.1	28.5	66.4	46.5	48.2	61.8	46.9	48.0
Text CoT Models									
Video-R1 [7]	7B	36.3	36.5	69.5	48.0	49.8	70.3	51.7	53.8
VideoChat-R1 [16]	7B	36.1	33.0	61.5	48.2	51.8	67.9	49.8	52.4
Video CoT Models									
Video-MTR [30] Concurrent work	7B	36.5	35.7	69.7	48.1	47.2	57.3	49.1	53.5
Rewatch-R1† [35] Concurrent work	7B	45.3	37.8	79.1	49.8	51.6	70.5	55.7	50.5
Conan SFT	7B	35.4	34.9	64.4	43.3	50.4	66.0	49.1	49.3
Conan	7B	42.7 (↑ 12.6)	44.6 (↑ 16.1)	81.0 (↑ 14.6)	51.0 (↑ 4.5)	52.3 (↑ 4.1)	72.8 (↑ 11.0)	57.4 (↑ 10.5)	54.9 (↑ 6.9)

Table 1. Main results of Qwen2.5-VL-7B-Instruct , Conan , and other baselines. **Avg. VR** and **Avg. LU** denote the average results on six video reasoning benchmarks and four long video understanding benchmarks, respectively. † indicates partial training on the training set of LongVideoReason (LVR) [3], and $*$ marks the multiple-choice subset.

to one epoch per stage with a global batch size of 32 and a learning rate of $1e$-5. The trained model is then used for the AIR RLVR phase, also trained for one epoch under the batch size of 24 and the learning rate of $1e$-6. The maximum completion length is set to $4,000$ tokens, with a generation temperature of 1.0 and a generation number G of 8. Each input video contains 16 initial frames, and the model can retrieve up to 8 additional frames per reasoning step. 2) Evaluation. The generation temperature is fixed at 1.0, and each sample is evaluated three times. The maximum number of new tokens is set to $4,000$ when reasoning traces are included and 128 for direct answering. The reasoning process is limited to three rounds. Videos are standardized to 16 frames for multi-step reasoning and 32 frames for long video understanding during evaluation, with a resolution of $448 \times 28 \times 28$.

Benchmarks & Baselines. We conduct comprehensive evaluations on six multi-step reasoning benchmarks(*i.e.*, MMR-V [37], Video-Holmes [6], VRBench [33], VCRBench [23] (multi-choice subset), Human-P&C [15], and LongVideoReason [3]) and four long video understanding benchmarks (*i.e.*, LongVideoBench [28], MLVU [36], LVBench [27], and Video-MME [8]), comparing Conan with a range of closed-source and open-source MLLMs. Descriptions of benchmarks and baselines are provided in Appendix C.1.

4.2. Main Results

The main evaluation results are shown in Table 1, while the full results of long video understanding are in Table 3 of Appendix. The key observations are summarized as follows.

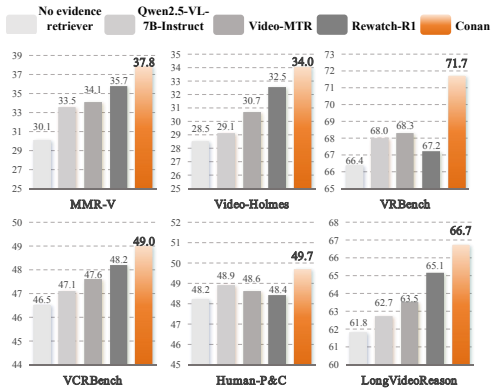


Figure 3. Performance comparison across different evidence retrievers, with Qwen2.5-VL-7B-Instruct as the question-answering model.

Overall Analysis. Across video reasoning benchmarks, Conan substantially surpasses its base model Qwen2.5-VL-7B-Instruct with an average accuracy gain of over 10%. Remarkably, Conan also outperforms the advanced GPT-4o on most benchmarks (*e.g.*, Video-Holmes, VRBench, Human-P&C, and LongVideoReason), underscoring its superior capabilities of multi-step, evidence-

	MMR-V	Video-Holmes	VRBench	VCRBench⋆	Human-P&C	LVR	Overall
Dataset design							
w/o multi-scale evidence	39.7	39.1	75.4	46.9	51.3	70.2	53.8
w/o difficulty sampling	40.0	40.8	78.9	48.4	50.7	72.3	55.2
Progressive cold-start							
w/o textual reasoning	42.2	43.9	**81.4**	49.8	51.7	**73.1**	57.0
w/o multimodal alignment reasoning	**43.5**	44.2	80.4	48.2	49.3	72.5	56.4
w/o vision-centric reasoning	39.0	36.5	75.2	50.2	48.7	68.5	53.0
w/o cold-start	39.1	36.8	73.3	46.3	47.9	62.3	51.0
AIR RLVR							
w/o identification reward	39.9	40.0	74.8	46.1	50.1	71.7	53.8
w/o retrieval reward	38.2	42.4	75.7	47.8	50.5	69.2	54.0
w/ text CoT	41.1	41.0	76.8	48.8	**52.8**	70.5	55.2
Conan	42.7	**44.6**	81.0	**51.0**	52.3	72.8	**57.4**

Table 2. Ablation results of Conan, where the best results are in **boldface**.

grounded reasoning. And the two text-CoT models (*e.g.*, Video-R1 and VideoChat-R1) perform notably worse than Conan, demonstrating the advantages of visual evidence identification and retrieval. Moreover, Conan consistently outperforms video-CoT approaches (*e.g.*, Video-MTR and Rewatch-R1), showcasing more precise evidence localization, and more effective frame retrieval of Conan. Notably, Conan also generalizes well to long video understanding tasks, with an average improvement of 6.9%, proving the robustness and scalability of our framework.

Conan as an Evidence Retriever. To further assess Conan's evidence localization capability, we compare different models used as evidence retriever, each selecting up to 16 key frames per video for Qwen2.5-VL-7B-Instruct to perform question answering. As shown in Figure 3, Conan yields the highest performance, highlighting its superior ability to identify informative visual evidence. Additional results under varying frame settings and base models are provided in Appendix C.2. And training dynamics can be found in Appendix C.3.

4.3. Ablation Study

To investigate the contribution of each component in our framework, we conduct a comprehensive ablation study with multiple Conan variants.

For the dataset design, we introduce two variants: 1) *w/o multi-scale evidence*, which merges the *context* type into the *distractor* category in multi-scale frame categorization; and 2) *w/o difficulty sampling*, which replaces the evidence difficulty-aware sampling with random sampling.

For the multi-stage progressive cold-start strategy, we develop four variants: 1) *w/o textual reasoning*, which excludes the textual reasoning stage; 2) *w/o multimodal alignment reasoning*, which skips the multimodal alignment reasoning stage; and 3) *w/o vision-centric reasoning*, which removes the vision-centric reasoning stage. 4) *w/o cold-start*, which bypasses the three cold-start stages and directly employs AIR RLVR to train the model on the Conan-RLVR-31k dataset.

For the AIR RLVR framework, we design three variants: 1) *w/o identification reward*, which discards the identification reward; 2) *w/o retrieval reward*, which removes the retrieval reward; and 3) *w/ text CoT*, which enforces single-round text-CoT paradigm performing pure textual reasoning and answering without additional retrieval in training and evaluation procedures.

The ablation results in Table 2 reveal several key findings: a) **Dataset design.** *w/o multi-scale evidence* underperforms Conan, confirming the benefits of multi-scale frame identification in providing richer contextual cues. Moreover, Conan outperforms *w/o difficulty sampling*, which validates that evidence difficulty-aware sampling effectively guides the model to progressively acquire multi-step reasoning abilities. b) **Progressive cold-start.** Conan consistently surpasses *w/o textual reasoning*, *w/o multimodal alignment reasoning*, *w/o vision-centric reasoning*, and *w/o cold start*,



Figure 4. A qualitative example from VRBench, including the QA pair and the reasoning traces of Video-R1 (Text CoT), Video-MTR (Video CoT), and Conan.

which demonstrates that the multi-stage progressive cold-start is crucial for gradually activating the model's multi-hop reasoning capabilities. c) **AIR RLVR.** Conan outperforms *w/o identification reward* and *w/o retrieval reward*, proving their effectiveness in refining evidence localization accuracy and frame retrieval efficiency, respectively. And *w/ text CoT* underperforms Conan, which validates the superiority of the Conan-style reasoning paradigm.

4.4. Qualitative Evaluation

Figure 4 presents a qualitative comparison on VRBench [33] among Video-R1, Video-MTR, and Conan. Video-R1 conducts pure textual reasoning without visual grounding, resulting in a hallucinated answer driven by linguistic priors. Video-MTR incorporates frame retrieval but fails to localize relevant evidence, leading to weak alignment between retrieved frames and the question. In contrast, Conan performs multi-step reasoning with accurate evidence grounding and efficient frame retrieval. In Round 1, it recognizes the absence of causal cues and performs random frames sampling to expand the search space. In Round 2, guided by contextual signals, it retrieves frames around key timestamps of player interactions. In Round 3, Conan identifies visual evidence of color-triggered game events (*e.g.*, team activities, flag captures) and integrates these clues to infer the correct option. This comparison highlights Conan's superior ability to localize, reason over, and act upon multi-scale visual evidence. More qualitative cases and error analyses are in Appendix C.4.

5. Conclusion and Future work

In this work, we present Conan, a novel framework that equips MLLMs with Conan-like visual reasoning through frame identification, evidence reasoning, and action decision. Employing the Conan-91k dataset, constructed via multi-scale frame categorization, reasoning trace construction, and evidence difficulty-aware sampling, we devise a multi-stage progressive cold-start strategy alongside a joint Identification-Reasoning-Action (AIR) RLVR framework to progressively cultivate robust multi-step reasoning abilities. In future work, we plan to extend Conan toward chain-of-frame reasoning, enabling dynamic frame generation during reasoning to tackle more complex video reasoning tasks.

Acknowledgements

We thank all anonymous reviewers for their insightful comments. This research was partially supported by the National Natural Science Foundation of China under Grant No. 92470205 and No. 62176002. Xu Sun is the corresponding author of this paper.

References

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015. 2, 1

[2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 2, 3, 5, 6, 1, 4

[3] Yukang Chen, Wei Huang, Baifeng Shi, Qinghao Hu, Hanrong Ye, Ligeng Zhu, Zhijian Liu, Pavlo Molchanov, Jan Kautz, Xiaojuan Qi, Sifei Liu, Hongxu Yin, Yao Lu, and Song Han. Scaling rl to long videos. In *Advances in Neural Information Processing Systems*, 2025. 2, 3, 6

[4] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024. 6, 3, 4, 5

[5] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24185–24198, 2024. 3

[6] Junhao Cheng, Yuying Ge, Teng Wang, Yixiao Ge, Jing Liao, and Ying Shan. Video-holmes: Can mllm think like holmes for complex video reasoning? *arXiv preprint arXiv:2505.21374*, 2025. 2, 3, 6, 1

[7] Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Benyou Wang, and Xiangyu Yue. Video-r1: Reinforcing video reasoning in mllms. *arXiv preprint arXiv:2503.21776*, 2025. 2, 6, 1, 3, 4

[8] Chaoyou Fu, Yuhan Dai, Yongdong 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. 3, 6, 4

[9] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE international conference on computer vision*, pages 5267–5275, 2017. 2, 1

[10] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. 2, 5, 1

[11] Songhao Han, Wei Huang, Hairong Shi, Le Zhuo, Xiu Su, Shifeng Zhang, Xu Zhou, Xiaojuan Qi, Yue Liao, and Si Liu. Videoespresso: A large-scale chain-of-thought dataset for fine-grained video reasoning via core frame selection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26181–26191, 2025. 1

[12] Zefeng He, Xiaoye Qu, Yafu Li, Siyuan Huang, Daizong Liu, and Yu Cheng. Framethinker: Learning to think with long videos via multi-turn frame spotlighting. *arXiv preprint arXiv:2509.24304*, 2025. 2, 1, 3, 4

[13] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 2, 6, 1, 3

[14] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 6, 3

[15] Keliang Li, Hongze Shen, Hao Shi, Ruibing Hou, Hong Chang, Jie Huang, Chenghao Jia, Wen Wang, Yiling Wu, Dongmei Jiang, Shiguang Shan, and Xilin Chen. Humanpcr: Probing mllm capabilities in diverse human-centric scenes. *arXiv preprint arXiv:2508.13692*, 2025. 3, 6

[16] Xinhao Li, Ziang Yan, Desen Meng, Lu Dong, Xiangyu Zeng, Yinan He, Yali Wang, Yu Qiao, Yi Wang, and Limin Wang. Videochat-r1: Enhancing spatio-temporal perception via reinforcement fine-tuning. *arXiv preprint arXiv:2504.06958*, 2025. 2, 6, 1, 3, 4

[17] Yunxin Li, Xinyu Chen, Baotian Hu, Longyue Wang, Haoyuan Shi, and Min Zhang. Videovista: A versatile benchmark for video understanding and reasoning. *arXiv preprint arXiv:2406.11303*, 2024. 1

[18] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004. 5

[19] Jingyuan Liu, Jianlin Su, Xingcheng Yao, Zhejun Jiang, Guokun Lai, Yulun Du, Yidao Qin, Weixin Xu, Enzhe Lu, Junjie Yan, et al. Muon is scalable for llm training. *arXiv preprint arXiv:2502.16982*, 2025. 3

[20] Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. Tempcompass: Do video llms really understand videos? In *Findings of the Association for Computational Linguistics*, pages 8731–8772, 2024. 1

[21] Yuanxin Liu, Kun Ouyang, Haoning Wu, Yi Liu, Lin Sui, Xinhao Li, Yan Zhong, Y Charles, Xinyu Zhou, and Xu Sun. Videoreasonbench: Can mllms perform vision-centric complex video reasoning? *arXiv preprint arXiv:2505.23359*, 2025. 2, 1

[22] Kun Ouyang, Yuanxin Liu, Haoning Wu, Yi Liu, Hao Zhou, Jie Zhou, Fandong Meng, and Xu Sun. Spacer: Reinforcing mllms in video spatial reasoning. *arXiv preprint arXiv:2504.01805*, 2025. 2, 1

[23] Yukun Qi, Yiming Zhao, Yu Zeng, Xikun Bao, Wenxuan Huang, Lin Chen, Zehui Chen, Jie Zhao, Zhongang Qi, and Feng Zhao. Vcr-bench: A comprehensive evaluation framework for video chain-of-thought reasoning. *arXiv preprint arXiv:2504.07956*, 2025. 2, 3, 6, 1

[24] Kimi Team, Yifan Bai, Yiping Bao, Guanduo Chen, Jiahao Chen, Ningxin Chen, Ruijue Chen, Yanru Chen, Yuankun Chen, Yutian Chen, et al. Kimi k2: Open agentic intelligence. *arXiv preprint arXiv:2507.20534*, 2025. 2, 4

[25] Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, et al. Kimi-vl technical report. *arXiv preprint arXiv:2504.07491*, 2025. 2, 6, 1, 3

[26] Subhashini Venugopalan, Huijuan Xu, Jeff Donahue, Marcus Rohrbach, Raymond Mooney, and Kate Saenko. Translating videos to natural language using deep recurrent neural networks. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1494–1504, 2015. 2, 1

[27] Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Shiyu Huang, Bin Xu, Yuxiao Dong, Ming Ding, and Jie Tang. Lvbench: An extreme long video understanding benchmark. *arXiv preprint arXiv:2406.08035*, 2024. 3, 6, 1, 4

[28] Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. *Advances in Neural Information Processing Systems*, 37:28828–28857, 2024. 2, 3, 6, 1, 4

[29] LLM-Core-Team Xiaomi. Mimo-vl technical report. *arXiv preprint arXiv:2506.03569*, 2025. 2, 1

[30] Yuan Xie, Tianshui Chen, Zheng Ge, and Lionel Ni. Video-mtr: Reinforced multi-turn reasoning for long video understanding. *arXiv preprint arXiv:2508.20478*, 2025. 2, 6, 1, 3, 4

[31] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. 3, 4

[32] Linli Yao, Haoning Wu, Kun Ouyang, Yuanxing Zhang, Caiming Xiong, Bei Chen, Xu Sun, and Junnan Li. Generative frame sampler for long video understanding. In *Findings of the Association for Computational Linguistics*, pages 17900–17917, 2025. 2, 3, 1

[33] Jiashuo Yu, Yue Wu, Meng Chu, Zhifei Ren, Zizheng Huang, Pei Chu, Ruijie Zhang, Yinan He, Qirui Li, Songze Li, et al. Vrbench: A benchmark for multi-step reasoning in long narrative videos. *arXiv preprint arXiv:2506.10857*, 2025. 2, 3, 6, 8, 1

[34] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986, 2023. 3

[35] Congzhi Zhang, Zhibin Wang, Yinchao Ma, Jiawei Peng, Yihan Wang, Qiang Zhou, Jun Song,

and Bo Zheng. Rewatch-r1: Boosting complex video reasoning in large vision-language models through agentic data synthesis. *arXiv preprint arXiv:2509.23652*, 2025. 2, 6, 3, 4

[36] Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Shitao Xiao, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. Mlvu: Benchmarking multi-task long video understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 13691–13701, 2025. 2, 3, 6, 1, 4

[37] Kejian Zhu, Zhuoran Jin, Hongbang Yuan, Jiachun Li, Shangqing Tu, Pengfei Cao, Yubo Chen, Kang Liu, and Jun Zhao. Mmr-v: What's left unsaid? a benchmark for multimodal deep reasoning in videos. *arXiv preprint arXiv:2506.04141*, 2025. 2, 3, 6, 1