Chain-of-Frames: Advancing Video Understanding in Multimodal LLMs via Frame-Aware Reasoning

Sara Ghazanfari[1*] **Francesco Croce**[2] **Nicolas Flammarion**[2]
Prashanth Krishnamurthy[1] **Farshad Khorrami**[1] **Siddharth Garg**[1]

[1]New York University, US [2]EPFL, Switzerland *Correspondence: sg7457@nyu.edu

Abstract

Recent work has shown that eliciting Large Language Models (LLMs) to generate reasoning traces in natural language before answering the user's request can significantly improve their performance across tasks. This approach has been extended to multimodal LLMs, where the models can produce chain-of-thoughts (CoT) about the content of input images and videos. In this work, we propose to obtain video LLMs whose reasoning steps are grounded in, and explicitly refer to, the relevant video frames. For this, we first create CoF-DATA, a large dataset of diverse questions, answers, and corresponding frame-grounded reasoning traces about both natural and synthetic videos, spanning various topics and tasks. Then, we fine-tune existing video LLMs on this chain-of-frames (CoF) data. Our approach is simple and self-contained, and, unlike existing approaches for video CoT, does not require auxiliary networks to select or caption relevant frames. We show that our models based on CoF are able to generate chain-of-thoughts that accurately refer to the key frames to answer the given question. This, in turn, leads to improved performance across multiple video understanding benchmarks.

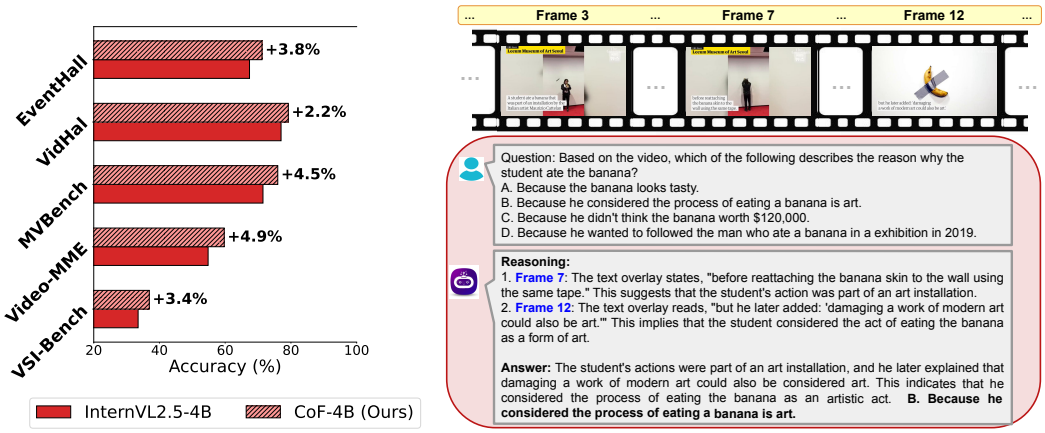
1 Introduction

Large Language Models (LLMs) are able to perform step-by-step reasoning, widely known as chain-of-thoughts (CoT) (Wei et al., 2022; Kojima et al., 2022). This capability has been implicitly integrated into state-of-the-art systems such as OpenAI's o1/o3 models (OpenAI, 2024) contributing to their remarkable performance. CoT reasoning has been also extended to multimodal LLMs (Hu et al., 2024; Wu and Xie, 2023): this presents new challenges compared to language-only domains, as the models need to attend to inputs from different modalities (Awal et al., 2023; Kil et al., 2024).

Recent work has also begun to explore the integration of CoT into video understanding tasks, where the input to a multimodal LLM consists of a *sequence* of images (the frames of the video) along with a text prompt. This makes reasoning on videos particularly complex, as the model needs to capture the semantics of the text prompt, understand temporal and causal relationships between frames, and reason about the video in its entirety. Existing approaches rely on complex inference frameworks with architecture modifications (Fei et al., 2024) or auxiliary networks (Han et al., 2024) at evaluation time to integrated reasoning into video LLMs. This makes using these models both more computationally expensive and less general (as they are specialized to some tasks), and deviates from the natural CoT prompting successfully applied to standard LLMs.

To remedy these limitations, in this work we propose chain-of-frames (CoF), a new frame-aware chain-of-thought reasoning approach for video LLMs that integrates temporal information directly into the CoT structure (see Fig. 1b). This enables the model to identify and refer to the most relevant frames while answering questions, in contrast to prior works that treat frame selection and reasoning as separate stages. Chain-of-frames is a simple and natural adaptation of the CoT paradigm in NLP to video understanding that does not require the auxiliary networks or complex inference pipelines of

39th Conference on Neural Information Processing Systems (NeurIPS 2025).



(a) CoF models vs. the baseline models.

(b) A CoF reasoning trace generated by CoF-4B model.

Figure 1: **(a)** Comparison of accuracy across multiple video understanding benchmarks between baseline model, InternVL2.5-4B and CoF-4B model: Our model consistently outperforms the baselines. **(b)** A chain-of-frames reasoning generated by our CoF-InternVL2.5-4B model, which includes the key frames to answer the question (from VIDEO-MME).

existing methods. Moreover, we propose an efficient data generation pipeline which allows use to collect a large dataset of CoF examples, named COF-DATA. To achieve this, a key element consists in leveraging a synthetic video dataset (Yi et al., 2020) to extract a large and diverse set of reasoning traces at virtually no cost. Then, we fine-tune a recent open-source video LLMs, InternVL2.5-4B (Chen et al., 2024) , on our COF-DATA. In an extensive evaluation on five established benchmarks, we show that our CoF models significantly outperform the original model (see Fig. 1a).

2 Chain-of-Frames: Reasoning on Videos via Frame References

To address the limitations, we propose **Chain-of-Frames (CoF)**, a simple yet effective approach introducing temporal grounding into the reasoning process. CoF consists of reasoning traces with explicit references to frames relevant to answering the given query. Concretely, we use the position of the frame in the video (e.g., "Frame 1", "Frame 2", ...) as an identifier. Unlike timestamps, this representation is agnostic to video duration and sample frequency, making it more consistent across diverse video data and potentially easier to learn.

Our approach offers four main benefits: Data availability, since training data can be easily generated from annotated real videos and CoF traces from synthetic videos at virtually no cost, unlike the complex pipelines of Wang et al. (2024b); Han et al. (2024); Simplicity, as reasoning traces are expressed fully in natural language without specialized formats (Fei et al., 2024; Han et al., 2024), making the method a direct extension of standard CoT in NLP; Temporal grounding, by explicitly tying reasoning to specific frames and thus strengthening video–reasoning alignment.

2.1 Chain-of-Frames training data collection (COF-DATA)

We construct chain-of-frames traces from both real and synthetic videos. For real videos, we use the training split of the VIDEOESPRESSO dataset (Han et al., 2024), which features videos from a wide range of sources and includes descriptions of key frames for each. To complement this, we use the training split of the CLEVRER dataset (Yi et al., 2020), which contains synthetic videos of simple 3D objects interacting within a controlled environment, along with rich annotations. We next describe the main steps of our data generation pipeline illustrated in Fig. 2. Additional details and examples are provided in App. B.

CoF from real videos (COF-DATA$_{\text{real}}$). VIDEOESPRESSO provides caption for key frames. After aligning frame IDs, we obtain data in the format shown in Fig. 2 (raw annotations). From these annotations, we generate triples of questions, answers and reasoning traces with frame references, by prompting an LLM using the raw annotations as input. In particular, we use Llama3.1-8B (Meta AI, 2024) (the full prompt is provided in App. B). This process yields multiple questions per video,

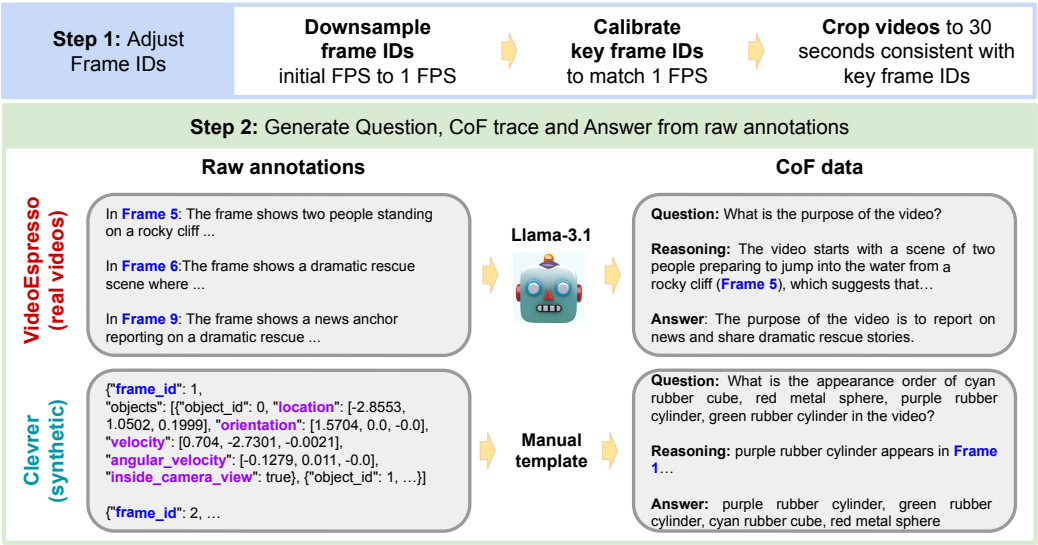


Figure 2: **Overview of our two-step pipeline for generating CoF-Data.** Step 1 adjusts the frame IDs while preserving frame-caption alignemnt. Step 2 utilizes raw annotations to generate CoF triplets (question, frame-aware reasoning trace, answer). For this, we leverage Llama3.1-8B with the real videos from VideoEspresso, a manual template with the synthetic videos from Clevrer.

often covering diverse parts of the videos and referencing different sets of frames. An example from CoF-Data$_{\text{real}}$ is illustrated in Fig. 2, and a complete training sample including the video is in Fig. 10.

CoF from synthetic videos (CoF-Data$_{\text{synth}}$). In each frame of Clevrer, every object is annotated with both fixed properties (shape, material, color) and situational attributes (e.g., velocity, location), see Fig. 2. These rich attributes enable the generation of three categories of quantitative questions: *object count*, *appearance order*, and *relative distance*, which complement the semantic questions obtained for real video. Notably, we can generate both questions, answers and chain-of-frames using a fixed manual template, since all the necessary information can be directly deduced from the the object-specific raw annotations. Examples from the categories are shown in Fig. 11 of App. E.

3 Experiments

Models selection. While chain-of-frames is a general approach, we find it particularly well-suited for the recent InternVL models (Chen et al., 2024; Zhu et al., 2025). These models introduce a novel format for videos, where frames are interleaved with text identifiers such as `Frame-1`, `Frame-2`, etc. (see Fig. 6). Details on base model and training configurations can be found in App. B.

Video benchmarks. We compare the video LLMs on five popular benchmarks that capture diverse aspects of video understanding. These benchmarks span a broad range of tasks, video types, and durations, providing a comprehensive evaluation of model capabilities. Details on each benchmark and the breakdown of the results over the fine-grained splits are available in App. C.

3.1 Chain-of-Frames vs other Chain-of-Thoughts variants.

We aim to evaluate how our chain-of-frames approach compares to alternative methods for incorporating reasoning into video LLMs, including prompting, fine-tuning and relevant baselines. To ensure consistency, we use the InternVL2.5-4B as baseline and compare the following models.

- **Original**: the InternVL2.5-4B model with default prompting,
- **Original + CoT Prompting**: the InternVL2.5-4B model with a prompt that encourages the model to perform intermediate reasoning before answering the question (see prompt in App. B).
- **SFT with QA only**: the InternVL2.5-4B model fine-tuned on the question-answers pairs from our CoF-Data without including the reasoning traces.
- **SFT with CoT**: the InternVL2.5-4B model fine-tuned on CoF-Data, where reasoning traces are included but references to specific frames are removed (e.g., *"In Frame 1..."* is replaced with with a generic *"In the video..."*). This approach mimics the standard CoT format.

Table 1: **Chain-of-Frames vs other Chain-of-Thoughts variants.** We compare different approaches to encourage reasoning in video LLMs, via either supervised fine-tuning (SFT) or prompting (see Sec. 3 for details). All models are obtained from InternVL2.5-4B. Fine-tuning on our chain-of-frames (CoF) data yields the best accuracy on all benchmarks.

Model	VSI-BENCH	VIDEO-MME	MVBENCH	VIDHAL	EVENTHALL
Original	31.8	54.9	70.8	74.0	62.5
Original + CoT Prompting	33.5	54.7	71.5	77.0	67.4
SFT with QA only	31.8	54.5	73.4	64.1	57.7
SFT with CoT	34.3	58.6	73.7	77.9	53.1
SFT with CoF (ours)	**36.9**	**59.7**	**76.1**	**79.2**	**71.2**

- **SFT with CoF**: the InternVL2.5-4B model fine-tuned on COF-DATA, i.e. our proposed approach.

For all baseline based on supervised fine-tuning (SFT), we report results using the best prompting strategy (either standard or CoT) for each benchmark. For our SFT with CoF model, we always use CoT prompting across all benchmarks. A complete comparison is provided in App. C.

Results. We report results for all models on the five benchmarks in Table 1. First, we observe that CoT prompting alone already improves accuracy on four out of five benchmarks compared to the original model. Second, fine-tuning on question-answers pairs without reasoning (SFT with QA only) gives mix results, possibly due to overfitting on the training data which might degraded the reasoning ability of the original InternVL2.5-4B. Next, training on reasoning traces without temporal grounding (SFT with CoT) improves the results across nearly all benchmarks, with the notable exception of EVENTHALLUSION. Finally, the model trained on the full COF-DATA (SFT with CoF), i.e., fine-tuned on the reasoning traces including frame references, achieves the highest accuracy across all benchmarks, with improvements ranging from 4.8% to 8.7% over InternVL2.5-4B.

3.2 Comparison to the baselines. In this section, we compare our approach with prior works introduced in the literature (Wang et al., 2024b; Han et al., 2024; Fei et al., 2024; Hu et al., 2025b). Since these models are not publicly available and do not report results on the five benchmarks used in our evaluation, we restrict our comparison to to VIDEO-MME and NEXTQA (Xiao et al., 2021) and based on the results they reported in the paper. As shown in Table 2, our CoF-based models consistently outperform these baselines, even when using smaller model such as CoF-InternVL2.5-4B Notably, compared to the most recent model, M-LLM, our method achieves a substantially larger gain on NEXTQA (4.9% vs. 0.8%), despite starting from a stronger backbone.

4 Conclusion

We have introduced chain-on-frames (CoF), a new approach to encourage video LLMs to produce temporally grounded reasoning before providing answering. Compared to existing works, CoF does not require complex ad-hoc inference frameworks or auxiliary models, and we show that its training data can be efficiently extracted by both real and synthetic videos. Our models fine-tuned on CoF data outperform across multiple benchmarks those obtained with alternative methods for reasoning. Overall, these features make CoF a viable option to further improve the reasoning capabilities of video LLMs. Exploring the effect of increasing the size and diversity of the training data, as well as the scale of the models, represent an exciting direction for future work.

Table 2: **Comparison to the baselines.**

Backbone	Model	VIDEO-MME	NEXTQA
Video-LLaVA-7B	Original	39.9	66.3
	Video-of-Thought	-	76.0 **9.7**↑
Qwen2-VL-7B	Original	58.1	77.6
	M-LLM	58.7 **0.6**↑	78.4 **0.8**↑
InternVL2.5-4B	Original	54.9	75.3
	SFT with CoF (ours)	59.7 **4.8**↑	79.6 **4.3**↑

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 7

Rabiul Awal, Le Zhang, and Aishwarya Agrawal. Investigating prompting techniques for zero-and few-shot visual question answering. *arXiv preprint arXiv:2306.09996*, 2023. 1

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2024. 7

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. Bge m3-embedding: Multilingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation, 2023a. 7

Zhe Chen, Yujie Wang, Yujie Zhang, Chunyuan Li, Yujing Wang, Bo Li, Ziwei Liu, and Chunyuan Li. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238*, 2023b. 7

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. 2, 3, 7, 10

Wey Yeh Choong, Yangyang Guo, and Mohan Kankanhalli. Vidhal: Benchmarking temporal hallucinations in vision llms. *arXiv preprint arXiv:2411.16771*, 2024. 10

Hao Fei, Shengqiong Wu, Wei Ji, Hanwang Zhang, Meishan Zhang, Mong-Li Lee, and Wynne Hsu. Video-of-thought: Step-by-step video reasoning from perception to cognition. In *ICML*, 2024. 1, 2, 4, 7

Chaoyou Fu, Yuhan Dai, Yondong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*, 2024. 10

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024. 7

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. *arXiv preprint arXiv:2411.14794*, 2024. 1, 2, 4, 7

Jian Hu, Zixu Cheng, Chenyang Si, Wei Li, and Shaogang Gong. Cos: Chain-of-shot prompting for long video understanding. *arXiv preprint arXiv:2502.06428*, 2025a. 7

Kai Hu, Feng Gao, Xiaohan Nie, Peng Zhou, Son Tran, Tal Neiman, Lingyun Wang, Mubarak Shah, Raffay Hamid, Bing Yin, et al. M-llm based video frame selection for efficient video understanding. *arXiv preprint arXiv:2502.19680*, 2025b. 4, 7

Yushi Hu, Weijia Shi, Xingyu Fu, Dan Roth, Mari Ostendorf, Luke Zettlemoyer, Noah A Smith, and Ranjay Krishna. Visual sketchpad: Sketching as a visual chain of thought for multimodal language models. In *NeurIPS*, 2024. 1

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. 7

Jihyung Kil, Farideh Tavazoee, Dongyeop Kang, and Joo-Kyung Kim. Ii-mmr: Identifying and improving multi-modal multi-hop reasoning in visual question answering. In *ACL Findings*, 2024. 1

Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. In *NeurIPS*, 2022. 1

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Chunyuan Li, and Ziwei Liu. Vila: On pre-training for visual language models. *arXiv preprint arXiv:2312.07533*, 2023a. 7

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 7

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. Mvbench: A comprehensive multi-modal video understanding benchmark, 2023b. 10

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *ECCV*, 2024. 7

Meta AI. Llama 3.1: Open foundation and instruction-tuned language models. https://huggingface.co/meta-llama/Llama-3.1-8B, 2024. Accessed: 2025-05-13. 2, 8

OpenAI. Openai o1/o3 models. https://openai.com, 2024. 1

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024a. 7

Yan Wang, Yawen Zeng, Jingsheng Zheng, Xiaofen Xing, Jin Xu, and Xiangmin Xu. Videocot: A video chain-of-thought dataset with active annotation tool. *arXiv preprint arXiv:2407.05355*, 2024b. 2, 4, 7

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*, 2022. 1

Penghao Wu and Saining Xie. V*: Guided visual search as a core mechanism in multimodal llms. In *CVPR*, 2023. 1

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *CVPR*, 2021. 4

Yifan Xue, Yuanhan Zhang, Bo Li, Ziwei Liu, and Chunyuan Li. Longvila: Scaling long-context visual language models for long videos. *arXiv preprint arXiv:2408.10188*, 2024. 7

Jihan Yang, Shusheng Yang, Anjali W. Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in Space: How Multimodal Large Language Models See, Remember and Recall Spaces. *arXiv preprint arXiv:2412.14171*, 2024. 10

Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B. Tenenbaum. Clevrer: Collision events for video representation and reasoning. In *ICLR*, 2020. 2

Jiacheng Zhang, Yang Jiao, Shaoxiang Chen, Jingjing Chen, and Yu-Gang Jiang. Eventhallusion: Diagnosing event hallucinations in video llms. *arXiv preprint arXiv:2409.16597*, 2024a. 11

Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. Long context transfer from language to vision. *arXiv preprint arXiv:2406.16852*, 2024b. 7

Yuanhan Zhang, Bo Li, Haotian Liu, Yong Jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. Llava-next: A strong zero-shot video understanding model. https://llava-vl.github.io/blog/2024-04-30-llava-next-video/, 2024c. 7

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Yuchen Duan, Hao Tian, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025. 3

A Related Work

Multimodal LLMs for videos. Multimodal Large Language Models have made substantial progress in integrating visual and textual modalities, enabling them to perform complex reasoning and achieve deep understanding across various data types, including videos (Chen et al., 2023b; Xue et al., 2024; Li et al., 2023a; Zhang et al., 2024b,c; Li et al., 2024; Bai et al., 2024). Among recent advancements, InternVL2.5 (Chen et al., 2024), LLaVA-NeXT-Video (Zhang et al., 2024c), and Qwen2-VL (Wang et al., 2024a) stand out for their strong video understanding capabilities. Both InternVL2.5 and LLaVA-NeXT-Video process videos in an image-text interleaved format, aligning sequences of video frames with language to form a unified multimodal stream. In addition to these open-source models, closed-source systems such as GPT-4o (Achiam et al., 2023) and Gemini-1.5 (Gemini Team et al., 2024) have also demonstrated impressive multimodal capabilities, although details of their architecture and training remain proprietary.

Chain-of-Thoughts for videos. Recent research on chain-of-thoughts for video understanding broadly falls into two categories: methods that provide explicit textual reasoning explanations (Wang et al., 2024b), and those that focus on identifying relevant frames to facilitate the process of generating the response (Han et al., 2024; Hu et al., 2025b,a). In the first category, VideoCoT (Wang et al., 2024b) introduces an active annotation tool to generate reasoning explanations, thereby encouraging models to explicitly reason through visual content. In the second category, several approaches emphasize frame selection as a pre-processing step to enhance reasoning efficiency. M-LLM (Hu et al., 2025b) proposes leveraging multimodal LLMs to identify the most relevant frames corresponding to the query. Likewise, Chain-of-Shot (Hu et al., 2025a) introduces a prompting strategy specifically tailored for understanding long-form videos by selecting key frames. VideoEspresso (Han et al., 2024) combines the ideas of core frame selection with fine-grained reasoning annotations, creating a large-scale dataset that supports more efficient and focused video reasoning. Finally, Fei et al. (2024) introduce Video-of-Thoughts, a complex five-step pipeline to generate the spatial-temporal scene graphs to answer multiple-choice questions about videos.

Limitations of reasoning on videos

Multimodal LLMs (Zhang et al., 2024c; Chen et al., 2024; Bai et al., 2024) process videos as a sequence of images (frames), which are encoded as image tokens by a vision encoder, then concatenated with the user prompt in natural language, and finally passed to a language model. While the base language model may be trained to produce reasoning traces, these are not specific for reasoning on videos.

To encourage chain-of-thought output in video LLMs, models must be fine-tuned on video-grounded reasoning traces. Wang et al. (2024b) introduce VideoCoT which used reasoning traces generated by LLMs and refined by human experts to describe video events. However, the traces lack explicit temporal grounding, that is, individual reasoning steps are not clearly aligned with the corresponding video frames. Additionally, the reliance on human and LLM annotations makes data generation expensive, limiting the dataset to only 11k samples. A different strategy is proposed by Han et al. (2024) who use multiple auxiliary models to generate CoT data. Lightweight multimodal LLMs select the core frames, GPT-4o (Hurst et al., 2024) identifies key elements in each and their relevance to the query, GroundingDINO (Liu et al., 2024) provides spatial annotations (e.g., bounding boxes), and the BGE-M3 retriever (Chen et al., 2023a) produces the temporal annotations. While effective during training, this approach cannot be deployed at inference time given its high complexity and cost (even the frame selection step depends on auxiliary LLMs). Alternatively, Video-of-Thoughts (Fei et al., 2024) introduces a complex five-step inference pipeline, which includes generating the spatial-temporal scene graphs for key frames, and is specialized for multiple-choice questions.

In summary, existing approaches to video reasoning with LLMs face three main limitations: *(i)* expensive training data generation, *(ii)* complex inference, possibly involving auxiliary models, and *(iii)* lack of explicit temporal grounding in the reasoning process.

B Experimental Details

B.1 Chain-of-Frames training data

Frame ID alignment. The original annotations include frame IDs, but, due to context length limitations of video LLMs, we downsample the videos while preserving the frame-annotation alignment. We first map each frame to its timestamp, and clip the video to the maximum duration allowed by the model (e.g., 30 seconds in our experiments) ensuring the segment includes all frames for which captions are available. We then re-calibrate the frame IDs to reflect their new positions within the clipped video.

CoF from real videos (CoF-Data$_{\text{real}}$). To generate question–reasoning–answer triplets, we prompt Llama-3.1-8B (Meta AI, 2024) using the instruction shown in Fig. 3, along with frame-aware video captions from the VideoEspresso dataset (see Fig. 2 for details). Notably, the raw video content is not included in this process. Two examples from CoF-Data$_{\text{real}}$ are shown in Fig. 10.

Prompt

```
Ask a question based on the narrative which is provided for a
    video. The questions should be answerable from the video
    description.
Start reasoning step-by-step like this:
Point out key elements from the video relevant to the question.
Break down the reasoning from those elements to the answer.
Include specific frame numbers as references to support your
    reasoning.
Answer clearly.
**Question**:
**Reasoning**:
**Answer**:
```

Figure 3: **Prompt for CoF-Data$_{\text{real}}$.** We prompt Llama-3.1-8B to generate questions, answers and reasoning traces with reference frames from the real videos of VideoEspresso. Notably, to generate our training data we do not use the videos but only their captions.

CoF from synthetic videos (CoF-Data$_{\text{synth}}$). The second portion of our training dataset is derived from the CLEVRER dataset, which includes detailed attributes for each object in every video frame. Specifically, given a frame ID and object ID, the `inside_camera_key` field indicates whether the object is visible in the frame, enabling us to determine when an object enters or exits the scene. The `velocity` attribute reflects whether an object is moving or stationary, while the `location` attribute provides its absolute or relative position, which can be leveraged to estimate distances or identify collisions. The final CoF-Data$_{\text{synth}}$ dataset comprises three categories of questions: *Object Count*, *Appearance Order*, and *Relative Distance*. Within the *Object Count* category, we define three subtypes: *(i) Collision-Based* ("How many collisions..."), *(ii) Motion State* ("How many moving objects..."), and *(iii) Temporal-Based*, where questions reference specific segments of the video ("After object A enters..."). The questions, answers and reasoning traces are generated with the manually written templates shown in Fig. 4, making the data collection process particularly simple and fast. Additional examples from CoF-Data$_{\text{synth}}$ are shown in Fig. 11.

Object Count Template

```
Question: How many collisions happen in this video?

Reasoning:
1. A collision happens in Frame <frame_id1> between <obj1_name>
   and <obj2_name>
2. ...
Answer:<#collisions> collisions happen in this video.
```

Appearance Order Template

```
Question: what is the appearance order of <object_list> in the
   video?

Reasoning:
1. <obj1_name> appears in Frame {frame_id}
2. ...
Answer: <sorted_object_list>
```

Relative Distance Template

```
Question: Measuring from the closest point of each object, when
   <obj_name_t> <action> the scene,
which of these objects (<all_objects_in_the_scene>) is closest to
   the <obj_name_t>?

Reasoning:
1. <obj_name> <action> the scene in Frame <frame_id>. In Frame
   <frame_id>, the distance between <obj_name_t> and <obj_name_i>
   is <distances[t][i]>.
2. ...
Answer: <obj_name_{min(distances[t])}>
% is the closet object to <obj_name_t>
```

Figure 4: **Templates for CoF-Data$_{\text{synth}}$.** To generate questions, answers and reasoning traces with reference frames from the annotations of the synthetic videos of CLEVRER we rely on fixed, manually written templates. We create three types of questions (object count, appearance order, relative distance) with different templates.

Final dataset (CoF-Data). From the generated chain-of-frames, we filter out samples where frames are referred in the question (as this does not happen in test time). Moreover, we reduce the number of samples with no frame references in the reasoning trace to give higher weight to more complex examples of reasoning. We nevertheless keep a non-negligible fraction of samples with no frame references since there might be, in the evaluation benchmarks, questions which do not require CoF-like reasoning, and we do not want to force the model to generate it when unnecessary. This yields a total of 164,186 samples, comprising 103,683 samples from the CoF-Data$_{\text{real}}$ dataset, which is based on real-world videos, and 60,503 samples from the CoF-Data$_{\text{synth}}$ dataset of synthetic videos. Fig. 5 shows the distribution of how many frames are referenced in the reasoning traces, both for the final dataset and the individual splits. The CoF-Data$_{\text{synth}}$ exhibits a more balanced distribution compared to the automatically generated CoF-Data$_{\text{real}}$: this highlights that using synthetic videos allows us to better control various aspects of the data.

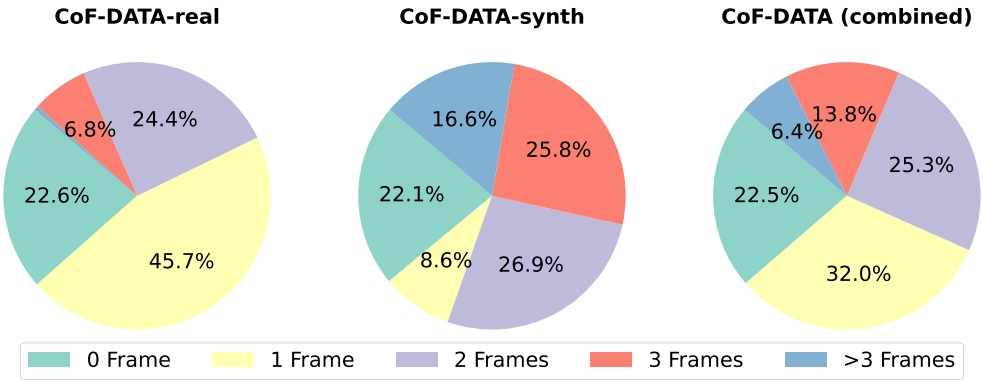


Figure 5: **Distribution of frame references in the Chain-of-Frames training data**. The left pie chart illustrates the distribution for CoF-DATA$_{\text{real}}$, having fewer frames per reasoning trace, whereas CoF-DATA-synth demonstrates a more balanced frame distribution due to controlled synthetic video generation. The right pie chart shows the overall distribution for the CoF-DATA.

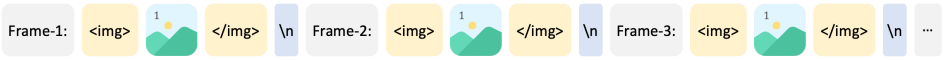


Figure 6: **Video encoding format.** InternVL models add a textual identifier before each frame, which is well-suite for our chain-of-frames reasoning. The illustration is taken from Chen et al. (2024).

B.2 Video benchmarks

VIDEO-MME. VIDEO-MME (Fu et al., 2024) stands out from existing benchmarks by offering a diverse range of video types, covering six primary visual domains and 30 subfields to support broad scenario generalizability. It also introduces variation in temporal length, including short (under 2 minutes), medium (4–15 minutes), and long (30–60 minutes) videos.

MVBENCH. Li et al. (2023b) present a comprehensive benchmark for multi-modal video understanding, encompassing 20 challenging tasks that require more than single-frame analysis. It is specifically designed to evaluate a model's ability to understand temporal dynamics across video sequences.

VSI-BENCH. This benchmark (Yang et al., 2024) is designed to quantitatively assess the visual-spatial intelligence of multimodal large language models. Built from over 5,000 high-quality question-answer pairs across 288 real-world indoor videos, VSI-BENCH spans diverse environments such as homes, offices, and industrial spaces. The benchmark covers eight tasks across three categories. First, the configurational tasks include object count, relative distance, relative direction, and route planning, testing a model's understanding of spatial layout and object relationships. Second, the measurement estimation tasks—object size estimation, room size estimation, and absolute distance estimation—assess a model's ability to infer scale and dimensions, which are crucial for embodied reasoning. Finally, the spatiotemporal task, appearance order, evaluates temporal understanding by requiring models to recall the sequence in which objects or areas appeared throughout the video. Out of the 8 tasks included in this benchmark, relative distance, appearance order, relative directory, and route planning come with multiple-choice questions while the other four require an open-ended quantitative answer. To better evaluate the proximity of model's prediction with the correct answer, Yang et al. (2024) propose the Mean Relative Accuracy (MRA). Given a model's prediction $\hat{y}$ and ground truth y, relative accuracy is calculated by:

$$\mathcal{MRA} = \frac{1}{10} \sum_{\theta \in \mathcal{C}} \mathbb{1} \left(\frac{|\hat{y} - y|}{y} < 1 - \theta \right).$$

where $\mathcal{C} = \{0.5, 0.55, \cdots, 0.95\}$ and denotes a range of confidence thresholds θ to calculate the relative accuracy.

VIDHAL. To evaluate video-based hallucinations in video LLMs, we use VIDHAL (Choong et al., 2024), a multiple-choice benchmark that features video instances drawn from public video

understanding datasets, covering a diverse array of temporal concepts and aspects—such as entity actions and event sequences.

EVENTHALLUSION. Zhang et al. (2024a) introduce EVENTHALLUSION as a binary-choice benchmark designed to systematically assess event-related hallucinations in state-of-the-art Video LLMs. From a hallucination attribution standpoint, it is specifically curated to evaluate a model's susceptibility to language priors and vision-language correlation biases.

CoT Prompting

```
Given a video and a question, Start reasoning step-by-step like
    this:
Point out key frames from the video relevant to the question.
Break down the reasoning from those frames to the answer.
Conclude your reasoning to the answer.

Question: <question>
```

Figure 7: **CoT prompt.** We show the prompt used for elicit reasoning for both the baseline and our fine-tuned models.

B.3 Chain-of-Frames Model

For InternVL2.5-4B, we fully fine-tune both the LLM and the projection modules, keeping the vision encoder frozen. Training is conducted on a single H100 node equipped with 4 GPUs, using a learning rate of 2×10^{-6}, a batch size of 2, and a single epoch.

C Additional Experiments

Effect of CoT prompting. An extended version of Table 1 is presented in Table 3. For all baselines, we report results using two prompting strategies, either standard (indicated by ⋆) or chain-of-thought (indicated by ♣, the prompt is shown in Fig. 7). For our SFT with CoF model, we consistently use CoT prompting. When considering InternVL2.5-4B, CoT prompting alone improves accuracy of the original models on four out of five benchmarks compared to the original model. However, this improvement does not hold for the SFT with QA only variant: we hypothesize that fine-tuning solely on QA data negatively impacts the reasoning capabilities of the baseline model. On the other hand, incorporating reasoning traces into the training data (SFT with CoT) generally enhances the model's reasoning capabilities, and using CoT prompting is beneficial except for the EVENTHALLUSION benchmark. Finally, our models (SFT with CoF) outperform the baseline across all benchmarks.

Table 3: **Effect of CoT prompting.** For all baselines, we report results using two prompting strategies, i.e. standard (indicated by ⋆) and chain-of-thought (indicated by ♣), while we fix CoT prompting for our CoF models.

Model	Prompt	VSI-BENCH	VIDEO-MME	MVBENCH	VIDHAL	EVENTHALL
InternVL2.5-4B						
Original	⋆	31.8	54.9	70.8	74.0	62.5
	♣	33.5	54.7	71.5	77.0	67.4
SFT with QA only	⋆	31.8	55.4	70.3	73.6	63.1
	♣	31.8	54.5	73.4	64.1	57.7
SFT with CoT	⋆	31.1	52.6	69.6	74.4	62.5
	♣	34.3	58.6	73.7	77.9	53.1
SFT with CoF (ours)	♣	36.9	59.7	76.1	79.2	71.2

Detailed results over benchmark splits. For completeness, we report the fine-grained results over the various splits of VSI-BENCH (Table 4), VIDEO-MME (Table 5), MVBENCH (Table 6)

Table 4: **Detailed results on the VSI-BENCH benchmark.** For the baselines, we report results using two prompting strategies, i.e. standard (indicated by ⋆) and chain-of-thought (indicated by ♣), while we fix chain-of-thoughts prompting for our CoF models.

Model	Prompt	Obj. Count	Abs. Dist.	Obj. Size	Room Size	Rel. Dist.	App. Order	Rel. Dir.	Route Plan	Avg
InternVL2.5-4B										
Original	⋆	29.2	31.2	45.5	20.4	35.4	23.2	41.4	27.8	31.8
	♣	36.0	17.1	38.1	29.8	34.2	30.3	52.2	29.9	33.5
SFT with QA only	⋆	22.7	30.7	44.0	26.0	36.3	22.2	39.4	33.0	31.8
	♣	34.9	18.6	38.8	23.2	37.0	28.2	47.1	26.3	31.8
SFT with CoT	⋆	31.5	22.0	41.6	27.9	36.8	21.2	40.4	27.3	31.1
	♣	39.1	19.5	36.1	26.9	36.1	30.1	57.1	29.4	34.3
SFT with CoF (ours)	♣	42.5	20.8	36.4	29.4	35.4	32.4	62.2	36.1	36.9

and EVENTHALLUSION (Table 7). Moreover, for the baseline models we report results using two prompting strategies, i.e. standard (indicated by ⋆) and chain-of-thought (indicated by ♣).

D Additional analyses of Chain-of-Frames

Influence of training data. Our training data COF-DATA combines chain-of-frames from both real and synthetic videos (total of 164k samples). To evaluate the impact of dataset diversity, we constructs dataset of the same size (164k samples) but from a single source, i.e. either COF-DATA$_{\text{real}}$ or COF-DATA$_{\text{synth}}$, and fine-tune InternVL2.5-4B on them. We compare these two models to standard CoF-InternVL2.5-4B (trained on COF-DATA) in Fig. 8. Using the combined dataset outperforms single-source datasets on all benchmarks except for EVENTHALLUSION, demonstrating the importance of diversity in the reasoning traces used for training. Between the two individual datasets, the model trained on synthetic videos outperforms the one trained on real videos in three out of five cases while being worse just in one. This suggests that further improvements could come from expanding the tasks covered by the synthetic datasets, considering the low cost of data generation.

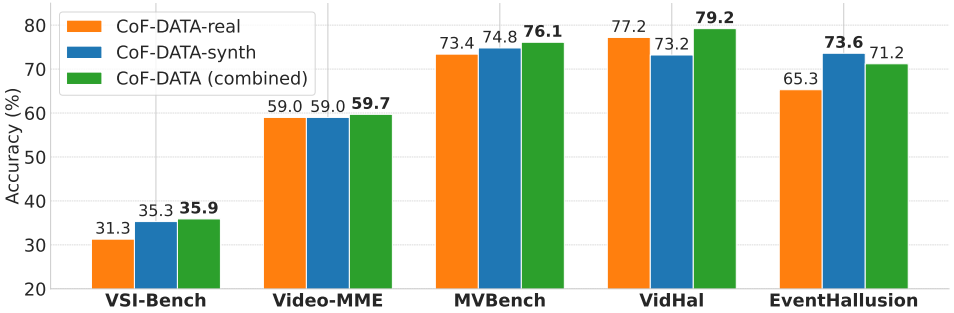


Figure 8: **Diversity of training data.** Combining CoF examples extracted from real and synthetic videos provides better results on almost all benchmarks compared to using a single source. For all models we use 164k training samples, and fine-tune an InternVL2.5-4B model.

CoF reasoning at inference time. To better understand how the CoF models utilize frame references during inference, we track the number of frames referenced in the reasoning trace per answer. In Fig. 9, we report how many answers, generated by our CoF-InternVL2.5-4B model across all evaluation benchmarks, contain N frames for $N = 1, \ldots, 10$. We can see that both models learn to produce, from a relatively small training set, reasoning traces which refer to frame IDs (see qualitative examples in App. E). In particular, CoF-InternVL2.5-4B includes at least one frame reference in 68.7% of the

Table 5: **Detailed results on the VIDEO-MME benchmark.** For the baselines, we report results using two prompting strategies, i.e. standard (indicated by ⋆) and chain-of-thought (indicated by ♣), while we fix chain-of-thoughts prompting for our CoF models.

Model	Prompt	Short (900)	Medium (900)	Long (900)	Avg
InternVL2.5-4B					
Original	⋆	64.9	52.7	47.2	54.9
	♣	64.0	53.2	47.0	54.7
SFT with QA only	⋆	68.0	53.6	44.8	55.5
	♣	66.8	53.1	43.6	54.5
SFT with CoT	⋆	64.3	51.8	41.8	52.6
	♣	70.4	55.7	49.6	58.6
SFT with CoF (ours)	♣	73.1	56.2	49.9	59.7
InternVL3-8B					
Original	⋆	73.0	61.7	52.1	62.3
	♣	75.3	65.3	59.0	66.6
SFT with CoF (ours)	♣	80.9	74.4	70.4	75.3

Table 6: **Detailed results on the MVBENCH benchmark.** For the baselines, we report results using two prompting strategies, i.e. standard (indicated by ⋆) and chain-of-thought (indicated by ♣), while we fix chain-of-thoughts prompting for our CoF models.

Model	Prompt	AA	AC	AL	AP	AS	CO	CI	EN	FA	MA
InternVL2.5-4B											
Original	⋆	89.5	54.0	44.0	75.0	82.9	62.5	79.0	29.0	46.0	97.5
	♣	88.5	50.0	46.5	77.0	81.4	67.0	78.0	34.5	60.5	99.0
SFT with QA only	⋆	90.0	55.0	44.0	76.5	81.9	63.5	75.0	33.0	46.0	98.5
	♣	90.5	50.0	48.5	78.5	84.0	67.5	80.0	38.5	71.5	99.5
SFT with CoT	⋆	87.0	53.0	35.5	76.5	81.4	62.0	76.5	31.5	43.5	98.0
	♣	90.5	46.5	57.5	85.0	83.5	67.0	80.5	38.5	73.5	98.5
SFT with CoF (ours)	♣	93.0	41.0	62.0	91.5	89.4	73.5	79.5	47.0	83.0	98.5

Model	Prompt	MC	MD	OE	OI	OS	ST	SC	UA	Avg
InternVL2.5-4B										
Original	⋆	88.5	73.0	96.5	83.5	39.5	92.0	57.5	85.0	70.8
	♣	86.5	72.5	96.0	81.5	40.5	91.5	58.0	78.0	71.5
SFT with QA only	⋆	87.5	75.0	96.5	82.5	41.0	91.5	59.5	85.5	70.3
	♣	86.5	75.5	96.5	86.5	42.0	92.0	52.5	82.0	73.4
SFT with CoT	⋆	89.0	72.5	96.5	82.0	38.0	91.5	56.5	82.0	69.6
	♣	86.5	72.0	95.5	86.5	40.5	92.0	52.5	80.5	73.7
SFT with CoF (ours)	♣	86.5	72.0	96.5	87.0	44.0	93.5	50.0	82.5	76.1

answers. Moreover, the CoF model can use frame references selectively rather than uniformly across tasks (distribution by benchmark in App. C), depending on the type of the questions.

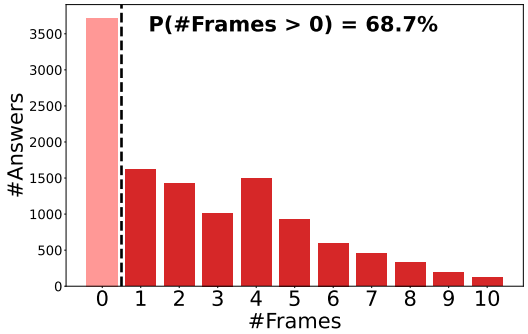


Figure 9: **CoF reasoning at inference time.** For CoF-InternVL2.5-4B (left plot), we show the distribution of how the number of frames generated at inference time: the video LLMs fine-tuned on CoF-DATA learn to produce chain-of-frames during evaluation.

E Additional Figures

This section presents additional samples from our training dataset along with inference examples. More specifically, Fig. 10 and Fig. 11 show samples from the COF-DATA$_{real}$ and COF-DATA$_{synth}$, respectively. To demonstrate the reasoning traces generated by our CoF models, we present the samples from VSI-BENCH and VIDHAL benchmarks in Fig. 12.

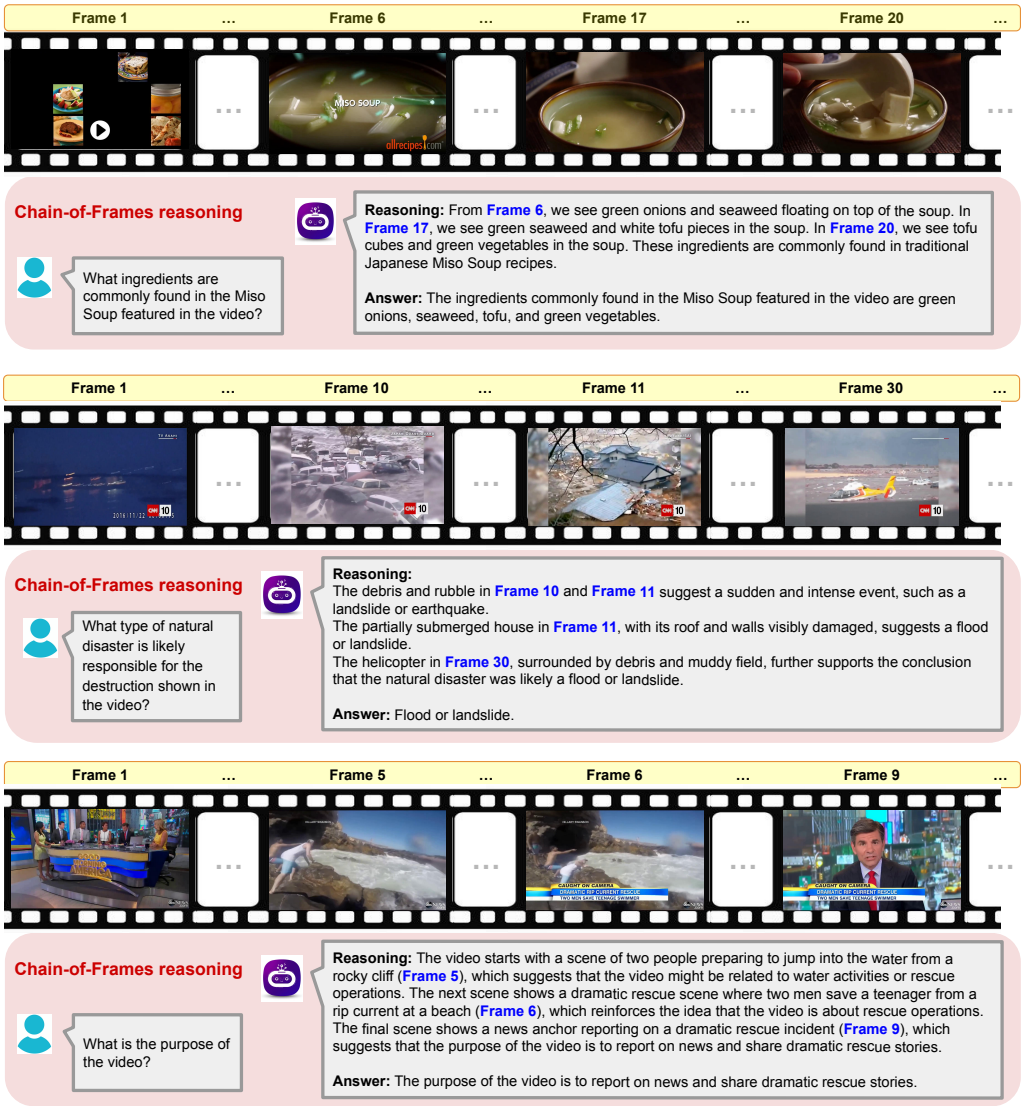


Figure 10: **CoF training data.** Examples of the training data from COF-DATA$_{real}$, including relevant frames, questions, answers, and reasoning traces.

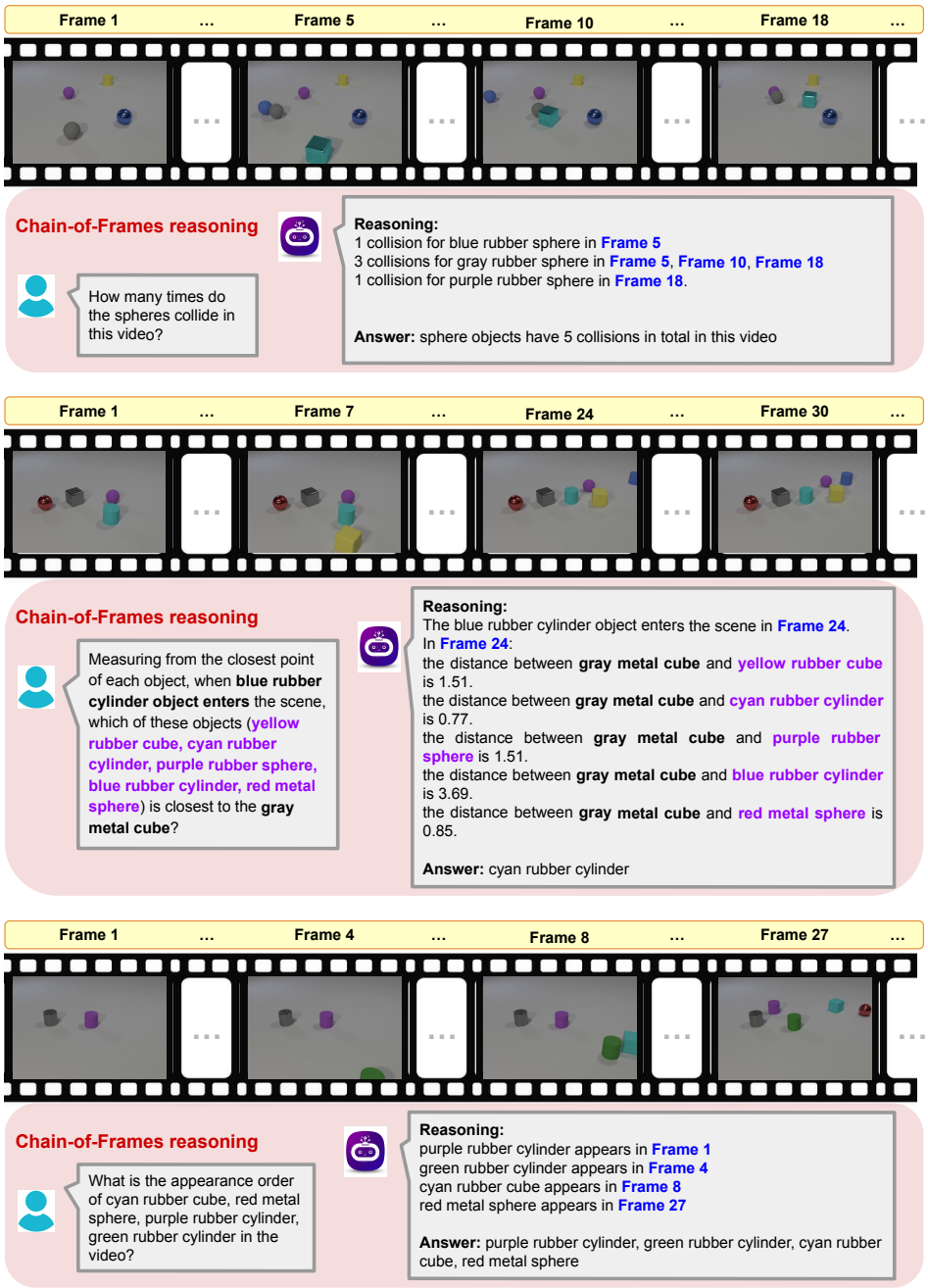
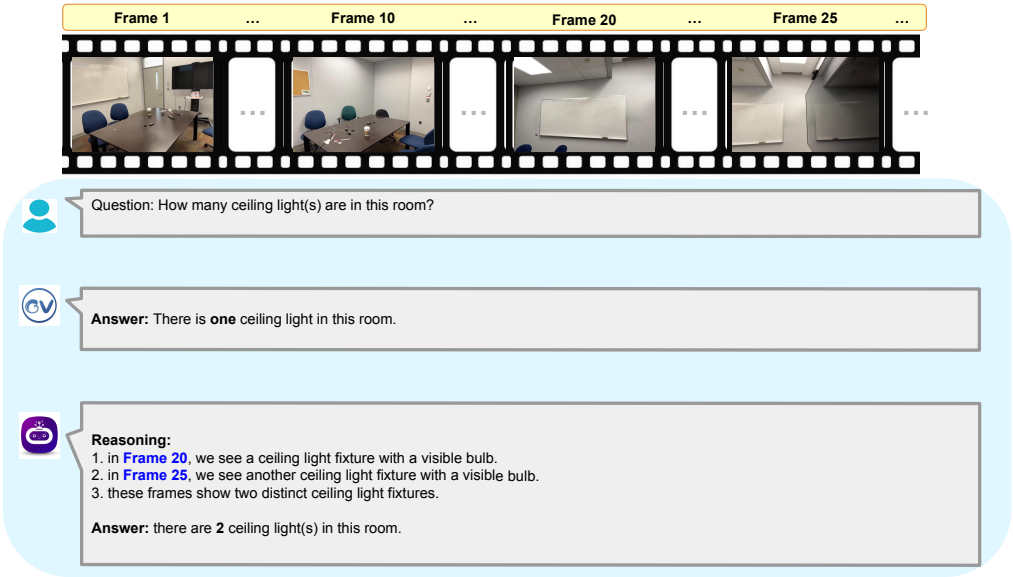
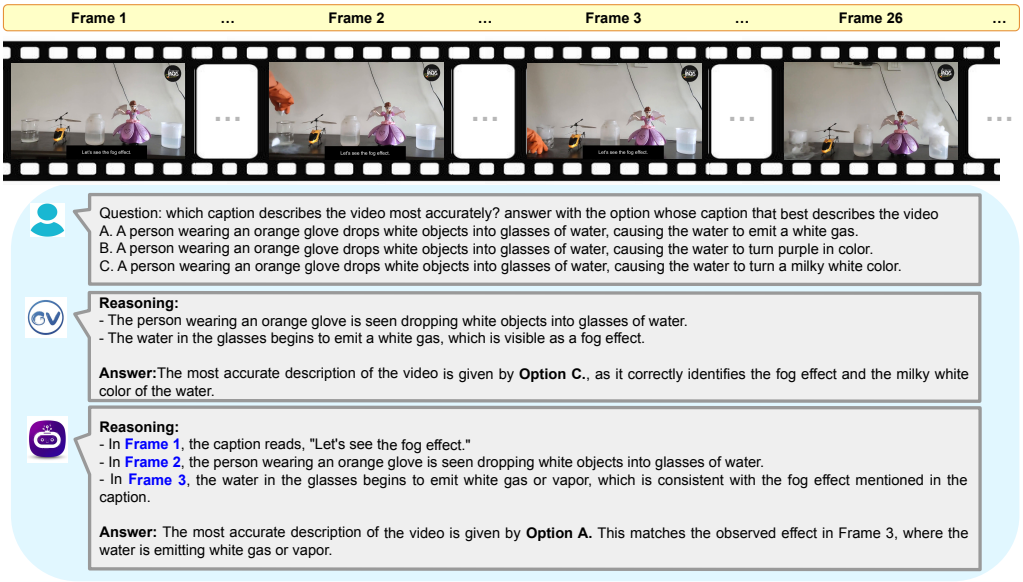


Figure 11: **CoF training data.** Examples of training data generated from the COF-DATA$_{\text{synth}}$ dataset, including relevant frames, questions, answers, and reasoning traces. The samples shown belong to the *Object Count* and *Relative Distance* categories, respectively.



(a) **VSI-BENCH benchmark.** We show the question (first box), the answer and possibly CoT reasoning of the original InternVL2.5-4B with CoT prompting (second box), and the answer with CoF reasoning of our CoF-InternVL2.5-4B model (third box).



(b) **VIDHAL benchmark.** We show the question (first box), the answer and possibly CoT reasoning of the original InternVL2.5-4B with CoT prompting (second box), and the answer with CoF reasoning of our CoF-InternVL2.5-4B model (third box).

Figure 12: **Inference examples.**

Table 7: **Detailed results on the VIDHAL and EVENTHALLUSION benchmarks.** For the baselines, we report results using two prompting strategies, i.e. standard (indicated by ⋆) and chain-of-thought (indicated by ♣), while we fix chain-of-thoughts prompting for our CoF models.

Model	Prompt	VIDHAL	EVENTHALLUSION			
			Entire	Misleading	Mix	Avg
InternVL2.5-4B						
Original	⋆	74.0	48.3	91.2	48.2	62.5
	♣	77.0	44.7	80.4	77.2	67.4
SFT with QA only	⋆	73.6	48.3	89.2	51.8	63.1
	♣	64.1	47.4	75.5	50.3	57.7
SFT with CoT	⋆	74.4	49.1	91.2	47.1	62.5
	♣	77.9	39.5	71.6	48.2	53.1
SFT with CoF (ours)	♣	79.2	49.1	85.3	79.3	71.2