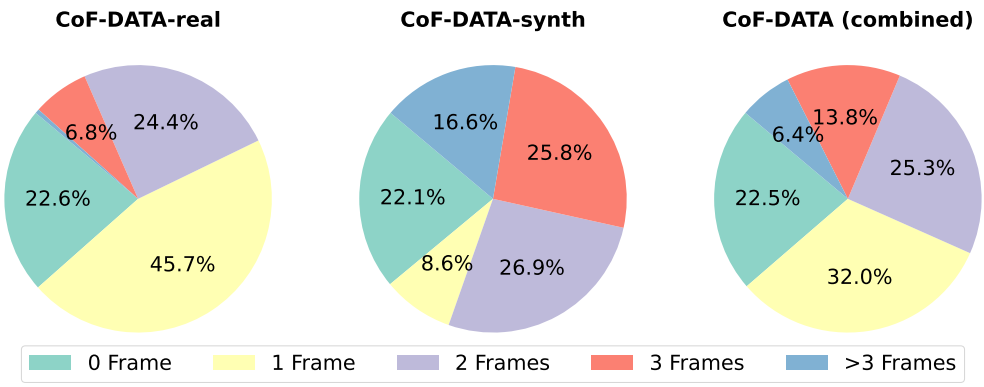


Figure 8. **Distribution of frame references in the Chain-of-Frames training data**. The left pie chart illustrates the distribution for CoF-DATA$_{\text{real}}$, having fewer frames per reasoning trace, whereas CoF-DATA-synth demonstrates a more balanced frame distribution due to controlled synthetic video generation. The right pie chart shows the overall distribution for the CoF-DATA.

A. Experimental Details

A.1. Chain-of-Frames training data

CoF from real videos (CoF-DATA$_{\text{real}}$). To generate question-reasoning-answer triplets, we prompt Llama-3.1-8B-Instruct [27] using the instruction shown in Fig. 9 along with frame-aware video captions from the VideoEspresso dataset (see Fig. 3 for details). Notably, the raw video content is not included in this process. Two examples from CoF-DATA$_{\text{real}}$ are shown in Fig. 13.

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. 10, making the data collection process particularly simple and fast. Examples from CoF-DATA$_{\text{synth}}$ are shown in Fig. 14.

Final dataset (CoF-DATA). Our training data, CoF-DATA, has 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. 8 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-CoF-DATA$_{\text{real}}$: this highlights that using synthetic videos allows us to better control various aspects of the data.

A.2. Video benchmarks

Video-MME. Video-MME [10] offers 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. [26] presents a comprehensive benchmark for multimodal 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 [38] 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: object count, relative distance, relative direction, route planning, object size estimation, room size estimation, absolute distance estimation, and appearance order. Out of these 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 the model's prediction with the correct answer, [38] proposes using mean relative accuracy ($\mathcal{MRA}$). 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 [8], 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. [40] introduce EVENTHALLUSION, from which we use the binary-choice questions designed to systematically assess event-related hallucinations in 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.

A.3. Chain-of-Frames model

CoF-InternVL. For InternVL2.5-4B, we fully fine-tune both the LLM and the projection modules, keeping the vision encoder frozen. In contrast, for InternVL3-8B, we adopt LoRA-based fine-tuning [15] to reduce memory consumption. All other training configurations remain consistent across both models. 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.

CoF-Phi-3.5-Vision-4B. To test the generalizability of our approach beyond the InternVL family, we employ Phi-3.5-Vision-4B [1], a mobile-scale multimodal LLM that demonstrates strong performance in language reasoning, as a third baseline model.

Phi-3.5-Vision-4B uses distinct, indexed image placeholder tokens such as <image$-i$>, instead of repeating a generic <image> token, because the model must uniquely align each image embedding with a specific position in the text sequence. Each placeholder is a separate token in the tokenizer, allowing the model to reliably map image embeddings to their corresponding locations and to correctly understand references to the images.

Phi-3.5-Vision's capability to understand references to the images makes it a good candidate for our base model. Similar to InternVL2.5-4B, we fully fine-tune both the LLM and the projection modules, keeping the vision encoder frozen. The other training configurations remain consistent across all baselines.

B. Additional Experiments

CoF-Phi-3.5-Vision-4B results. Tab. 4 compares the baseline Phi-3.5-Vision-4B model with its CoF-enhanced variant. Across all benchmarks, CoF-Phi-3.5-Vision-4B consistently surpasses the corresponding base model. The largest improvements appear on the VSI-BENCH benchmark, where CoF-Phi-3.5-Vision-4B achieves better performance compared to LLaVA-OneVision-7B and Qwen2-VL-7B (Tab. 1) despite using substantially fewer parameters. These results demonstrate that our chain-of-frames approach generalizes effectively across architectures and model families.

Effect of CoT prompting. An extended version of Tab. 3 is presented in Tab. 4. 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. 11). For our SFT with CoF models, we always use CoT prompting. When considering InternVL2.5-4B, CoT prompting alone improves the 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. CoT prompting improves the results also for the original InternVL3-8B on most benchmarks. Finally, our models (SFT with CoF) outperform the baseline across all benchmarks.

Detailed results over benchmark splits. For completeness, we report the fine-grained results over the various splits of VSI-BENCH (Tab. 5), VIDEO-MME (Tab. 6), MVBENCH (Tab. 7), and EVENTHALLUSION (Tab. 8). Moreover, for the baseline models, we report results using two prompting strategies, i.e., standard (indicated by ⋆) and chain-of-thought (indicated by ♣).

Detailed results of CoF reasoning at inference time. In Fig. 12, we show statistics of how many frames are referenced in the reasoning traces generated by CoF-InternVL3-8B. To complement Fig. 7, we report the frequency for each benchmark separately. We see that the number of frames mentioned varies across benchmarks, e.g., the cases where no frames are referenced significantly decreases on the hallucination benchmarks VIDHAL and EVENTHALLUSION.

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
InternVL3-8B						
Original	⋆	41.0	62.3	72.0	80.9	72.1
	♣	40.2	66.5	74.3	61.6	73.9
SFT with CoF (ours)	♣	51.3	73.7	77.1	79.5	78.7
Phi-3.5-Vision-4B						
Original	⋆	26.6	50.2	48.7	52.7	55.2
	♣	26.4	39.0	46.6	43.1	51.3
SFT with CoF (ours)	♣	32.8	52.6	51.8	55.5	59.2

Table 4. **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.

This suggests that our CoF models learn to modulate the reasoning traces and the frame references depending on the task.

C. Additional Figures

This section presents additional samples from our training dataset along with inference examples. More specifically, Fig. 13 and Fig. 14 show samples from the COF-DATA$_{\text{real}}$ and COF-DATA$_{\text{synth}}$, respectively. To illustrate the reasoning traces generated by our CoF models and compare them to the answers of the baseline models, we present samples from VSI-BENCH and MVBENCH benchmarks in Fig. 15 and samples from hallucination benchmarks in Fig. 16.

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
InternVL3-8B										
Original	⋆	58.6	28.5	49.5	43.3	47.0	38.5	31.3	31.4	41.0
	♣	55.2	29.5	38.1	32.6	42.8	47.1	47.5	28.4	40.2
SFT with CoF (ours)	♣	61.8	34.2	37.7	26.7	66.8	43.4	83.9	55.7	51.3
Phi-3.5-Vision-4B										
Original	⋆	27.3	23.4	30.3	21.4	26.5	17.5	33.7	33.0	26.6
	♣	27.3	24.1	32.6	22.6	20.5	20.9	35.2	28.3	26.4
SFT with CoF (ours)	♣	31.5	27.9	37.9	26.3	33.5	22.7	47.3	35.1	32.8

Table 5. **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.

Prompt

```
Ask a question based on the narrative that 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 9. **Prompt for COF-DATA$_{real}$.** 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.

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 10. **Templates for CoF-DATA$_{synth}$.** 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.

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

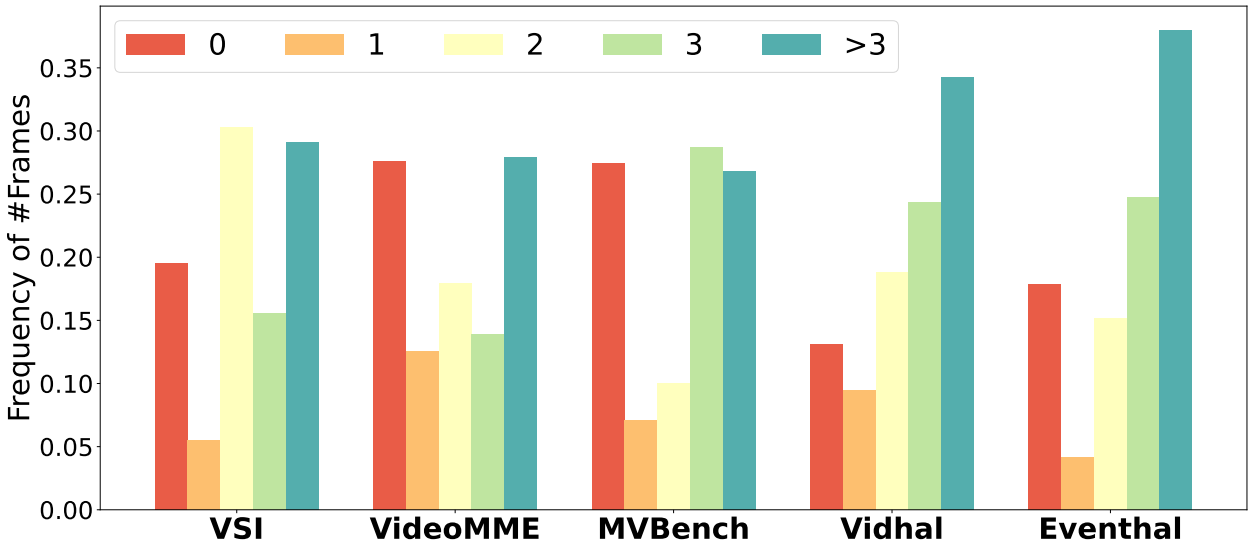


Figure 12. **CoF reasoning at inference time.** For each benchmark, we show the frequency of the number of frames referenced in the reasoning traces of CoF-8B.

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)	♣	79.3	71.7	70.0	73.7
Phi-3.5-Vision-4B					
Original	⋆	61.4	50.6	38.7	50.2
	♣	46.0	34.9	36.1	39.0
SFT with CoF (ours)	♣	63.1	51.9	42.7	52.6

Table 6. **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	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
InternVL3-8B											
Original	⋆	90.0	42.0	44.5	83.0	82.45	75.5	78.0	38.5	45.0	98.0
	♣	77.5	45.0	42.5	87.5	85.1	80.5	89.0	33.5	47.0	99.0
SFT with CoF (ours)	♣	96.5	50.5	49.5	89.5	91.0	91.5	77.5	45.5	59.5	96.5
Phi-3.5-Vision-4B											
Original	⋆	65.0	52.5	32.0	45.5	40.96	48.5	36.5	40.0	36.0	65.5
	♣	61.0	49.5	27.5	43.0	37.8	42.0	35.0	36.5	36.5	61.5
SFT with CoF (ours)	♣	68.0	51.5	38.5	49.5	42.02	49.5	39.5	44.0	38.5	68.0

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
InternVL3-8B										
Original	⋆	60.5	89.0	96.97	85.5	39.5	92.5	69.0	81.0	71.7
	♣	63.5	89.0	97.47	87.5	41.5	92.5	73.0	73.5	72.5
SFT with CoF (ours)	♣	67.0	91.0	90.4	90.5	45.5	94.0	78.0	84.0	77.1
Phi-3.5-Vision-4B										
Original	⋆	35.0	37.0	57.1	45.0	31.0	88.5	46.0	75.0	48.7
	♣	35.5	38.5	53.0	43.5	28.5	89.0	47.5	72.5	46.6
SFT with CoF (ours)	♣	36.5	45.5	61.1	46.5	33.5	90.0	53.3	76.5	51.8

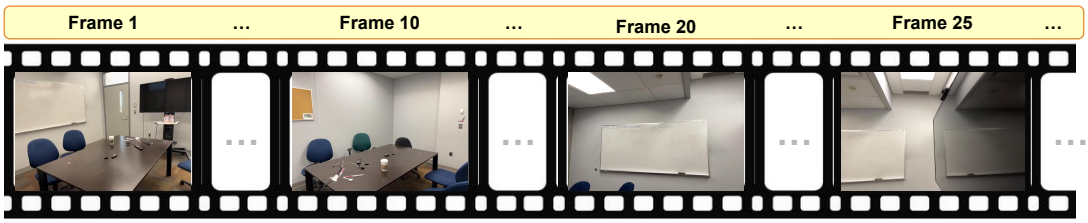
Table 7. **Detailed results on the MVBENCH benchmark.** For the baselines, we report results using two prompting strategies, i.e. standard (indicated by ⋆) and chain-of-thought (indicated by ♣), while we fix chain-of-thoughts prompting for our CoF models.

Model	Prompt	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
InternVL3-8B						
Original	⋆	80.9	52.6	91.2	72.5	72.1
	♣	61.6	57.0	94.1	70.5	73.9
SFT with CoF (ours)	♣	79.5	57.9	92.2	86.0	78.7
Phi-3.5-Vision-4B						
Original	⋆	52.7	27.2	72.6	65.8	55.2
	♣	43.1	22.8	64.7	66.3	51.3
SFT with CoF (ours)	♣	55.5	34.2	73.5	70.0	59.2

Table 8. **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.

Figure 13. **CoF training data.** Examples of the training data from CoF-DATA$_{\text{real}}$, including relevant frames, questions, answers, and reasoning traces.

Figure 14. **CoF training data.** Examples of training data generated from the CoF-Data$_{synth}$ dataset, including relevant frames, questions, answers, and reasoning traces. The samples shown belong to the *object count* and *relative distance* categories, respectively.



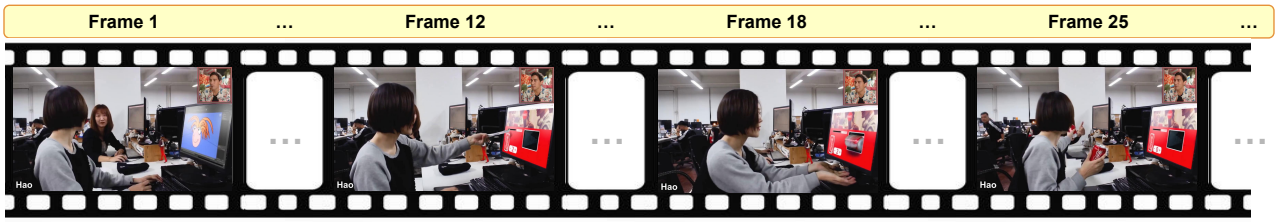
Question: How many ceiling light(s) are in this room?

Answer: There is **one** ceiling light in this room.

Reasoning:
1. in **Frame 20**, we see a ceiling light fixture with a visible bulb.
2. in **Frame 25**, we see another ceiling light fixture with a visible bulb.
3. these frames show two distinct ceiling light fixtures.

Answer: there are **2** ceiling light(s) in this room.

(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).



Question: What makes the actions seen in the video impossible in real life?
(A) The woman can actually put money into the computer screen and get a drink.
(B) The computer screen can transform into a vending machine.
(C) Physical objects cannot enter or leave the computer screen in reality.
(D) Physical objects are capable of passing through digital barriers.

Reasoning:
The video shows show a computer screen acting as a vending machine, which is not possible in real life. The woman is able to put money into the screen and receive a drink, which is only possible in the digital world.

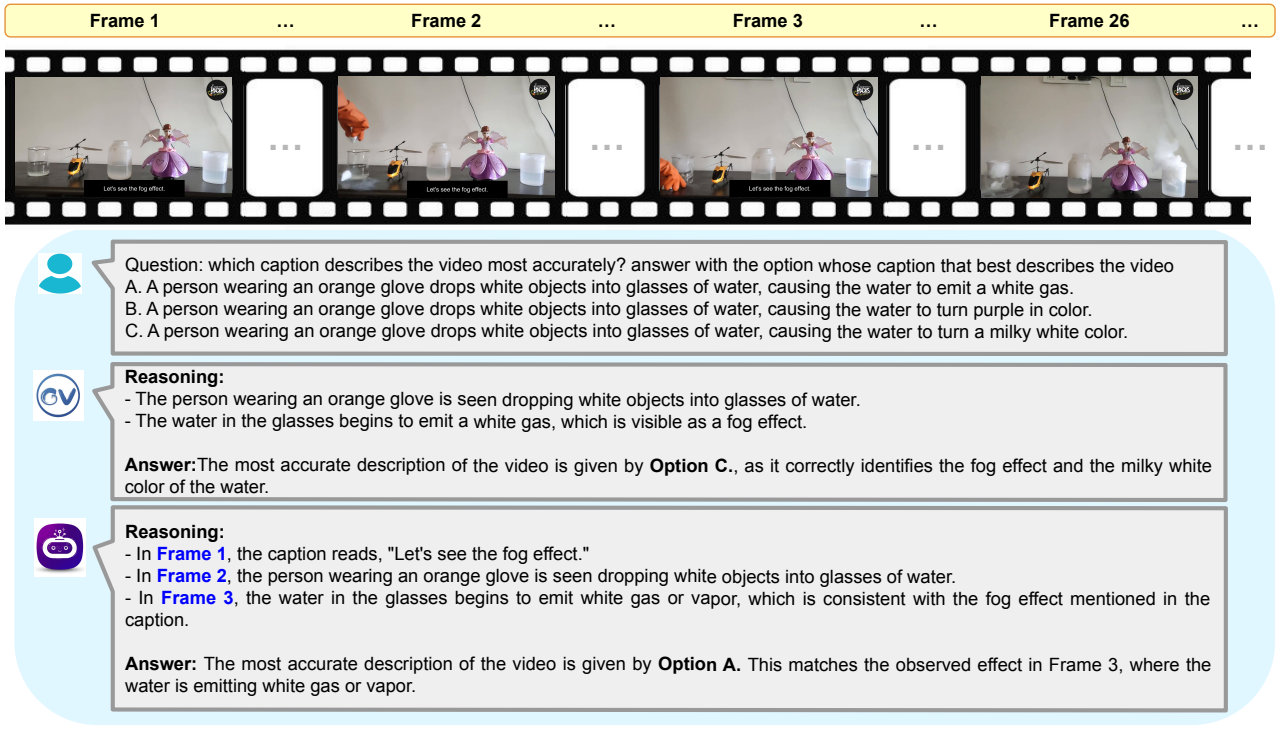
Answer: (B) The computer screen can transform into a vending machine.

Reasoning:
In **Frame 18**, the woman is shown putting a can into the computer screen, which is not possible in real life.
Break down the reasoning from those frames to the answer.
The key frame in **Frame 18** shows the woman putting a can into the computer screen, which is not possible in real life. This suggests that the actions seen in the video are not possible in real life because physical objects cannot enter or leave the computer screen.

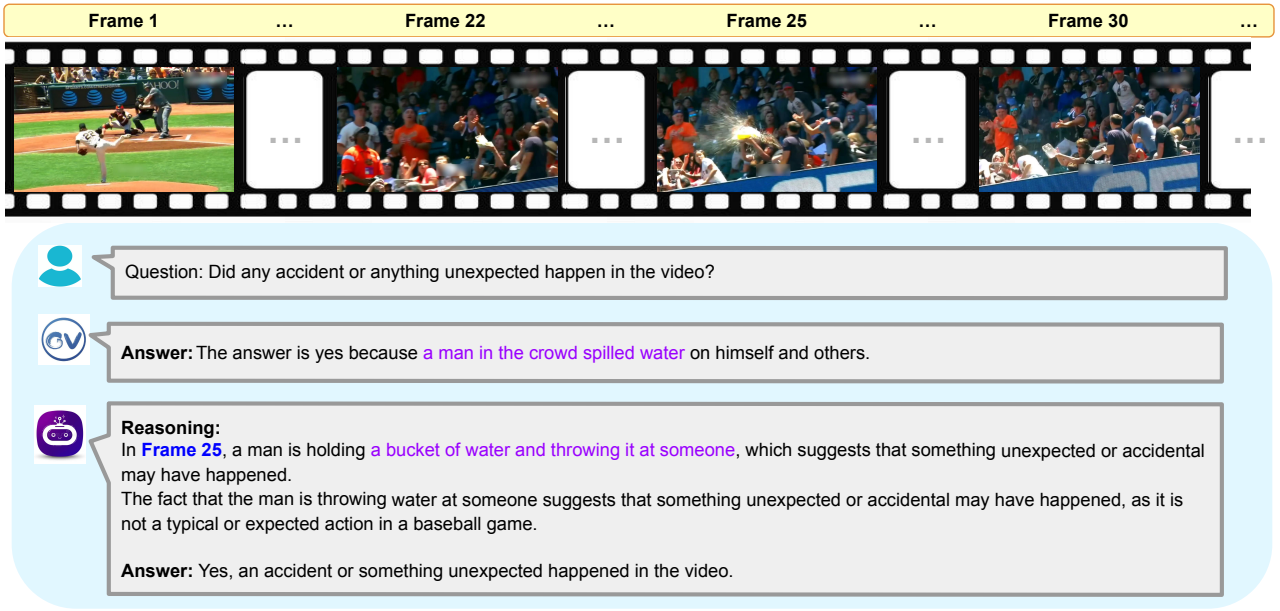
Answer: (C) Physical objects cannot enter or leave the computer screen in reality.

(b) **MVBench benchmark.** We show the question (first box), the answer and possibly CoT reasoning of the original InternVL3-8B with CoT prompting (second box), and the answer with CoF reasoning of our CoF-InternVL3-8B model (third box).

Figure 15. **Inference examples.**



(a) **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).



(b) **EVENTHALLUSION benchmark.** We show the question (first box), the answer, and possibly CoT reasoning of the original InternVL3-8B with CoT prompting (second box), and the answer with CoF reasoning of our CoF-InternVL3-8B model (third box).

Figure 16. **Inference examples.**