

VIDEOP2R: Video Understanding from Perception to Reasoning

Supplementary Material

We provide additional details and illustrations for our main content in the following section:

- Process-Aware CoT Generation (Sec. 7).
- Evaluation Setup (Sec. 8)
- Ablation Study on Judge Model (Sec. 9)
- Perception Effectiveness Experiment (Sec. 10)
- RL Training Trend (Sec. 11).
- Think-Answer Mismatch (Sec. 12).
- More Qualitative Results (Sec. 13).
- Impact of Model Size and Dataset Composition (Sec. 14)

7. Details of Process-Aware CoT Generation and Data Analysis

7.1. Prompt Used

Figure 7 illustrates the prompt template for process-aware CoT generation. We employ Qwen2.5-VL-72B-Instruct with a temperature of 0 for the generation.

```
Prompt Template for Process-aware CoT Generation

{Question}

You are required to answer the question using the visual content provided. Your output must include three parts: <observation>, <think>, and <answer>. Do not stop after the first part. Follow the instructions below for each section:

Step 1: Visual Observation
• First, analyze what the question is asking.
• Identify what type of visual information is necessary to answer it (e.g., objects, actions, text, spatial relations, etc.).
• Write a single paragraph that begins with: "Based on the question, I need to notice ..."
• Then directly continue with your visual observations in the same paragraph.
• Ensure that the motivation ('I need to notice...') and the actual observations are logically connected.
• Wrap this section between <observation> and </observation> tags.

<observation>
Based on the question, I need to notice ... [Continue your observation here in the same paragraph.]
</observation>

Step 2: Reasoning
• Use your observations to reason step by step toward the answer.
• Begin with: "Based on the observations, ..."
• Engage in internal dialogue (e.g., 'let me think', 'hmm', 'wait a second', etc.).
• It's encouraged to include self-reflection or verification, such as: 'Does that make sense?', 'Let me doublecheck...', etc.
• Wrap your reasoning between <think> and </think> tags.

<think>
Based on the observations, ... [Insert your reasoning here.]
</think>

Step 3: Final Answer
• Provide a concise and specific final answer.
• Ensure it is logically consistent with both the observations and the reasoning.
• Wrap your answer between <answer> and </answer> tags.

<answer>
[Insert your final answer here.]
</answer>
```

Figure 7. Prompt Template for Process-aware CoT Generation. We use the same prompt for training and inference.

Figure 8 illustrates the prompt template for observation sufficiency verification. We use Claude 3.7 to judge the sufficiency of the observation segment.

7.2. Data Source and Metric for CoT Verification

Our data source [14] encompasses five distinct question types to enhance the model’s flexibility and its generalization across diverse tasks and formats: (1) Multiple

Prompt Template for Observation Sufficiency Verification

You will receive a question, its correct answer, and a set of observations extracted from {data_type}. Your task is to decide—strictly based on the observations—whether the correct answer can be deduced. Do **not** assume any information beyond what is provided.

```
## Question
{Question}

## Correct Answer
{CorrectAnswer}

## Observations
{Observations}

## Instruction
1. Carefully read the question and the correct answer.
2. Briefly explain whether (and how) the observations support the correct answer.
3. Finally output your judgement, either <judgement>Yes</judgement> or <judgement>No</judgement>.
```

Figure 8. Prompt Template for Observation Sufficiency Verification. We use the same prompt for perception correctness judgment in RL stage.

Choice, (2) Numerical QA, (3) OCR, (4) Free-form QA, and (5) Regression. Each data sample includes a question, data source, correct answer, and optional choices when applicable (e.g., for multiple-choice questions).

In the subsequent CoT Verification stage, task-specific accuracy metrics are adopted to assess annotation reliability, and samples below a preset threshold of 0.6 are filtered out. The task-specific metrics are listed as follows:

- **Multiple Choice**: 1 if the predicted option matches the ground truth; 0 otherwise.
- **Numerical QA**: 1 for exact match with the reference value.
- **OCR**: reward based on Word Error Rate (WER) between prediction and reference.
- **Free-form QA**: reward is the average of ROUGE-1, ROUGE-2, and ROUGE-L scores.
- **Regression**: reward = 1 – relative error between prediction and ground truth.

These task-specific metrics are also used for computing the reasoning accuracy reward in the RL stage.

7.3. Data Statistic

Adapting our generation pipeline to the data source yields VIDEOP2R-CoT-162K, consisting of 162,062 image and video visual QA pairs with high-quality annotations on perception and reasoning. We present the data statistics in Tab. 3. The dataset covers both image and video modalities, and spans multiple question types including multiple-choice, numerical, OCR, free-form, and regression. Multiple-choice questions constitute the majority, providing stable evaluation signals, while the inclusion of numerical, OCR, and free-form questions introduces diverse reasoning skills such as count-

ing, reading, grounded description, and open-ended inference. This heterogeneous composition enables comprehensive assessment of process-aware perception and reasoning across modalities.

To analyze our constructed data, we visualize the embedding distributions (using UMAP [35]) of perception and reasoning annotations (Fig. 9). The two clusters are clearly separated, indicating that our annotated data inherently distinguishes perception and reasoning.

7.4. Word Count and Word Cloud Analysis

Figure 10 presents the word length distribution and word cloud visualization of VIDEOP2R-CoT-162K. As shown on the left of the figure, perception and reasoning annotations exhibit a comparable number of words across the entire annotation set, suggesting a balanced contribution of both processes. The word clouds further highlight the intrinsic difference in focus between the two processes: perception annotations are dominated by video-centric terms such as “video”, “person”, and “observing”, reflecting their emphasis on factual and descriptive content; in contrast, reasoning annotations frequently contain introspective expressions such as “double check” and “make sense”, which indicate deeper reflective reasoning.



Figure 9. Embeddings visualization of VIDEOP2R-CoT-162K

7.5. Annotation Examples

We provide annotation examples in Figs. 11 and 12 to illustrate how our annotations explicitly separate perception from reasoning. Figure 11 presents a video QA example where the perception segment successfully captures the key visual cue (the zigzag pattern), and the reasoning segment then derives the correct answer based on this evidence. Figure 12 presents an image QA example in which the perception segment accurately extracts the numerical information from the table, and the reasoning segment performs the required mathematical reasoning

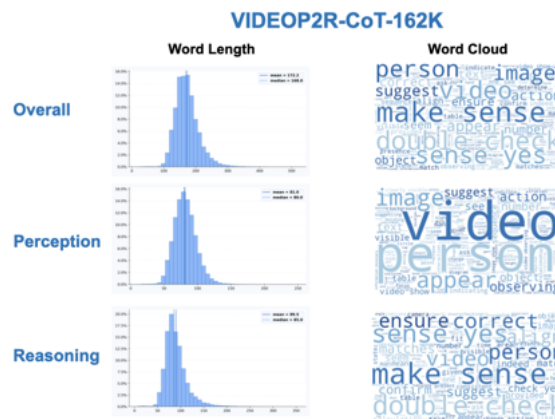


Figure 10. Word length (Left) and Word cloud (Right) Visualization for VIDEOP2R-CoT-162K.

over these numbers, followed by validation to double-check the final answer.

Question: In which direction does the person pour the oil into the wok?

Options: A. In a diagonal line. B. In a circular motion. C. In a zigzag pattern. D. In a straight line.

<observation>Based on the question, I need to notice the direction in which the oil is poured into the wok. Observing the video, the oil is being poured in a way that creates a zigzag pattern across the surface of the wok. This is evident from the way the oil spreads and the lines it forms as it hits the wok's surface. **</observation>**

<think>Based on the observations, the oil is being poured in a zigzag pattern. Let me think, this pattern is distinct and not a straight line, diagonal line, or circular motion. The zigzag pattern is clearly visible as the oil spreads across the wok. Does that make sense? Yes, it does. Let me double-check the options provided, and the zigzag pattern matches option C. **</think>**

<answer>C</answer>

Figure 11. An Annotation Example of the Video QA Sample

Question: What is the total number of computers in all the classrooms combined?

Answer: 55

Computers in classrooms	
Teacher	Number of computers
Miss Murphy	20
Mr. Schmidt	10
Ms. Rojas	19
Ms. Cohen	1
Mr. Watson	5

<observation>Based on the question, I need to notice the number of computers in each classroom as listed for each teacher. The table provides the necessary data: Miss Murphy has 20 computers, Mr. Schmidt has 10, Ms. Rojas has 19, Ms. Cohen has 1, and Mr. Watson has 5. These numbers are essential for calculating the total number of computers across all classrooms. **</observation>**

<think>Based on the observations, I need to add up the number of computers for each teacher to find the total. Let me think: 20 (Miss Murphy) + 10 (Mr. Schmidt) + 19 (Ms. Rojas) + 1 (Ms. Cohen) + 5 (Mr. Watson) equals 55. Does that make sense? Let me double-check the addition: 20 + 10 is 30, plus 19 is 49, plus 1 is 50, and plus 5 is indeed 55. The calculation seems correct. **</think>**

<answer>55</answer>

Figure 12. Annotation Example of the Image QA Sample

Table 3. Distribution of question types across VIDEOP2R-CoT-162K.

	Question Type					Sum
	Multiple Choice	Numerical	OCR	Free-form	Regression	
Image	47,091	18,476	4,014	2,501	693	72,775
Video	86,910	1,371	–	1,006	–	89,287
Sum	134,001	19,847	4,014	3,507	693	162,062

8. Experiment Set up

8.1. Implementation Details

The whole two-stage training is conducted on 8× NVIDIA A100 GPUs. For efficiency, we limit the video input to 16 frames at a resolution of $128 \times 28 \times 28$ during training, where 28×28 denotes the patch size and 128 the number of patches. For the SFT stage, we use a batch size of 8 with gradient accumulation = 2. For the RL stage, we adopt a batch size of 56^1 (8 rollouts per sample). We use Claude 3.7 Sonnet for observation judgment to speed up the training process. We note that since video token processing dominates training time, the additional text-only judgment introduces little overhead: GRPO takes 16.5 hours for 1k steps, while VIDEOP2R takes 18 hours on the same hardware.

During inference, we increase the number of frames and resolution to 32 and $256 \times 28 \times 28$, respectively, and apply the decoding configuration (top-p = 0.001, temperature = 0.01) consistent with the Qwen2.5-VLA official demo. During both training and inference, we adopt the same prompt (Fig. 7) as in process-aware CoT generation, and use the prompt shown in Fig. 8 for perception accuracy judgment.

8.2. Main Table Evaluation Setup

This section introduces the evaluation benchmarks used in Tab. 1 and the evaluation metrics. We selected seven widely used video understanding and reasoning benchmarks to provide a comprehensive analysis of VideoP2R:

- VSI-Bench [62] is a video-based benchmark designed to evaluate models’ visual–spatial reasoning capability. It includes two types of questions: (1) numerical and (2) multiple-choice. Numerical questions are evaluated using Mean Relative Accuracy (MRA), while multiple-choice questions are evaluated using Accuracy (ACC). Following the original benchmark protocol, we report the overall performance as **the average of MRA and ACC**.
- VideoMMU [20] is a multi-modal, multi-

disciplinary video benchmark, designed to evaluate models’ ability to acquire and apply knowledge from expert-level lecture videos. In our experiments, models are evaluated with **accuracy** over all questions.

- MMVU [72] is an expert-level multi-disciplinary video understanding benchmark aimed at assessing models’ capability to perform domain-specific reasoning across diverse scientific and technical fields. In our experiments, models are evaluated with **accuracy** over **all multiple-choice questions** (1858 of 3000).
- VCR-Bench [40] is a benchmark crafted to assess video Chain-of-Thought reasoning. VCR-Bench selected and integrated data from multiple existing video benchmarks. In our experiments, models are evaluated with **accuracy** over all **multiple-choice questions** (510 of 1034).
- MVBench [25] is a multi-modal video understanding benchmark designed to stress test models’ temporal reasoning capabilities across diverse domains. In our evaluation, models are assessed using **accuracy** on multiple-choice QA derived from temporally grounded tasks.
- TempCompass [32] is a temporal reasoning benchmark designed to dissect video LLMs’ ability to perceive dynamic changes over time. It constructs paired videos that share identical static content but differ in temporal aspects (e.g., speed, direction) to prevent shortcut solutions based on static frames. In our evaluation, models are measured by **accuracy** over temporal reasoning questions under the official protocols.
- Video-MME [15] is a comprehensive multi-modal evaluation benchmark for video-centric large language models, designed to assess their analysis capabilities across diverse video types and modalities. We evaluate using the official metrics and configuration, reporting **accuracy** over the QA pairs **without subtitles**.

For all result numbers of *Open-Source Models* in Tab. 1, we use the reported number in the original paper. For all result numbers of *RFT Models*, we run the evaluation locally. We additionally include Qwen2.5-VL-72B in Tab. 8 as an upper bound for our model. While VIDEOP2R still trails Qwen2.5-VL-72B on average, it

¹To accelerate training, we integrate vLLM [23] for sampling, assigning one GPU exclusively for sample generation and the remaining seven for model updates.

significantly boosts the base model (Qwen2.5-VL-7B) and even outperforms Qwen2.5-VL-72B on MVBench, underscoring the effectiveness of our approach.

9. Ablation Study on Judge Model

Table 4 presents the results of using different judge models for perception correctness judgement. We conduct the same two-stage training process, but only change the Claude3.7 to Llama3.1 [11] families for providing perception correctness judgement.

Compared with the base model, all VIDEOP2R variants using different judge models achieve consistent improvements, confirming the effectiveness of perception reward supervision. To further assess judge reliability, we randomly annotate the perception correctness of 200 samples with human labels and evaluate each judge’s decision accuracy. We observe a clear upward trend in accuracy as the judge model becomes larger and more capable. Additionally, the fact that even Llama3.1-8B attains reasonable reliability on this relatively simple perception correctness judgement suggests that perception correctness can be robustly handled by current LLMs, and our pipeline is broadly applicable across a wide range of judge models. Moreover, the positive correlation between judge capability and the downstream performance of the trained model indicates that stronger judges provide more reliable perception feedback and lead to larger gains, with Claude3.7 achieving the highest agreement with human annotations and the best overall process-aware performance.

10. Details of the Perception Examination

10.1. Prompt Used and Detailed Set up

The perception examination experiment involves three types of experiments on either text or video domains. We compare the zero-shot performance of Qwen2.5-VL-7B across different input settings and examine how perception segments influence its answers: (i) performance on text-only questions, (ii) performance with both text and video inputs, and (iii) performance when the text-only prompt is augmented with a perception segment generated by VIDEOP2R or Qwen2.5-VL-7B. We used the prompt “*Prompt for Qwen Inference*” (Fig. 13 Top) for (i) and (ii). The prompt “*Prompt for Qwen Inference with Perception Segment*” (Fig. 13 Bottom) is used for (iii). For (iii), We use the same prompt in Fig. 7 to get the perception segment from VIDEOP2R or Qwen2.5-VL-7B first and then augment the segment within the prompt for inference.

10.2. Full Results

Table 5 presents the full results of the perception examination experiment (Fig. 4).

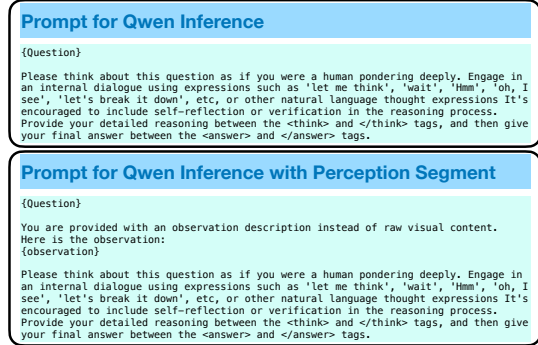


Figure 13. Prompt Template for Perception Examination Experiment.

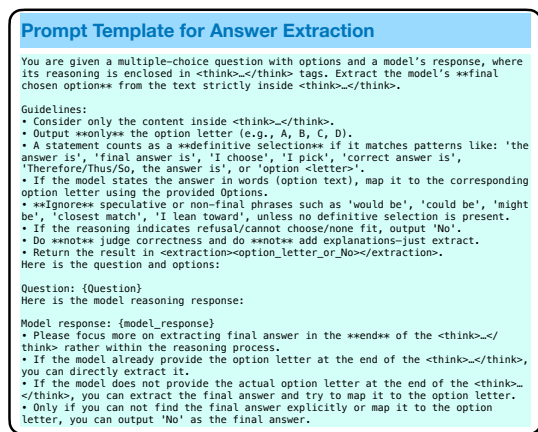


Figure 14. Prompt Template for Answer Extraction.

10.3. Examples of Qwen Inference Output

We present examples of Qwen’s outputs under different configurations in our perception examination experiment in Fig. 21. When given only the text question (Top Left), Qwen fails to perform meaningful reasoning due to the absence of video information and resorts to guessing from the answer choices. When conditioned on the text question plus Qwen’s own perception segment (Top Right), it mentions some relevant visual cues (e.g., the person placing books into the backpack) but omits critical details such as the exact number of books, resulting in unreliable reasoning. Even with access to both the text question and the video input (Bottom Left), Qwen still produces inaccurate perception, confusing the top pocket with the main compartment; this misperception propagates into the reasoning process and yields an incorrect answer. In contrast, the perception segment generated by VIDEOP2R is clear and sufficient (Bottom Right), explicitly capturing both the number of books and their correct placement. This improvement can be attributed to PA-GRPO, which guides the perception target to information directly relevant to answer-

Table 4. Ablation studies of VIDEOP2R on judge models. “Judge Acc.” reports perception decision accuracy on 200 human-labeled samples. Best results within each group are in **bold**.

Model	Judge Acc.	Video Reasoning				Video Understanding			Avg
		VSI.	VideoMMMU	MMVU	VCR.	MV.	TempCom.	VideoMME	
• Base Models									
Qwen2.5-VL(7B)	–	30.1	48.1	60.0	44.3	59.0	72.6	56.6	52.9
• Judge Model									
VIDEOP2R (Llama3.1-8B)	82	39.0	52.2	64.0	49.2	64.7	73.8	59.2	57.4
VIDEOP2R (Llama3.1-70B)	88	35.8	52.4	64.6	50.2	65.0	74.5	60.5	57.6
VIDEOP2R (Llama3.1-405B)	91	38.2	54.4	64.5	49.2	66.5	75.0	58.4	58.0
VIDEOP2R (Claude3.7)	95	36.8	55.0	65.4	51.0	68.1	74.5	60.0	58.7

Table 5. Comparison of Qwen2.5-VL-7B with and without perception segments augmentation (video/text modalities).

Model	Video Reasoning				Video Understanding			Avg
	VSI.	VideoMMMU	MMVU	VCR.	MV.	TempCom.	VideoMME	
• Video + Text								
Qwen2.5-VL(7B)	30.1	48.1	60.0	44.3	59.0	72.6	56.6	52.9
Qwen2.5-VL(7B) + Obs.	29.7	54.1	61.5	49.4	65.4	73.4	58.8	56.0
Qwen2.5-VL(7B) + Obs.(Qwen)	30.4	53.1	61.5	48.6	61.6	72.8	56.4	54.9
• Text								
Qwen2.5-VL(7B)	22.1	34.4	45.8	33.5	30.9	46.2	33.1	35.1
Qwen2.5-VL(7B) + Obs.	33.8	51.5	61.0	48.4	63.6	72.9	57.5	55.5
Qwen2.5-VL(7B) + Obs.(Qwen)	28.8	45.8	60.0	41.5	49.1	67.3	48.0	48.6
VIDEOP2R	36.8	55.0	65.4	51.0	68.1	74.5	60.0	58.7

ing the question, thereby enabling Qwen to arrive at the correct answer even without direct access to the video.

11. RL Training Dynamics

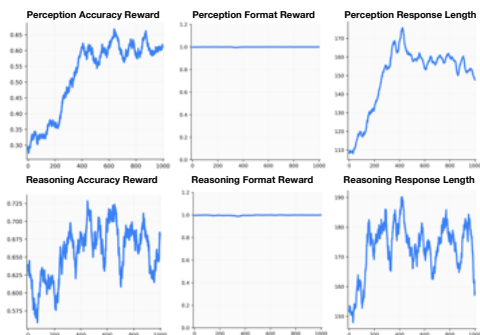


Figure 15. RL training Dynamics of VIDEOP2R

We provide the full RL training dynamics of VIDEOP2R in Fig. 15 to comprehensively illustrate our RL stage. Both the perception accuracy reward and the reasoning accuracy reward exhibit an overall increasing trend, indicating that the model progressively improves its ability to produce correct perception and reasoning traces. The perception format reward and reasoning for-

mat reward remain close to 1 throughout training, showing that the model consistently adheres to the process-aware inference template and maintains stable format compliance. Since the length reward is conditioned on both accuracy and format rewards, we instead visualize the lengths of the perception and reasoning segments during RL. We observe an initial increase followed by a decrease in both segments, indicating that the model adaptively adjusts its outputs and eventually converges to concise yet sufficiently informative perception and reasoning traces.

12. Think-Answer Mismatch Analysis

12.1. Pilot Experiment

We conduct a pilot experiment with Claude 3.7 to assess the reliability of answer extraction using the prompt in Fig. 14. We first sample 400 responses to multiple-choice questions and ask human annotators to extract the model’s answers from the `<think>` segments. The annotators achieve 95% agreement, with most disagreements arising from cases where the reasoning is unclear and the model appears to guess the answer. Using the same setting, Claude 3.7 reaches 96.5% accuracy, confirming the reliability of this assessment.

12.2. Think-Answer Mismatch Example

Figure 16 presents an example of Think-Answer Mismatch, where the model conducts correct reasoning in `<think>` but produces an incorrect final answer in `<answer>`. This mismatch highlights how relying solely on final-answer rewards can reinforce unfaithful or inconsistent behavior, underscoring the necessity of process-aware rewards in PA-GRPO.

Think-Answer Mismatch Example

Question: For the video example, if we change student 7's Test 2 score from 79 to 89, what will be the new t value.

Options:
 A. -1.030
 B. 1.020
 C. -0.500
 D. -1.250
 E. -0.950
 F. -0.300
 G. 0.750
 H. 1.030
 I. -1.500
 J. -0.600

<think>To solve this problem, we need to recalculate the t-value after changing the test score for student 7 from 79 to 89. The t-value is calculated using the formula:

$$T = \frac{\bar{d}}{SE(\bar{d})}$$
 where:
 \bar{d} is the mean of the differences,
 $SE(\bar{d})$ is the standard error of the difference.
 First, let's calculate the new mean of the differences and the new standard error of the difference.
 1. Calculate the new mean of the differences:
 - Original differences: -7, 5, 12, 8, 10, -6, -11
 - New difference for student 7: -11 + 10 = -1
 - New differences: -7, 5, 12, 8, 10, -6, -1
 - Mean of new differences: $\bar{d} = \frac{-7 + 5 + 12 + 8 + 10 + (-6) + (-1)}{7} = \frac{19}{7} \approx 2.71$
 2. Calculate the new standard error of the difference:
 - Original standard deviation of differences: $s_d = 9.325$
 - New standard deviation of differences: $s_d' = \sqrt{\frac{\sum(d_i - \bar{d})^2}{n-1}}$
 - Sum of squared differences: $((-7)^2 + 5^2 + 12^2 + 8^2 + 10^2 + (-6)^2 + (-1)^2) = 49 + 25 + 144 + 64 + 100 + 36 + 1 = 419$
 - New standard deviation: $s_d' = \sqrt{\frac{419}{6}} \approx 9.05$
 - Standard error: $SE(\bar{d}') = \frac{s_d'}{\sqrt{n}} = \frac{9.05}{\sqrt{7}} \approx 3.44$
 3. Calculate the new t-value:
 - New t-value: $T = \frac{2.71}{3.44} \approx 0.79$
The closest option to 0.79 is 0.750. **</think>**

<answer>H**</answer>**

Figure 16. Example of Think-Answer Mismatch.

12.3. Think-Answer Mismatch Results

Table 6² reports Think-Answer Mismatch rates across benchmarks. All results are computed on the multiple-choice subsets of each benchmark.


13. More Qualitative Results of VIDEOP2R

13.1. Success Case


We provide two additional success cases of VIDEOP2R in Fig. 19. In the left example, VIDEOP2R effectively tracks key visual information throughout the video: in the `<observation>` segment, it identifies the three positions where the yellow clothing is presented, supporting subsequent reasoning. In the right example, VIDEOP2R accurately captures relevant visual cues, including the person's gestures and the background context, and leverages them to produce the correct final answer.

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

Answer: 2



Key-frame Missing



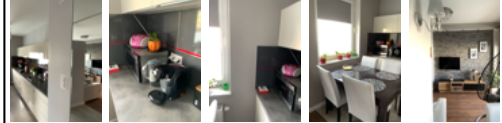
<observation>Based on the question, I need to notice the number of chairs present in the room. Observing the video, I see a room with a table and a chair. The chair is positioned next to the table, and there are no other chairs visible in the room. The room also includes a sofa, a coffee table, and a kitchen area, but these do not contribute to the count of chairs. The focus is on identifying the chairs, and only one chair is clearly visible in the video. The room appears to be a living space with minimal furniture, and the chair is the only piece of furniture that fits the description of a chair. The video does not show any additional chairs in the background or any other areas of the room. **<observation>**
<think>Based on the observations, the room contains only one chair. Let me think, the question specifically asks for the number of chairs in the room. The video clearly shows one chair next to the table, and there are no other chairs visible in the room. Does that make sense? Let me double-check the video again to ensure there are no other chairs hidden or partially visible. After rechecking, it is confirmed that there is only one chair in the room. The other furniture items like the sofa and coffee table do not count as chairs. **<think>**
<answer>1**</answer>**

Figure 17. Failure Cases of Key-frame Missing in VIDEOP2R

Question: If I am standing by the stool and facing the sofa, is the tv to my front-left, front-right, back-left, or back-right?

The directions refer to the quadrants of a Cartesian plane (if I am standing at the origin and facing along the positive y-axis)?

Options: A. front-right, B. back-left, C. back-right, D. **front-left**



<observation>Based on the question, I need to notice the spatial relationship between the stool, the sofa, and the TV. Observing the video, a table is positioned on my left, holding an oven, a blender, and a green plant. To the left of this table is a window with the curtain half open, and several plants placed along the window threshold; nothing notable is visible outside. Another table is placed close to the window, accompanied by four white chairs and four bowl mats with speckled decoration. The stool is positioned near the center of the room, and the sofa is located towards the back-right corner of the room. The TV is mounted on the wall to the left of the sofa, which means it is not directly in front of the sofa but rather to its left. **<observation>**
<think>Based on the observations, the TV is to the left of the sofa, which means it is in the back-left position relative to the sofa. Let me think, if I am standing by the stool and facing the sofa, the TV would be to my back-left. Does that make sense? Yes, it does. Let me double-check the options, and A. back-left seems to be the correct choice. **<think>**
<answer>A**</answer>**

Figure 18. Failure Cases of Overly detailed visual configuration in VIDEOP2R

13.2. Failure Case

We identify two representative types of failure cases for VIDEOP2R, illustrated in Fig. 17 and Fig. 18. (1) **Key-frame missing.** During inference, we adopt uniform frame sampling, which may omit question-critical key frames containing essential visual evidence [6]. As shown in Fig. 17, the question asks for the number of chairs in the room. However, the sampled frames only include the view where one chair appears on the left side of the table, while the key frame showing another chair on the right side is missed. In this case, the question becomes unsolvable given the incomplete observations. Increasing the number of sampled frames can mitigate this issue, and an adaptive sampling strategy [19] can further reduce the risk of missing question-

²VersaVid-R1 has too few available traces on VSI-Bench for meaningful statistics.

Table 6. Think–Answer Mismatch rates across models and benchmarks (lower is better).

Model	Video Reasoning				Video Understanding			Avg
	VSI.	VideoMMMU	MMVU	VCR.	MV.	TempCom.	VideoMME	
Qwen2.5-VL(7B)	4.9	12.2	3.5	6.3	7.8	4.6	5.1	6.3
Video-R1 (SFT)	2.0	2.7	3.2	1.0	0.6	0.5	0.8	1.6
VideoRFT (SFT)	5.4	4.2	9.1	5.0	7.3	1.6	3.6	5.2
VideoP2R (SFT)	0.9	4.2	0.9	1.0	1.4	0.9	1.2	1.5
Video-R1	25.5	19.0	18.8	26.7	21.7	14.5	24.3	21.5
VideoChat-R1	13.6	11.9	11.9	13.3	12.4	7.3	11.3	11.7
VersaVid-R1*	–	21.5	11.1	18.6	16.7	9.5	15.6	15.5
VideoRFT	23.5	22.9	12.9	18.5	15.1	10.9	14.4	16.9
VIDEOP2R	6.8	9.6	7.9	7.6	7.4	6.4	7.1	7.5

critical evidence, which we leave for future work. **(2) Overly detailed visual configuration.** The second failure type arises when questions require tracking an excessive number of fine-grained visual details, which is particularly common in VSI-Bench [62]. Questions in VSI-Bench demand precise modeling of object layouts and relative positions across multiple regions, often exceeding the length targets ([128,320]) used during VIDEOP2R’s training. When the required descriptions fall outside this familiar length regime, the model tends to compress or drop critical perceptual details, leading to incomplete observations and subsequent reasoning errors. As shown in Fig. 18, the question targets the relative positions among the stool, sofa, and TV, while accurately specifying this configuration also requires the locations of surrounding reference objects, such as the table near the stool and the table next to the window. VIDEOP2R initially exhibits high-quality perception (e.g., correctly identifying the tables and window) but gradually introduces errors for later elements (e.g., the stool and sofa), which ultimately leads to failure in subsequent reasoning. We propose to have a more dynamic length reward system [51] in future work.

13.3. From Base Model to VIDEOP2R: A Stepwise Capability Evolution

We present a representative example in Fig. 20 to illustrate how the model’s perception and reasoning capabilities evolve from the base Qwen2.5-VL-7B to VIDEOP2R-SFT and finally to the RL-optimized VIDEOP2R. The question asks where the cat stays for the longest time. In the video the cat briefly starts on the stool and then spends the remaining time on the robot’s thigh. The base Qwen2.5-VL-7B only captures the early details where the cat is on the stool and misses the later details on the robot’s thigh, leading to an incorrect answer. VIDEOP2R-SFT attends to both the stool and the robot’s thigh but misjudges the relative durations.

In contrast, VIDEOP2R produces a comprehensive and faithful perception trace that correctly tracks the cat’s locations and time spent, even explicitly ruling out distractors such as the carpet and nest, thereby enabling reliable downstream reasoning. This example demonstrates that each stage (SFT, RL) of the VIDEOP2R framework is both effective and necessary, enabling stable, long-term improvement in process-aware perception and reasoning.

13.4. Impact of Task Difficulty

We analyze VIDEOP2R’s adaptability across difficulty levels using the fine-grained categorizations of our benchmarks Sec. 8.2. VideoMMMU [20] spans three reasoning difficulties from basic knowledge identification (*Identify*) to interpretation (*Interpret*) and adaptation (*Adapt*), while Video-MME [15] categorizes perception difficulty by video duration (*Short*, *Medium*, *Long*). As shown in Tab. 7, VIDEOP2R obtains its largest gains on the more challenging splits: +3.66% on *Interpret* (VideoMMMU) and +3.24% on *Medium* videos (Video-MME), while the smallest gains occur on the easiest splits (+2.31% on *Identify*, +1.66% on *Short*). These results indicate that our method is most beneficial for challenging tasks, while providing moderate gains for simpler tasks that already fall within the model’s intrinsic capability boundary.

14. Impact of Model Size and Dataset Composition

We report the full results in Tab. 8, including VIDEOP2R’s adaptation to a smaller model. VIDEOP2R also scales effectively to smaller architectures: applying our pipeline to Qwen2.5-VL-3B improves average accuracy from 38.4% to 53.3%, surpassing even the Qwen2.5-VL-7B baseline (52.9%).

18:44To demonstrate that VIDEOP2R-CoT-162K is

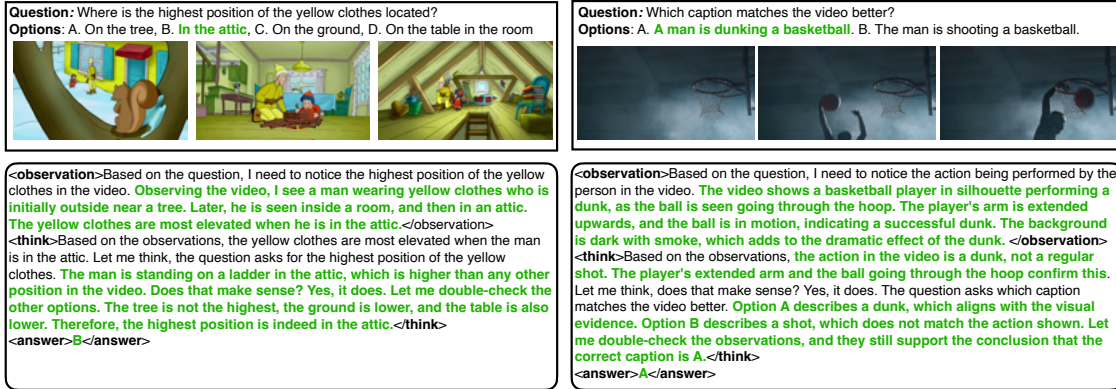


Figure 19. Success Cases of VIDEOP2R

Table 7. Performance across difficulty levels on Video-MME (perception, by video duration) and VideoMMMU (reasoning). Best result of each column is in **bold** (all numbers in %).

Model	Video-MME (Perception)			VideoMMMU (Reasoning)		
	Short	Medium	Long	Identify	Interpret	Adapt
Qwen2.5-VL-7B	68.49	55.92	51.06	68.67	48.67	42.00
VIDEOP2R	70.80	59.16	53.65	70.33	52.33	44.33
Δ	+2.31	+3.24	+2.59	+1.66	+3.66	+2.33

Table 8. Performance comparison on video reasoning and understanding benchmarks. Best/second-best result of each column is in **bold/underline**. Missing entries indicate unreported results (all numbers unit in %).

Model	Video Reasoning				Video Understanding			Avg
	VSI	VideoMMMU	MMVU	VCR	MV	TempCom.	VideoMME	
Qwen2.5-VL-72B	37.2	67.0	73.4	54.9	61.7	74.9	64.5	61.9
Open-Source 7B Models								
LLaVA-OneVision-7B	32.4	33.8	49.2	–	56.7	–	58.2	–
LongVA-7B	29.2	23.9	–	–	–	56.9	52.6	–
Video-UTR-7B	–	–	–	–	58.8	59.7	52.6	–
VideoLLaMA2-7B	–	–	44.8	–	54.6	–	47.9	–
Qwen2.5-VL-3B	26.9	34.3	51.5	35.1	47.0	23.5	50.7	38.4
Qwen2.5-VL-7B	30.1	48.1	60.0	44.3	59.0	72.6	56.6	52.9
RFT on Qwen2.5-VL-7B								
Video-R1	<u>35.8</u>	52.3	63.8	49.0	63.9	73.2	59.3	<u>56.8</u>
VideoChat-R1	33.9	<u>54.0</u>	63.0	49.0	<u>67.9</u>	72.5	57.7	<u>56.9</u>
Time-R1	29.0	51.0	62.9	49.6	63.1	73.7	59.3	55.5
VersaVid-R1	33.7	51.9	64.3	<u>49.8</u>	62.9	<u>74.0</u>	58.8	<u>56.5</u>
VideoRFT	36.8	51.1	68.5	49.6	62.1	73.7	<u>59.8</u>	<u>57.4</u>
VIDEOP2R (3B) (Ours)	38.7	45.3	58.7	45.3	63.2	66.4	55.9	53.3
VIDEOP2R (Ours)	36.8	55.0	<u>65.4</u>	51.0	68.1	74.5	60.0	58.7

beneficial and generalizes to other RFT pipelines, we compare it against Video-R1-CoT-165K, the CoT dataset used for SFT in Video-R1 [14]. Specifically, in the SFT stage, we train Qwen2.5-VL-7B on each dataset separately, and in the RL stage, we apply Video-R1’s temporal GRPO (T-GRPO) on top of both SFT checkpoints under identical settings. As shown in Tab. 9, training on VIDEOP2R-CoT-162K yields 55.6% aver-

age accuracy in the SFT stage, outperforming Video-R1-CoT-165K (53.9%) by 1.7%. After applying T-GRPO, our dataset further improves to 57.0%, still surpassing the Video-R1 counterpart (56.8%). These results suggest that VIDEOP2R-CoT-162K not only provides a stronger foundation in the SFT stage, but also generalizes well to other RFT pipelines.

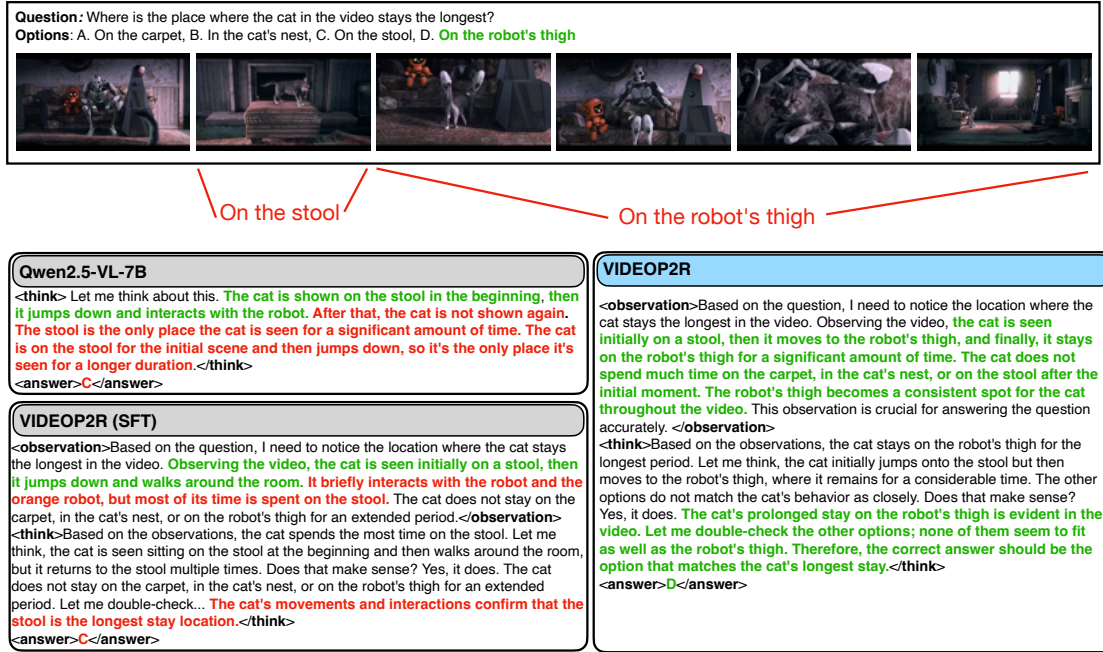
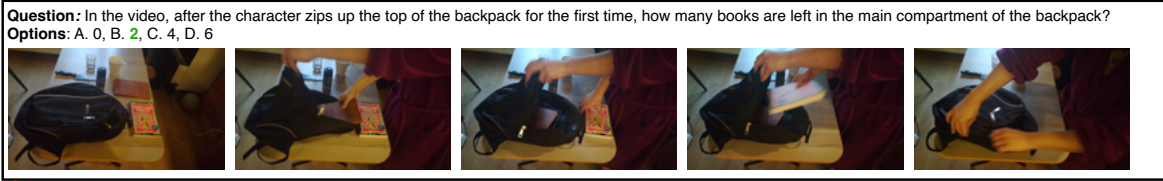


Figure 20. From base Qwen2.5-VL-7B to VIDEOP2R-SFT and VIDEOP2R: a representative example illustrating the stepwise improvement in model's perception and reasoning.

Table 9. Comparison between VIDEOP2R-CoT-162K and Video-R1-CoT-165K (all numbers in %). Best result of each column is in **bold**.

Method	Video Reasoning				Video Understanding			Avg
	VSI.	VideoMMMU	MMVU	VCR.	MV.	TempCom.	VideoMME	
SFT Stage								
SFT (Video-R1-CoT-165K)	33.3	49.4	63.5	45.5	60.5	69.9	55.4	53.9
SFT (VIDEOP2R-CoT-162K)	35.2	53.7	61.6	46.9	62.3	72.4	57.2	55.6
RL Stage								
SFT+T-GRPO (Video-R1-CoT-165K)	35.8	52.3	63.8	49.0	63.9	73.2	59.3	56.8
SFT+T-GRPO (VIDEOP2R-CoT-162K)	39.9	50.0	62.8	48.9	64.5	73.5	58.2	57.0



Two books were placed in the main compartment of the backpack.

The top of the backpack was zipped.

Qwen with the text question only

<think>The question asks how many books are left in the main compartment after this action. **However, the video itself is not provided, so I can't directly observe the movement.** If we assume that the character is packing books into the backpack, and the top compartment is being closed, it's likely that the books are being placed in the main compartment.

Given the options:
 A. 0 - This seems unlikely because the character would need to have some books to pack.
 B. 2 - This is a small number, but it's possible.
 C. 4 - This is a reasonable number, considering a typical backpack can hold several books.
 D. 6 - This is a larger number, but it's still possible depending on the size of the backpack.

Without more specific information about the number of books or the size of the backpack, the most reasonable answer based on typical backpack usage would be a moderate number. Therefore, the most likely answer is C. **</think>**

<answer>C</answer>

Qwen with the text question plus the perception segments (Qwen)

<observation>Based on the question, I need to notice the number of books left in the main compartment of the backpack after the character zips it up for the first time. **The video shows a person placing books into a backpack and then zipping it up. The books are visible on the table before being placed in the backpack, and the person's hands are shown closing the backpack's top zipper.** **</observation>**

<think>Let me think about this. The question asks specifically about the number of books left in the main compartment of the backpack after the first zip. **The observation mentions that the person places books into the backpack and then zips it up. However, it doesn't specify how many books are placed in the backpack or how many are left after the first zip.** Without more specific information about the number of books initially placed or the number left after the first zip, it's impossible to determine the exact number of books left in the main compartment.

Since the question doesn't provide enough details to make a definitive choice, and the options given are specific numbers, it seems the question might be based on a misunderstanding or lack of information. **</think>**

No answer

Qwen with the text question plus the video input

<think>Let me think through this step-by-step. Initially, there are two books on the table: one red and one with a yellow cover. **The person picks up the red book and places it into the backpack. Then, they pick up the yellow book and place it into the backpack as well. After these actions, the person zips up the top of the backpack, which presumably closes the main compartment.** At this point, both books should be inside the backpack.

Now, let's consider the question: "After the character zips up the top of the backpack for the first time, how many books are left in the main compartment of the backpack?"

Since both books were placed inside the backpack and the top was zipped up, there should be no books left outside the backpack. **Therefore, the number of books left in the main compartment of the backpack is zero.** **</think>**

<answer>A</answer>

Qwen with the text question plus the perception segments (VIDEOP2R)

<observation>Based on the question, I need to notice the number of books left in the main compartment of the backpack after the character zips up the top for the first time. **In the video, the character is seen placing books into the backpack and then zipping it up. After the first zip, there are two books visible in the main compartment of the backpack.** **</observation>**

<think>
 The question asks for the number of books left in the main compartment of the backpack after the character zips up the top for the first time. The observation clearly states that after the first zip, there are two books visible in the main compartment. This directly corresponds to option B. 2.

Let me think through this step-by-step:
 1. The character places books into the backpack.
 2. The character zips up the top of the backpack for the first time.
 3. After the first zip, the observation mentions that there are two books visible in the main compartment.
 4. The options provided are A. 0, B. 2, C. 4, and D. 6.
 5. The observation aligns perfectly with option B. 2.

Therefore, the number of books left in the main compartment of the backpack after the first zip is 2. **</think>**

<answer>B</answer>

Figure 21. Examples of Perception Examination: Top Left: Qwen with the text question only; Top Right: Qwen with the text question plus the perception segments from Qwen; Bottom Left: Qwen with the text question plus the video input; Bottom Right: Qwen with the text question plus the perception segments from VIDEOP2R. Green text denotes correct visual information or reasoning traces, while red text denotes incorrect or insufficient visual information or reasoning traces.